UNIVERSIDAD COMPLUTENSE DE MADRID
FACULTAD DE INFORMÁTICA
Departamento de Ingeniería del Software e Inteligencia Artificial

TESIS DOCTORAL

Gestión de colecciones digitales con esquemas de catalogación reconfigurables

MEMORIA PARA OPTAR AL GRADO DE DOCTOR

PRESENTADA POR

Joaquin Gayoso Cabada

Directores

José Luis Sierra Rodríguez
Ana María Fernández-Pampillón Cesteros
Antonio Sarasa Cabezuelo.

Madrid, 2018

UNIVERSIDAD COMPLUTENSE DE MADRID
FACULTAD DE INFORMÁTICA

Departamento de Ingeniería del Software e Inteligencia Artificial

GESTIÓN DE COLECCIONES DIGITALES CON ESQUEMAS DE CATALOGACIÓN RECONFIGURABLES

TESIS DOCTORAL

Presentada por:
Joaquín Gayoso Cabada

Bajo la dirección de los Doctores:
José Luis Sierra Rodríguez
Ana María Fernández-Pampillón Cesteros
Antonio Sarasa Cabezuelo

Madrid, 2017

GESTIÓN DE COLECCIONES DIGITALES CON ESQUEMAS DE CATALOGACIÓN RECONFIGURABLES

Memoria que presenta para optar al título de Doctor en Informática:
Joaquín Gayoso Cabada

Bajo la dirección de los Doctores:
José Luis Sierra Rodríguez
Ana María Fernández-Pampillón Cesteros
Antonio Sarasa Cabezuelo

Universidad Complutense de Madrid
Facultad De Informática
Departamento de Ingeniería del Software e Inteligencia Artificial

Madrid, 2017

Para mis sobrinos Jesús y Eva

Agradecimientos

Agradezco el apoyo recibido durante estos años por parte de todos los miembros de mi grupo de investigación ILSA en la Facultad de Informática de la Universidad Complutense de Madrid. También a los grupos de investigación LEETHI y LOEP pertenecientes también a la Universidad Complutense, y a la Fundación El Caño de Panamá, sin los que no habría podido realizar parte de los experimentos expuestos en los trabajos.

A título personal, deseo agradecer a mis directores José Luis Sierra, Ana Fernández-Pampillón, Antonio Sarasa, y compañeros de grupo de investigación Alfredo Fernández Valmayor, Daniel Rodríguez, Bryan Temprado y César Ruiz por darme la oportunidad de desarrollar estos años de investigación con ellos sobre este campo, esfuerzo que concluye en esta tesis, y por todo lo que me han enseñado sobre cómo ser un buen investigador.

Dentro de la universidad también deseo dar las gracias a mis compañeros del "Aula16": Toni, Dan, Iván, Víctor, Jesús, Pablo, Cristina y Marta con los que he compartido muchas comidas, y cafés, a lo largo de estos años divagando sobre informática. También quiero dar las gracias a mis actuales compañeros del "420bip": Susana, Vicky, Carlos y Noelia, que me han visto dando los últimos remates estos meses a esta tesis y me han ayudado en todo lo que han podido.

Quiero agradecer también a los amigos que he creado en Madrid en estos años: Gorka Suárez, Alan Somoza y Raquel García, que no sólo son mis amigos, sino que han sido mis compañeros en esta carrera de fondo que es la informática. También se lo agradezco a mis amigos de fuera del gremio: Jorge Martín, Manuel Salinero, Guillermo García, Antonio Oeo, Manuel Vera, Marta Sáenz, Sara Isabel y Mariola Lorente, sin los que muchas veces no podría haber desconectado del trabajo para poder tomar una cerveza sin hablar de nada que tenga que ver con la informática.

Pero especialmente, deseo dar las gracias a mi familia: mis padres, Joaquín y Concepción; mis suegros, Rosa Sánchez y Pedro García; mi hermana Inmaculada, mis cuñados Juan García, Esmeralda García y Raúl García. Una mención especial en esta tesis tienen mis sobrinos Jesús Colomo y Eva Gayoso, que me han dado las alegrías y los besos necesarios para acabar esta tesis. Pero sobre todo deseo agradecérselo a mi esposa Lourdes García, sin la cual nada de esto habría sido posible, desde que fue mi compañera de prácticas durante la carrera, pareja y novia durante el master y esposa durante algo más del último año de la tesis doctoral, con la que a veces, por trabajo y nervios, no he podido estar tanto tiempo como quisiera.

Esta tesis se ha llevado a cabo en el contexto de los proyectos HUM14_251 (Programa de Ayudas de la Fundación BBVA a Proyectos de Investigación, Convocatoria 2014), y TIN2014-52010-R (Subprograma de Investigación Orientada a los Retos de la Sociedad, convocatoria 2014 del Plan Nacional de I+D+i), así como en el del proyecto "Collaborative Annotation of Digitalized Literary Text" del Digital Humanities Award Program de Google (convocatorias 2010 y 2011).

Joaquín Gayoso Cabada

Madrid, 14 de marzo del 2017

Índice de Contenidos

Resumen

El *esquema de catalogación* de una colección digital sirve para catalogar (describir y clasificar) adecuadamente los objetos que la conforman. Dicha catalogación es esencial para habilitar la explotación efectiva de la colección, tanto por parte del sistema de gestión de la misma (*repositorio digital*), como por parte de las herramientas externas encargadas de recuperar, reproducir y agregar los objetos. Efectivamente, funcionalidades básicas como la recuperación de recursos a partir de consultas, o la navegación guiada, dependen todas ellas de la adecuada catalogación de los recursos.

De esta forma, en el dominio de las bibliotecas digitales se han propuesto múltiples estándares para la catalogación de objetos digitales, cuyo propósito último es garantizar la interoperabilidad entre los distintos repositorios y aplicaciones que manipulan los objetos. No obstante, la adopción de esquemas de catalogación preestablecidos puede no ser una solución satisfactoria en aquellos escenarios donde los esquemas de catalogación, en lugar de establecerse *a priori*, son artefactos que cambian y evolucionan a lo largo de todo el ciclo de vida de la colección, con el fin de adaptarse a requisitos y necesidades de catalogación cambiantes, que no se conocen al inicio de la producción de la colección, sino que aparecen y maduran al mismo tiempo que la colección crece y evoluciona. Este ha sido el caso en nuestras experiencias colaborando con distintos grupos de Humanidades en distintos dominios (arqueología, literatura convencional, literatura digital, escritura creativa, etc.), experiencias que han mostrado que la definición inductiva de esquemas de catalogación (es decir, la evolución concurrente de esquemas y colecciones), lejos de ser una excepción es la realidad habitual en la creación de colecciones muy específicas, frecuentemente orientadas a la investigación y docencia.

Con el fin de permitir que los esquemas de catalogación evolucionen y se adapten a las necesidades cambiantes que surgen durante el ciclo de vida de una colección, es necesario proporcionar mecanismos que permitan *reconfigurar* dichos esquemas conforme la colección cambia. Esta tesis aborda este problema de la gestión de colecciones de objetos digitales con esquemas de catalogación que pueden reconfigurarse dinámicamente.

De esta forma, el primer aspecto abordado en esta tesis tiene que ver con la *dimensión lógica* del problema de la reconfiguración de esquemas: ¿cómo se puede expresar de manera efectiva dicha reconfiguración? Para ello, la tesis propone enfocar el problema desde el nivel *estructural* o *sintáctico* de los esquemas de catalogación, entendiendo la reconfiguración como la aplicación de un conjunto básico de operaciones de edición sobre la estructura de los

esquemas (de forma similar a como, por ejemplo, el procesamiento de un lenguaje informático se centra en la sintaxis abstracta del lenguaje). Más concretamente, la tesis propone centrarse en representaciones arborescentes de los esquemas, y en utilizar un conjunto básico de operaciones de edición sobre dichas representaciones (añadido, renombrado y eliminación de nodos, cambio de filiación de nodos en la jerarquía, y fusión de nodos para resolver problemas de sinonimia). La reconfiguración en sí será llevada a cabo por los expertos en el dominio de la colección, utilizando editores adecuados, expertos que cuidarán de que las reconfiguraciones tengan sentido a niveles semánticos y pragmáticos (de forma similar a como ocurre con el marcado descriptivo de un documento, utilizando, por ejemplo, un lenguaje de marcado descriptivo definido mediante SGML o XML). La tesis muestra la factibilidad de esta propuesta mediante la aplicación del enfoque a distintos tipos de esquemas de catalogación, tanto esquemas basados en *metadatos* como basados en *vocabularios controlados* (listas de términos, taxonomías y tesauros facetados).

El segundo aspecto abordado por la tesis tiene que ver con la *dimensión humana* del proceso de reconfiguración: ¿cómo guiar al experto que realiza la reconfiguración del esquema en esta actividad? Para ello, la tesis propone aplicar un enfoque basado en *análisis de datos*, según el cual la colección catalogada se analiza para inducir una estructura comparable con el esquema original. El experto, entonces, puede comparar la estructura inferida a partir de la colección con el esquema inicialmente propuesto, y utilizar las diferencias encontradas para, bien diagnosticar usos potencialmente erróneos del esquema durante el proceso de catalogación, bien para reconfigurar el esquema con el fin de adaptar el mismo a la realidad de su uso. Este enfoque se ha aplicado al caso particular de la catalogación de anotaciones sobre textos literarios clasificadas mediante taxonomías. La técnica de análisis de datos utilizada ha sido el *análisis de conceptos formales*, lo que ha permitido inducir organizaciones reticulares de las anotaciones. Dichos retículos pueden compararse de manera significativa con las taxonomías originales, ofreciendo a los expertos, de esta forma, una guía valiosa de cara a la reconfiguración.

Por último, el tercer aspecto abordado en esta tesis se refiere al *factor tecnológico* del proceso de reconfiguración. Efectivamente, para que el sistema de gestión de colecciones pueda llevar a cabo funcionalidades básicas como la navegación o la búsqueda a las que se ha hecho referencia anteriormente, es necesario disponer de índices apropiados de los objetos. Dichos índices se construyen a partir de la catalogación de los mismos. Por tanto, si el esquema de catalogación se reconfigura, dichos índices pueden invalidarse parcial, o incluso totalmente. Para evitar este efecto es necesario proponer índices que permanezcan invariantes a la

reconfiguración de los esquemas. No obstante, esto puede acarrear, a su vez, una merma considerable en el rendimiento, frente a índices dependientes de un esquema específico. En esta tesis se proponen propuestas de indexado que permiten minimizar el impacto del rendimiento producido por las reconfiguraciones en el caso de una funcionalidad crítica del sistema de gestión de colecciones: la navegación guiada por los esquemas de catalogación. Para ello, la tesis comienza modelizando dicha navegación mediante un autómata finito determinista, el *autómata de navegación*, y muestra cómo el enfoque puede aplicarse con los distintos tipos de esquemas de catalogación contemplados (tanto esquemas de metadatos, como los basados en vocabularios controlados). Mientras que dicho autómata puede proporcionar directamente el índice requerido, se muestra cómo, en el peor de los casos, su tamaño puede crecer exponencialmente con el tamaño de la colección. Por tanto, se proponen alternativas de indexado que permiten recrear dinámicamente las partes relevantes de dicho autómata durante la navegación: una alternativa clásica, basada en *índices invertidos*, y una alternativa más sofisticada, basada en *dendrogramas*. La tesis analiza el comportamiento en la práctica de ambas alternativas, y muestra, por último, cómo la alternativa basada en dendrogramas puede mejorar sustancialmente la basada en índices invertidos.

Abstract

The *cataloging scheme* of a digital collection serves to adequately catalog (describe and classify) the objects that comprise it. Such cataloging is essential to enabling the effective exploitation of the collection, both by the collection management system (*digital repository*), and by the external tools used for retrieving, reproducing and adding objects. Indeed, basic functionalities such as the retrieval of resources from queries, or guided navigation, all depend on the proper cataloging of resources.

In this way, in the domain of digital libraries, several standards have been proposed for the cataloging of digital objects, whose ultimate purpose is to guarantee interoperability among the different repositories and applications that manipulate the objects. However, the adoption of pre-established cataloging schemata may not be a satisfactory solution in those scenarios where cataloging schemata, instead of being established a priori, are artifacts that change and evolve throughout the entire collection's life cycle (the rationality of this evolving nature is to adapt to changing cataloging requirements and needs, which are not known at the beginning of the collection, but appear and mature as the collection grows and evolves). We realized this fact during our collaboration with different humanities groups in different domains (archeology, conventional literature, digital literature, creative writing, etc.). From these experiences we learned that the inductive definition of cataloging schemata (i.e., the concurrent evolution of schemata and collections), far from being an exception, is the norm for the production of very specific collections, often oriented to research and teaching.

In order to allow cataloging schemata to evolve and adapt to the changing needs that arise during the life cycle of a collection, it is necessary to provide mechanisms to *reconfigure* these schemata as the collection changes. This thesis addresses this problem of managing collections of digital objects with cataloging schemata that can be dynamically reconfigured.

In this way, the first aspect addressed in this thesis has to do with the *logical dimension* of the schemata reconfiguration problem: how can this reconfiguration be effectively described? For this purpose, the thesis proposes approaching the problem from the *structural* or *syntactic* level of the cataloging schemata, understanding the reconfiguration as the application of a basic set of editing operations on the structure of the schemata (in a similar way as, for instance, computer language processing is focused on abstract syntax). More specifically, the thesis proposes focusing on tree-like representations of schemata, and using a basic set of editing operations on such representations (the addition, renaming and deletion of nodes, changing node parents in the hierarchy, and merging nodes to avoid synonymy). The

reconfiguration itself will be carried out by experts in the collection domain, using suitable editors, experts who will ensure that the reconfigurations make sense at semantic and pragmatic levels (similarly to what happens with descriptive document markup, as promoted by SGML or XML). The thesis demonstrates the feasibility of this proposal by applying the approach to different types of cataloging schemata: *metadata* schemata, and *controlled vocabularies* (term lists, taxonomies and faceted thesauri).

The second aspect addressed by the thesis has to do with the *human dimension* of the reconfiguration process: how to guide the expert who reconfigures the scheme along this reconfiguration activity? For this purpose, the thesis proposes applying a data analysis approach, according to which the cataloged collection is analyzed to induce a structure comparable to the original scheme. The expert can then compare the inferred structure with the initially proposed scheme. The differences found can serve to diagnose potentially misleading uses of the scheme during the cataloging process. More importantly, these differences can also serve to reconfigure the schema in order to adapt it to the way in which it is actually being used. This approach was applied to the particular case of digital annotation in literary texts catalogued according to taxonomies. The data analysis technique used was formal concept analysis. It allowed the researchers to infer the lattice-like organizations of the annotations. Such lattices can be compared to the original taxonomies in a significant way, which provides experts with valuable information for reconfiguration.

Finally, the third aspect addressed in this thesis refers to the *technological factor* of the reconfiguration process. Indeed, for the collection management system to perform basic functionalities, such as navigation or search, it is necessary to maintain appropriate indices of the objects. These indices are built from the cataloging of these objects. Therefore, if the cataloging scheme is reconfigured, these indices may be partially, or even completely, invalidated. To avoid this effect, it is necessary to propose indices that remain invulnerable to schemata reconfiguration. However, this can, in turn, lead to a considerable decrease in performance, compared to using scheme-specific indices. This thesis recommends indexing proposals that minimize the impact on performance caused by reconfigurations in the case of a critical function of the collection management system: navigation guided by the cataloging schemata. For this purpose, the thesis begins by modeling such a navigation using a deterministic finite automaton, the navigation automaton, and shows how the approach can be applied to the different types of cataloging schemata envisioned (both metadata schemata and schemata based on controlled vocabularies). While such an automaton can directly provide the index required, this thesis shows how, in the worst case, its size can grow exponentially with

the size of the collection. Therefore, indexing alternatives are proposed that allow the relevant parts of the automaton to be dynamically recreated during navigation: a classic alternative, based on *inverted indices*, and a more sophisticated alternative based on *dendrograms*. The thesis analyzes the practical behavior of both alternatives, and shows, finally, how the dendrogram-based alternative can substantially improve the one based on inverted indices.

Capítulo 1 - Introducción

1.1 Motivación de la investigación

Esta tesis aborda el problema de la reconfiguración dinámica de los esquemas de catalogación de colecciones de objetos digitales. Los objetos digitales son un tipo de recurso digital constituido por un contenido y una descripción de dicho contenido (i.e. metadatos). Los objetos digitales y las colecciones de objetos digitales se describen y se clasifican mediante los esquemas de catalogación. Los esquemas de catalogación, en consecuencia, constituyen una herramienta vital para la explotación (i.e. exploración, recuperación, utilización) de las colecciones de objetos digitales en un sistema de gestión de colecciones digitales, sistemas que, actualmente, se conocen como repositorios digitales. Efectivamente, las funcionalidades más características soportadas por un repositorio digital, tales como la navegación por los contenidos de las colecciones o la búsqueda de objetos digitales, dependen, en última instancia, de las descripciones de los objetos.

A este respecto, la tendencia generalizada para llevar a cabo la catalogación de objetos en colecciones digitales es hacia el establecimiento de estándares que normalicen la manera en la que los recursos son descritos. Dos ejemplos paradigmáticos de este tipo de estándares son el estándar Dublin Core (D. C. M. I. DCMI, 2012) ampliamente utilizado tanto a nivel global en la web como en dominios especializados como en el de las Bibliotecas Digitales, y LOM (IEEE, 2001) definido para el dominio específico de colecciones de materiales educativos. La adopción de estándares hace posible abordar un aspecto básico en la gestión de colecciones de objetos digitales como es el de la *interoperabilidad* entre distintos repositorios de colecciones digitales, así como entre otras aplicaciones orientadas a explotar los contenidos de los repositorios como, por ejemplo, los reproductores de contenidos de los objetos digitales. No obstante, el principal problema de este tipo de estándares surge cuando no se adaptan satisfactoriamente a las necesidades de creación, actualización y preservación de colecciones en un determinado dominio. Este problema aparece con frecuencia, por ejemplo, en el dominio de las Humanidades Digitales (D. Berry, 2012). Efectivamente, en los escenarios de trabajo humanístico, a pesar de que existen protocolos internacionales para la catalogación de las obras y colecciones de las bibliotecas y museos tradicionales, también existen dificultades para su utilización en la creación de colecciones digitales con un propósito específico, que normalmente surgen y se utilizan en el ámbito de la investigación y la docencia. En estos contextos, los esquemas de catalogación necesitan ser revisados y modificados con frecuencia

durante la creación de las colecciones de objetos digitales e, incluso, a lo largo de toda su vida (José Luis Sierra, Fernández-Valmayor, Guinea, & Hernanz, 2006). En este sentido, existe la necesidad de encontrar mecanismos que faciliten la *creación inductiva* de esquemas de catalogación a partir de los objetos digitales que se van incorporando de forma dinámica a las colecciones. De esta forma, en lugar de adoptar un esquema de catalogación preestablecido, estos mecanismos deben permitir formular los esquemas más apropiados para cada escenario, y, lo que resulta más importante, *reconfigurar* dichos esquemas a lo largo de todo el ciclo de vida de la colección, con el fin de adaptarse a los nuevos objetos digitales que vayan incorporándose a las colecciones. La reconfiguración de esquemas de catalogación necesita:

- Disponer de modelos y mecanismos de esquemas de catalogación *reconfigurables*, que los expertos encargados en la creación y la preservación de las colecciones puedan utilizar para adaptar los esquemas a las distintas necesidades y circunstancias que surgen durante la vida de dichas colecciones.
- Disponer de enfoques que permitan ofrecer a los expertos, en la medida de lo posible, la información necesaria para planificar, de manera fundamentada, las reconfiguraciones de los esquemas.
- Dotar a las plataformas de gestión del aparato necesario para minimizar los efectos que las reconfiguraciones en los esquemas puedan tener sobre el rendimiento global de las mismas. Esto es especialmente crítico en escenarios en los que las reconfiguraciones en los esquemas son frecuentes, y en los que los usuarios esperan ver reflejadas en el acto las reconfiguraciones en el resto de las funcionalidades del sistema de gestión.

En este sentido, esta tesis tiene como propósito abordar cada una de estas tres necesidades de la gestión de colecciones digitales con esquemas de catalogación reconfigurables. La tesis se encuadra en una línea de investigación de creación y mantenimiento de repositorios digitales en dominios especializados iniciada hace más de quince años en la Facultad de Informática de la Universidad Complutense de Madrid por el Prof. Fernández-Valmayor, y continuada por el Grupo de Investigación en Ingeniería de Lenguajes Software y Aplicaciones (ILSA), en cuyo seno se ha desarrollado la presente tesis.

1.2 Objetivos y planteamiento de la línea de investigación

Tal y como se ha indicado anteriormente, la tesis se centra en el problema de la reconfiguración de esquemas de catalogación de colecciones digitales, abordando tres aspectos críticos relacionados con dicho problema: (i) cómo deben ser los modelos en los que se basan los esquemas que pueden ser reconfigurados dinámicamente a lo largo del ciclo de vida de las colecciones, (ii) cómo puede guiarse a los creadores y preservadores de las colecciones en el proceso de reconfiguración, y (iii), cómo deben organizarse internamente los sistemas de gestión de colecciones para hacer frente a las reconfiguraciones. De esta forma, se plantean los siguientes objetivos:

- El primer objetivo se centra en el *aspecto lógico* del problema y consiste en proponer mecanismos de reconfiguración para distintos tipos de esquemas de catalogación.
- El segundo objetivo se centra en el *aspecto humano* del proceso de reconfiguración, planteando la propuesta de enfoques que guíen a los expertos en la creación y mantenimiento de colecciones en la reconfiguración de los esquemas.
- El tercer objetivo se centra en el *aspecto tecnológico*, a través de la búsqueda de propuestas organizativas y arquitectónicas que permitan a las plataformas de gestión reaccionar dinámicamente a las reconfiguraciones en los esquemas.

Para abordar el primer objetivo, en lugar de buscar un modelo de esquema de catalogación reconfigurable *universal* (lo que, a su vez, sería similar a tratar de formular *otra* propuesta de normalización más), se plantea como solución analizar la reconfiguración para distintos tipos de esquemas de catalogación representativos, tanto desde el punto de vista *descriptivo* (esquemas basados en *metadatos*), como desde el punto de vista *temático* (esquemas basados en *vocabularios controlados*). Para ello, se propone centrar los esfuerzos en los aspectos *estructurales* de los esquemas, y entender la reconfiguración como un proceso de edición de dichas estructuras de catalogación. Dicha edición permitirá operaciones del tipo de *cambiar* la posición de determinados elementos de información en la estructura de catalogación, *renombrar*, *fusionar*, *modificar* y *eliminar* elementos o *crear* otros nuevos. A este respecto, se piensa que este enfoque, con el que ya se ha tenido cierta experiencia en trabajos previos en el grupo de investigación (más concretamente, en la reconfiguración de esquemas de metadatos en un sistema para la creación de repositorios digitales en dominios

especializados), podrá extrapolarse de manera significativa a distintos tipos de esquemas de catalogación, tanto esquemas de metadatos, como distintos tipos de vocabularios controlados.

El segundo objetivo se aborda planteando un enfoque basado en técnicas de *análisis de datos*. Para ello, utilizando técnicas adecuadas, se analizarán las colecciones catalogadas con el fin de inducir la forma *real* en la que los expertos han utilizado los esquemas de catalogación para catalogar los objetos digitales de dichas colecciones. De esta manera, el propósito será, en cierta medida, ser capaces de *inducir* los esquemas de catalogación que *realmente* se han utilizado durante la catalogación para crear las descripciones de los objetos. Esto permitirá a los expertos comparar los esquemas creados inicialmente por ellos con los inducidos del uso de los mismos en la catalogación. En dicha comparación podrán, por ejemplo, descubrir conceptos en los esquemas que no se están utilizando, conceptos que se están utilizando erróneamente, conceptos considerados equivalentes por los catalogadores, etc. Con ello, los expertos podrán planificar las posibles reconfiguraciones de una manera *fundamentada* en el uso real de los esquemas, con el fin de adecuar estos esquemas a las necesidades reales de catalogación de los objetos.

Por último, el tercer objetivo implicará encontrar propuestas de indexación de colecciones que produzcan índices en los que basar las distintas funcionalidades de los sistemas de gestión (i.e., navegación, búsqueda, etc.) y que, a su vez, permanezcan invariantes a las distintas reconfiguraciones de los esquemas. Dichos índices deberán, a su vez, soportar eficientemente las citadas funcionalidades, evitando en la medida de lo posible que el rendimiento del sistema se degrade tras cada reconfiguración. Así mismo, los requisitos temporales y espaciales asociados al mantenimiento de dichos índices deberán ser razonables desde un punto de vista práctico.

1.3 Estructura de la memoria

La Universidad Complutense de Madrid contempla la posibilidad de organizar la memoria de una tesis doctoral en torno a un compendio de artículos editados y publicados. La presente memoria sigue dicho formato. De esta forma, el grueso de esta tesis doctoral se presenta como un compendio integrado por los siguientes artículos:

- Gayoso-Cabada, J., Ruiz, C., Pablo-Nuñez, L., Sarasa-Cabezuelo, A., Goicoechea-de-Jorge, M., Sanz-Cabrerizo, A., & Sierra-Rodriguez, J.-L. (2012). A flexible model for the collaborative annotation of digitized literary works. En Proceedings

of the 2012 Digital Humanities Conference, DH 2012 (pp. 190–193). (Joaquin Gayoso-Cabada et al., 2012)

- Gayoso-Cabada, J., Sanz-Cabrerizo, A., & Sierra, J.-L. (2013). @Note: An Electronic Tool for Academic Readings. En Proceedings of the 1st International Workshop on Collaborative Annotations in Shared Environment: Metadata, Vocabularies and Techniques in the Digital Humanities, DH-CASE 2013 (paper 17, 4 pags). Florence, Italy: ACM (Workshop del 13th ACM Symposium on Document Engineering, DocEng 2013, simposio **CORE A** en la edición 2013 del CORE). (Joaquín Gayoso-Cabada, Sanz-Cabrerizo, & Sierra, 2013)
- Cigarrán-Recuero, J., Gayoso-Cabada, J., Rodríguez-Artacho, M., Romero-López, M.-D., Sarasa-Cabezuelo, A., & Sierra, J.-L. (2014). Assessing semantic annotation activities with formal concept analysis. Expert Systems with Applications, 41(11), 5495–5508. (**ISI WoK JCR 2014**: **2.24**. *Computer Science, Artificial Intelligence*: 29/123 **Q1**; *Engineering, Electrical & Electronic*: 48/249 **Q1**; *Operation Research & Management Science*: 12/81 **Q1**). (Cigarrán-Recuero et al., 2014)
- Gayoso-Cabada, J., Rodríguez-Cerezo, D., & Sierra, J.-L. (2016). Browsing Digital Collections with Reconfigurable Faceted Thesauri. En Proceedings of the 25th International Conference on Information Systems Development, ISD 2016 (pp. 378-389) (**CORE 2014: A**). (Joaquín Gayoso-Cabada, Rodríguez-Cerezo, & Sierra, 2016a)
- Gayoso-Cabada, J., Rodríguez-Cerezo, D., & Sierra, J.-L. (2016). Multilevel Browsing of Folksonomy-Based Digital Collections. En Proceedings of the 17th Conference on Web Information Systems Engineering Part II, WISE 2016 (pp. 43-51). Lecture Notes in Computer Science 10042, Springer. (**CORE 2014: A**). (Joaquín Gayoso-Cabada, Rodríguez-Cerezo, & Sierra, 2016b)
- Gayoso-Cabada, J., Rodríguez-Cerezo, D., & Sierra, J.-L. (2016). Learning object repositories with dynamically reconfigurable metadata schemata. En Proceedings of the XVIII International Symposium on Computers in Education, SIIE 2016 (6 pags). IEEE Computer Society. (Joaquín Gayoso-Cabada, Rodríguez-Cerezo, & Sierra, 2016c)

De esta forma, esta memoria, aparte de esta introducción, desarrolla los siguientes capítulos:

- El Capítulo 2 presenta una revisión de conceptos implicados en este trabajo de tesis. Para ello comienza introduciendo los principales conceptos relativos a la creación de colecciones de objetos digitales y a su catalogación. Seguidamente se centra en los distintos tipos de esquemas de catalogación normalmente utilizados en la descripción de objetos digitales: *esquemas de metadatos*, y *vocabularios controlados*. A continuación, aborda los aspectos relativos a las tecnologías más relevantes utilizadas en el soporte de este tipo de esquemas. Seguidamente aborda los aspectos relativos a la reconfiguración de esquemas de catalogación. Por último, resume los trabajos previos llevados a cabo en el Grupo de Investigación, y que han servido como punto de partida para el desarrollo de esta tesis.
- El Capítulo 3 plantea los objetivos principales de esta tesis, así como la planificación del trabajo ideada para acometer dichos objetivos. Estos objetivos, y la planificación del trabajo planteada para acometerlos, han sido introducidos con anterioridad en la sección 1.2 .
- El Capítulo 4 describe las diferentes contribuciones aportadas por los artículos anteriormente aludidos, que acompañan a este trabajo de tesis, a la hora de acometer los objetivos planteados. La sección 1.4 resume estas contribuciones de acuerdo con los objetivos planteados en la sección 1.2 .
- El Capítulo 5 describe las conclusiones finales de este trabajo de tesis, y plantea diferentes líneas de trabajo futuro que surgen de los resultados obtenidos al acometer los objetivos planteados.
- Por último, el Capítulo 6 contiene los diferentes artículos que constituyen el compendio de artículos de esta tesis, en su versión original.

1.4 Resumen de las contribuciones de esta tesis

Los objetivos planteados en este trabajo de tesis cristalizan en tres contribuciones que se desarrollarán a lo largo de los siguientes capítulos. En esta sección se resumen estas contribuciones y los artículos en los que se han publicado los resultados de dichas contribuciones.

La primera contribución consiste en una propuesta de mecanismos de reconfiguración de distintos tipos de esquemas de catalogación, tanto basados en metadatos, como basados en vocabularios controlados. Dicha propuesta se centra, como ya se ha indicado, en los aspectos

estructurales, es decir, en el nivel *sintáctico* de los esquemas. A este respecto, resulta notorio cómo es posible aplicar recurrentemente el mismo conjunto de operaciones de edición de estructuras, básicamente, edición de estructuras arborescentes, a los distintos tipos de esquema: operaciones de creación, eliminación, renombrado, fusión y reubicación de nodos en la jerarquía. Los artículos que registran los resultados de esta contribución son (Joaquin Gayoso-Cabada et al., 2012), (Joaquín Gayoso-Cabada et al., 2013), (Joaquín Gayoso-Cabada et al., 2016a), (Joaquín Gayoso-Cabada et al., 2016b) y (Joaquín Gayoso-Cabada et al., 2016c).

La siguiente contribución se centra en una propuesta para dar soporte a los expertos del dominio durante la reconfiguración de los esquemas de catalogación. La propuesta en sí se centra en un tipo particular de esquema de catalogación basado en vocabularios controlados, y se ajusta a la citada estrategia basada en análisis de datos. Se muestra cómo es posible inducir estructuras significativas de las colecciones, comparables con los esquemas de catalogación propuestos, y cómo dichas estructuras son de utilidad para mejorar los esquemas de partida, y también para corregir posibles malos usos durante el proceso de catalogación. Esta contribución se registra en (Cigarrán-Recuero et al., 2014).

Por último, la tercera contribución se refiere a la organización interna de las plataformas de gestión de colecciones con el fin de amortiguar adecuadamente el potencial impacto en el rendimiento provocado por las reconfiguraciones de los esquemas. La contribución en sí se centra en una funcionalidad crítica de este tipo de plataforma: la navegación por las colecciones guiada por los esquemas. Se propone un modelo conceptual que sirve de base para la indexación de las colecciones, y se proponen mecanismos prácticos de indexación que permiten recrear de manera razonablemente eficiente el comportamiento derivado de dicho modelo. Esta contribución queda registrada en los siguientes trabajos: (Joaquín Gayoso-Cabada et al., 2016a), (Joaquín Gayoso-Cabada et al., 2016b) y (Joaquín Gayoso-Cabada et al., 2016c).

Todas estas contribuciones, y su relación con los artículos que acompañan y avalan este trabajo de tesis, se describen en profundidad en el Capítulo 4.

Capítulo 2 - Estado de la cuestión

Este capítulo revisa los aspectos más relevantes para la investigación desarrollada en esta tesis doctoral. La sección 2.1 introduce la problemática relativa a la gestión de colecciones de objetos digitales y a su catalogación. La sección 2.2 revisa los aspectos relativos a la catalogación de objetos digitales mediante metadatos. La sección 2.3 revisa los aspectos relativos a la catalogación mediante vocabularios. La sección 2.4 revisa las tecnologías más relevantes para la representación de esquemas de catalogación. La sección 2.5 discute el aspecto relativo a la reconfiguración de colecciones digitales. La sección 2.6 resume los trabajos previos realizados en el Grupo de Investigación en el que se ha realizado esta tesis que son relevantes de cara a la misma. Por último, la sección 2.7 concluye el capítulo.

2.1 Colecciones de objetos digitales y su catalogación

2.1.1 *Introducción a la catalogación bibliográfica y documental*

En el Área de Biblioteconomía y Documentación la catalogación es "el proceso de descripción y organización de una colección de recursos con el fin de que sean fácilmente recuperables" (Lamarca Lapuente, 2006). Se entiende por *recurso* (también denominado *recurso bibliográfico*) "una entidad, tangible o intangible, que recoge el contenido intelectual y/o artístico y que está concebida, producida y/o editada como una unidad, constituyendo la base de una descripción bibliográfica única. Los recursos incluyen texto, música, imágenes fijas y en movimiento, gráficos, mapas, grabaciones sonoras y videograbaciones, datos o programas electrónicos, incluyendo los publicados de forma seriada" (IFLA & BNE, 2014).

La catalogación comienza con la descripción de cada recurso en lo que se denomina *registro bibliográfico*[1] y termina con la confección de un *catálogo*, que es el índice ordenado de registros bibliográficos que representan los fondos de una biblioteca y que constituyen su memoria (Carrión Gútiez, 1988).

Un registro bibliográfico está formado por un conjunto de campos (o atributos) y valores (Figura 1). Su objetivo es describir de forma única un recurso bibliográfico o documental (en adelante recurso) para permitir su identificación, selección y localización rápida y precisa. Se pueden distinguir tres tipos de campos-valores (Rodríguez Bravo, 2011): (i) los de identificación y descripción del recurso, (ii), los de localización en la colección o en

[1] También se denomina "asiento bibliográfico", pero es un término más antiguo (Rodríguez Bravo, 2011).

la biblioteca (punto de acceso principal o encabezamiento, puntos de acceso secundarios, signatura y número de registro) y, (iii) los de indexación, basados en vocabularios. Los campos-valores de descripción facilitan la selección por parte de los usuarios de los recursos buscados, mientras que los de localización e indexación servirán para la búsqueda y recuperación del recurso.

Ver signatura/s Registro del catálogo
Título **Universo y planetas**
Autor Martín Ávila, Pablo
Editor: Libsa
Fecha de pub.: D.L. 2016
Páginas: 128 p.
ISBN: 9788466231213
Información de ejemplar: 2 ejemplares disponibles en Sede de Alcalá.

FONDOS

Sede de Alcalá	Código de barras	Tipo de material	Localización
12/1124175	1105409829	Fondo moderno (posterior a 1958 inclusive)	Salón General-Petición anticipada
DL/2438426	1105409830	Préstamo restringido	Ejemplar de conservación

Universo y planetas
Martín Ávila, Pablo

N.º depósito legal: M 10544-2016 Oficina Depósito Legal Madrid
ISBN: 978-84-662-3121-3
CDU: 52
Autor personal: Martín Ávila, Pablo
Título: Universo y planetas / Pablo Martín Ávila
Publicación: Alcobendas, Madrid : Libsa, D.L. 2016
Descripción física: 128 p. : il. col. ; 28 cm
Tipo de contenido: Texto (visual)
Tipo de medio: sin mediación
Encabez. materia: Astronomía

Figura 1. Ejemplo de registro bibliográfico del catálogo de la Biblioteca Nacional Española[2]

Las formas de representación de un registro bibliográfico, es decir el conjunto de campos y valores, son varias y han ido cambiando a lo largo del tiempo desde que comenzó la práctica de la catalogación. En España, por ejemplo, la catalogación comienza aproximadamente en el siglo XVIII, con las reglas de catalogación elaboradas por Pedro García, bibliotecario de la Real Biblioteca (BNE, 2011a). Actualmente, se utilizan esquemas bibliográficos normalizados con el fin de facilitar el acceso universal a los recursos y la

[2] Fuente: http://catalogo.bne.es/

compartición de recursos entre instituciones. Instituciones nacionales (AENOR) e internacionales (FID, IFLA, ISO, ANSI y Unesco), en este sentido, han creado estándares para el formato de los registros bibliográficos, y, también, sobre cómo aplicarlos: las reglas de catalogación, la terminología y los vocabularios que describen el contenido de los registros. Actualmente, a nivel internacional conviven los estándares ISBD[3] (IFLA, 2011), el más antiguo y creado por la IFLA[4] y el estándar angloamericano RDA[5] (Coyle & Hillmann, 2007; JSC, 2014; Lazarinis & Fotis Lazarinis, 2015) desarrollado por el RDA *Steering Committee*[6], más reciente y basado en los modelos conceptuales de organización de la información bibliográfico de la IFLA (BNE, 2014). La BNE[7], por ejemplo, utiliza ISBD, aunque recientemente se ha comprometido, durante los años 2017 y 2018, a crear registros bibliográficos también en RDA para catalogar sus materiales (BNE, 2016).

Además de los estándares de registro bibliográfico ISBD o RDA, existen también propuestas normalizadas para codificar los registros en formato electrónico. Entre las más utilizadas destacan, el formato MARC[8] (Lazarinis & Fotis Lazarinis, 2015; LC & NDMSO, 1999), el nuevo formato BIBFRAME[9] que pretende sustituir a MARC (Kroeger, 2013) y el formato Dublin Core (D.-L. working group DCMI, 2000). Otros formatos pueden consultarse en (Chan, 2007). Estos formatos, básicamente, utilizan etiquetas para denotar los diferentes campos (Figura 2). Por ejemplo, en la Figura 2 la etiqueta "*490 0"* denota el campo título, y el "*700"* los autores. Su objetivo es que las descripciones permitan un tratamiento automático lo más eficiente posible, tanto por las aplicaciones de gestión bibliográfica, como por las bases de datos bibliográficas, y por los motores de búsqueda de la web. Los estándares MARC y Dublin Core se presentarán con más detalle en la sección 2.2 al ser formatos relevantes para este trabajo de tesis.

[3] International Standard Bibliographic Description

[4] International Federation of Library Associations and Institutions (http://www.ifla.org/)

[5] Resource Description and Access

[6] Anteriormente llamado Joint Steering Committee (JSC) (http://www.rda-rsc.org/)

[7] Biblioteca Nacional de España (http://www.bne.es)

[8] Machine Readable Cataloging (https://www.loc.gov/marc/)

[9] Bibliographic Framework Initiative (http://www.loc.gov/bibframe/)

Respecto a la catalogación, ésta puede ser descriptiva o temática (Figura 3). En realidad, cada uno de estos tipos son procesos de catalogación diferentes que responden a objetivos también diferentes (Martínez de Sousa, 2004):

- La *catalogación descriptiva* es la fase del proceso de catalogación que añade al registro bibliográfico la identificación, descripción y localización del recurso. Toda esta información constituye un ejemplo palpable de lo que, en el contexto de los modernos sistemas de información, se denominan *metadatos*. La catalogación descriptiva tiene como objetivo facilitar la identificación de los recursos en las colecciones cuando el usuario conoce algún rasgo del recurso.

```
LEADER 00000nam a2200061 i 4500
008    010620s2000    mx a          100 0 spa c
020    968-36-8130-1
035    (OCoLC)760407739
040    SpMaUCBY
080    025.315:004(083.74)
080    025:004.738.5(063)
080    004.73:027
080    027:004.73
099 1  UCM2OCLC|b20150601
245 00 Internet, metadatos y acceso a la información en
       bibliotecas y redes en la era electrónica /|cCompiladores
       Filiberto Felipe Martínez Arellano, Lina Escalona Ríos ;
       Traducción de los documentos en inglés Filiberto F.
       Martínez Arellano
250    1ª ed.
260    México :|bUniversidad Nacional Autónoma de México,|c2000
300    X, 112 p. :|bil. ;|c24 cm.
490 0  Sistematización de la información documental;|v1
505    Los trabajos presentados en el XVI Coloquio Internacional
       de Investigación
650 04 Catalogación bibliográfica|xAutomatización|xNormas
650 04 Internet|xAplicaciones en bibliotecas|xCongresos
650 04 Metadatos
650 04 Redes de información bibliotecaria
700 1  Martínez Arellano, Filiberto Felipe,|eed. lit.
700 1  Escalona Ríos, Lina,|eed. lit.
907 00 vp|b0
961 00 ae|b0|c20020116
```

Figura 2. Registro bibliográfico ISBD en formato MARC

- La *catalogación temática* consiste en describir el contenido temático (o tópicos) del recurso utilizando el léxico de un *vocabulario* de referencia. La catalogación temática tiene como objetivo la clasificación conceptual, por temas o tópicos, de los recursos y constituye uno de los puntos de acceso al recurso que es útil cuando el usuario quiere, o bien explorar los recursos sobre un tema, o bien, cuando necesita buscar un recurso del que no conoce ningún rasgo descriptivo o identificativo, pero sí el contenido. Para ello, durante esta catalogación se añaden un conjunto de términos del vocabulario de referencia al registro bibliográfico del recurso.

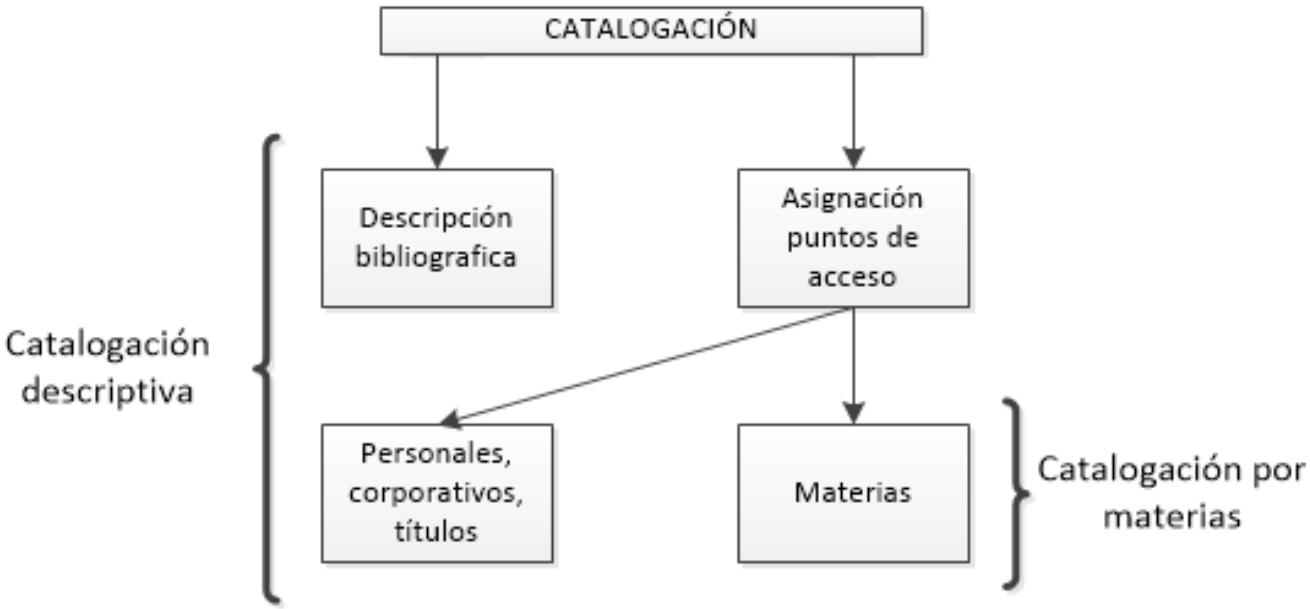


Figura 3. Partes de la catalogación[10]

En lo que se refiere a los vocabularios utilizados durante la catalogación temática, dichos vocabularios son normalmente, *controlados*, lo que significa que se ha seleccionado previamente el léxico (i.e. conjunto de palabras) de forma que cada palabra, en este caso denominada *término*, está normalizada y designa de forma no ambigua y precisa a un único concepto, tema o tópico de la colección de recursos. Ejemplos de vocabularios controlados son los tesauros, taxonomías y clasificaciones y, un tipo de ontologías denominadas ontologías terminológicas. Estos tipos de vocabularios se revisarán en la sección 2.3 . Los términos del vocabulario se utilizan en el registro bibliográfico como valores de los campos denominados "encabezamientos de materia". Los encabezamientos de materia pueden tener una organización jerárquica, distinguiendo entre materias principales y secundarias. En la Figura 1, por ejemplo, se puede observar un encabezamiento primario, el campo "*Encabezamiento de materia*" con el valor "*Astronomía*".

Asimismo, conviene tener en cuenta que en el contexto bibliográfico y documental, es común el uso del término *lenguaje documental* para denotar a los vocabularios controlados (Rodríguez Bravo, 2011). Como en el caso de la catalogación descriptiva, en la catalogación temática se recomienda el uso de lenguajes documentales o vocabularios controlados de

[10] Fuente: (Garrido Arilla, 1996)

referencia como la LCC[11], el DDC22[12], o la CDU[13] o de vocabularios controlados estándares como ISO 472:2013 Plastics – Vocabulary (ISO, 2013). Además, para el uso de estos vocabularios los estándares ISBD y RDA incluyen reglas para la catalogación temática de los recursos. La BNE, por ejemplo, utiliza para la catalogación y clasificación temática de sus recursos los *Encabezamientos de Materia de la Biblioteca Nacional de España (EMBNE)* (BNE, 2013) y la *Clasificación Decimal Universal* en su edición de 2015 (BNE, 2015).

Actualmente, además del uso de vocabularios controlados en la clasificación temática se utilizan vocabularios no controlados (Rodríguez Bravo, 2011). Los vocabularios no controlados están formados por léxico del lenguaje natural que puede no estar normalizado e incluso ser polisémico. Normalmente, se generan de forma inductiva y colaborativa cuando se necesita describir el contenido de los recursos de una colección conforme se van creando. Dos ejemplos conocidos son los vocabularios no controlados de los ámbitos de, (i) las publicaciones académicas en línea, cuando se permite que los autores libremente elijan y asignen palabras clave a sus producciones y, (ii) la web social, mediante la creación o selección y utilización de etiquetas en lenguaje natural por los internautas para designar a los recursos que publican y comparten en la web. Estos tipos de vocabularios, denominados *folksonomías*, se revisarán también con más detalle en la sección 2.3 .

En resumen, es en el área de Biblioteconomía y Documentación donde, originalmente, se genera y se aportan soluciones a la necesidad de disponer de herramientas fiables para la representación, organización y recuperación de colecciones de los recursos preservados y gestionados por las bibliotecas, museos o centros documentales. Estas herramientas se denominan *catálogos* y, básicamente, están formados por registros bibliográficos que representan e identifican de forma única a los recursos y a las colecciones de recursos. Un registro está formado por una serie de pares campo y valor que pueden estar organizados jerárquicamente y que cumplen la misión de identificar, describir, indexar y localizar cada recurso. En base a esta información los catálogos se configuran de dos formas: *descriptiva* y *temática*. La primera representa, organiza y recupera los recursos a partir de sus propiedades, mientras que la segunda lo hace a partir de su contenido. Ambas formas se complementan para

[11] Library of Congress Classification (https://www.loc.gov/catdir/cpso/lcc.html)

[12] Dewey Decimal System (https://www.oclc.org/en/dewey/features/summaries.html)

[13] Clasificación Decimal Universal (http://www.udcc.org/index.php/site/page?view=about)

asegurar la mayor accesibilidad posible a los recursos: se puede acceder a los recursos conociendo alguna de sus propiedades o su contenido. Los catálogos se construyen de forma normalizada, utilizando reglas y esquemas de representación consensuados internacionalmente. El fin es poder intercambiar recursos e información sobre los recursos y facilitar, a los usuarios, el acceso a los recursos mediante mecanismos universales. El problema es que, a pesar de los esfuerzos por la normalización, no existe un acuerdo común sobre los modelos de representación, organización y, en consecuencia, de recuperación de recursos bibliográficos y documentales. Esto deriva en que, actualmente, se están utilizando varios esquemas de registro bibliográfico descriptivo y temático que limitan y encarecen el intercambio y reutilización de recursos e información sobre los recursos. El problema se agrava, además, con los continuos avances tecnológicos que obligan a realizar continuas actualizaciones en los esquemas de catalogación en poco tiempo para encontrar una forma de representación común y consensuada. Este problema del ámbito bibliográfico y documental, se reproduce de forma todavía más patente en el contexto de los sistemas de información distribuidos en la web, generando importantes dificultades para la reutilización, mantenimiento y migración de las colecciones de recursos. Su resolución constituye uno de los principales objetivos y motivaciones de este trabajo de tesis doctoral.

En las siguientes secciones y a lo largo de la tesis se continuará utilizando los términos *catálogo* y *esquema de catalogación* para denotar, como en el área de Bibliografía y Documentación, las herramientas y mecanismos de representación, identificación, localización, organización y recuperación de los recursos digitales en el ámbito global de la web. También se seguirán utilizando los dos tipos de esquemas de catalogación descritos, los esquemas descriptivos que, en la web se denominan *metadatos*, y los esquemas de catalogación temáticos que, de forma general, se denominan *vocabularios*. Finalmente, dejaremos de usar el término *recurso* en su sentido bibliográfico para usar el término *objeto digital* que es un tipo de recurso digital con unas características técnicas bien definidas.

2.1.2 *De la catalogación de colecciones de recursos a la catalogación de colecciones de objetos digitales*

La incorporación de internet en la sociedad, en la década de 2000, incluyó un nuevo contexto de difusión y acceso a los recursos digitales para el cual no habían sido previstos y no son totalmente aplicables los sistemas de catalogación tradicionales. Básicamente, los recursos

que se gestionaban internamente en cada centro bibliográfico o documental tenían un formato de contenido determinado (p.e., texto, imagen, programa, objeto, etc.) claramente definido por los estándares bibliográficos (p.e., ISBD (IFLA & BNE, 2014)) y un identificador y una localización física o digital también claramente definidos y estables en sus registros bibliográficos. Con la incorporación de internet como espacio de producción, difusión y compartición de los recursos aparece el problema del mantenimiento y validación de catálogos de recursos (IFLA & BNE, 2014; Rodríguez Bravo, 2011). Este problema se genera, (i) por una parte, por la continua evolución de los recursos que cambian en formatos, estructura o localización –incluyendo la desaparición de los mismos- y esto en relativamente cortos periodos de tiempo, y (ii) por otra parte, por la creciente descentralización y heterogeneidad de las colecciones de recursos web.

En este nuevo contexto, los conceptos de recursos y los sistemas de almacenamiento y gestión de colecciones de recursos han evolucionado hacia los conceptos de objetos digitales y repositorios digitales (Gerolimos, Papadourakis, Nikitakis, & Sitas, 2011; Kahn & Wilensky, 2006; Pal & others, 2016; Park & Tosaka, 2010). Sin embargo, a nivel lógico, los conceptos de vocabulario (controlado y no controlado), catálogo y esquema de catalogación no han cambiado significativamente, aunque se han aplicado nuevas soluciones tecnológicas para su implementación como son los lenguajes y esquemas XML o RDF.

2.1.3 *Objeto digital y colecciones de objetos digitales*

El cambio de recurso (bibliográfico) a objeto digital viene motivado por dos factores: (i) la necesidad de tener en cuenta la continua aparición de nuevos formatos digitales de los recursos digitales con contenido informacional (BNE, 2011b; Rodríguez, 2007) y, (ii), la necesidad de utilizar sistemas de identificación y localización universales en la web que sean independientes del tiempo y la localización física del recurso (Kahn & Wilensky, 2006).

Un objeto digital es, desde el punto de vista tecnológico, una instancia de un tipo abstracto de datos formado por dos componentes básicos: datos y metadatos (Kahn & Wilensky, 2006). Normalmente, los metadatos son, al igual que los registros bibliográficos, una serie de pares campo-valor que tienen como fin identificar, describir y localizar los objetos digitales en la web. Incluyen un identificador universal único y persistente para cada objeto digital. Este identificador es un *string* que ha sido generado por lo que se denomina un *sistema autorizado de registro*. Está formado por dos partes separadas, la primera identifica al productor del objeto digital (i.e., universidad, editorial o museo) y la segunda identifica al

objeto. Los identificadores persistentes aseguran que el objeto será localizado en internet aunque cambie de ubicación (URL). Para asignar un identificador a un objeto digital, el productor del objeto digital solicita a un sistema autorizado de registro un identificador o propone uno, en cuyo caso el sistema comprueba la unicidad del mismo antes de confirmar el registro. Sistemas de registros autorizados ampliamente utilizados son el Sistema Handle[14] o el Sistema DOI[15] (Figura 4). Por su parte, los datos del objeto digital constituyen su contenido informacional. Estos datos pueden ser simples o compuestos: texto, imagen, archivos, urls, o incluso otros objetos digitales. Esto implica que un objeto digital, puede, categorizarse como *simple* (elemental) o *compuesto*. Los objetos simples son aquellos en donde entre sus datos no se encuentran otros objetos digitales. Los objetos compuestos en cambio englobarán como parte, o todo, en sus datos a otros objetos, habitualmente en forma de referencia.

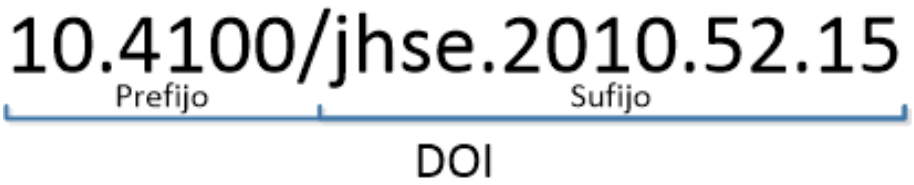


Figura 4. Identificador DOI[16]

Una colección de objetos digitales comparte un mismo modelo (o modelos) de metadatos. Al igual que ocurre con los recursos bibliográficos, es posible utilizar modelos de metadatos estándares para facilitar la recuperación y reutilización de los objetos y colecciones. Además, en el entorno web, el uso de metadatos estándares facilita la preservación de las colecciones (Caplan, 2010; Caplan, Kehoe, & Pawletko, 2010), facilitando especialmente las migraciones a nuevos sistemas de almacenamiento y gestión (Nilsson, Baker, & Johnston, 2008; Tarrant et al., 2009; Ternier et al., 2008). Existen múltiples estándares incluso dentro de un mismo ámbito de conocimiento lo que, como en el contexto Bibliográfico, limita las posibilidades de reutilización de colecciones de objetos digitales y la interoperabilidad entre repositorios. Entre las propuestas de estándares de metadatos destacan Dublin Core[17](D. C. M. I. DCMI, 2012; Mahdi Taheri & Hariri, 2012; Nilsson et al., 2008), por ser un estándar sencillo,

[14] Sistema Handle (https://www.handle.net/)

[15] Digital Object Identifier (http://www.doi.org/)

[16] Fuente: https://biblioteca.ua.es/es/propiedad-intelectual/imagenes/pi/doi.gif

[17] The Dublin Core Metadata Initiative (http://dublincore.org/)

de dominio general y, uno de los más utilizados por los repositorios digitales (Park & Tosaka, 2010) . En el ámbito educativo destaca el modelo de metadatos IEEE-LOM (LOM-ES[18] en España). En Biblioteconomía y Ciencias de la Documentación destacan, por ser los más usados, los modelos METS[19], MARC-21 y MODS[20]. Otros formatos de metadatos estándar de diferentes dominios son Darwing Core[21] para biología, ISO19115 [22]para datos geográficos o CDWA[23] para arte y arquitectura. Finalmente, es interesante resaltar que, a pesar de no existir propuestas estándares respecto a cómo organizar e implementar los datos de un objeto digital por todas las posibilidades que podrían incluirse, sí existe en el contexto educativo propuestas de empaquetamiento de objetos digitales (i.e. *objetos de aprendizaje* en este contexto) que, como por ejemplo SCORM[24], facilitan el intercambio e integración de colecciones de objetos digitales en diferentes repositorios o sistemas de publicación.

2.1.4 *Los repositorios digitales*

Un repositorio digital (en adelante *repositorio*) es un sistema de almacenamiento y recuperación de objetos digitales en línea (o en redes distribuidas privadas) (Kahn & Wilensky, 2006). Los repositorios disponen, como mínimo, de mecanismos para añadir nuevos objetos a la colección (depósito) y para acceder a ellos (acceso). El mecanismo de depósito de objetos puede tener diferentes formas de trabajar: (i) crear y almacenar un objeto digital a partir de los datos, el identificador persistente y, opcionalmente, los metadatos; (ii) crear y almacenar un objeto a partir de los datos y, opcionalmente de los metadatos; o, (iii), solicitar los datos y metadatos, excepto el identificador, generar una petición de identificador a un Servicio de Registro Autorizado y cuando lo recibe, crear y almacenar el objeto. El mecanismo de acceso se denota con el acrónimo inglés *RAP*, Protocolo de Acceso al Repositorio (*Repository Access Protocol*), y define las operaciones para acceder tanto a los metadatos como a los datos del objeto. Las operaciones básicas de un RAP según (Park & Tosaka, 2010) son: (i) un mecanismo

[18] Perfil de Aplicación LOM-ES V1.0 (http://educalab.es/recursos/lom-es)

[19] Metadata Encoding and Transition Standard (http://www.loc.gov/standards/mets/)

[20] Metadata Object Description Schema (http://www.loc.gov/standards/mods/)

[21] Darwin Core (http://rs.tdwg.org/dwc/index.htm)

[22] ISO 19115:2003 (http://www.iso.org/iso/catalogue_detail?csnumber=26020)

[23] Categories for the Description of Works of Art (http://www.getty.edu/research/publications/electronic_publications/cdwa/)

[24] Sharable Content Object Reference Model (http://www.adlnet.gov/adl-research/scorm/)

de acceso a los documentos a través de su identificador, (ii) un mecanismo de almacenamiento de los objetos digitales y su descripción de manera permanente y (iii) un mecanismo de acceso a los objetos digitales de modo que, a través de los valores de descripción de los objetos, se pueda dar acceso a los objetos que contienen en su descripción esos valores (Kahn & Wilensky, 2006).

Los repositorios se identifican de forma oficial con nombre únicos asignados y aprobados por lo que se denomina Autoridad de nombres globales o locales (*global naming authority* o *local naming authorities)*. En los repositorios españoles, para la creación de nombres de repositorios con dominio español, la autoridad pública local que lo gestiona es la Entidad Pública Empresarial Red.es[25]. Dentro de cada dominio, el dueño del dominio será el encargado de dar la autorización para el uso de nombres de subdominio dentro de su sistema. A nivel internacional la entidad proveedora hasta 1998 de nombres y números en Internet era IANA[26]; actualmente esta actividad está desarrollada por la fundación sin ánimo de lucro ICANN[27].

La gestión de un repositorio de objetos digitales se puede considerar dividida en dos partes principales: (i) la gestión de los objetos y sus descriptores y, (ii), la gestión del acceso a los objetos (Bluhm, Getting, Hayft, & Walz, 2006). Los mecanismos de gestión de los objetos digitales y sus descriptores deben permitir al menos: la edición, borrado y creación de los objetos de una colección y de la estructura que los organiza dentro de la colección. Los mecanismos de gestión del acceso a objetos digitales pueden permitir el control de visibilidad de objetos y ciertos campos en función de diferentes roles que se pueden gestionar desde la aplicación (Fernández-Valmayor Crespo, Fernández-Pampillón Cesteros, & Varadero Software Factory, 2013). Estos sistemas de gestión del acceso también pueden contener mecanismos de control de la indexación de los objetos y sus descriptores en la colección. Estos índices de acceso a través de identificadores y descriptores de objetos digitales permiten implementar los servicios de acceso a los objetos digitales descritos en las operaciones básicas RAP (Kahn & Wilensky, 2006).

[25] Entidad Pública Empresarial Red.es (http://www.red.es/redes/)

[26] Internet Assigned Numbers Authority (http://www.iana.org/)

[27] Internet Corporation for Assigned Names and Numbers (https://www.icann.org/)

Estado de la cuestión

Para la recuperación de los objetos digitales (Baeza-Yates, Ribeiro-Neto, & others, 1999; Chowdhury, 2010; Salvador Oliván, 2008) el sistema puede generar estructuras o índices (claves de asignación directa a los objetos digitales a través de un elemento descriptor). Uno de estos campos candidato a indexar son los identificadores únicos que deben aparecer para cada objeto digital de la colección. Dependiendo del sistema, pueden proporcionarse índices extra sobre los elementos descriptivos de los objetos que permitan la rápida recuperación de los objetos para los procesos de navegación y búsqueda por valores controlados. Debido a ello, las diferentes aplicaciones que gestionan cada colección pueden contener sistemas de configuración de estos índices de manera que el experto pueda configurarlos para ciertos valores, o conjuntos de los mismos. Además, estos índices pueden implementarse internamente de múltiples maneras (Ambroziak, 2002; Cauffman, Thompson, & Cauffman, 1994). Algunos ejemplos son el uso de índices invertidos, *clustering*, tablas de búsquedas frecuentes, etc.

Con el fin de poder interconectar repositorios para mostrar, o compartir, o comparar, información, los repositorios hacen uso de modelos y servicios estándar de importación, exportación y visualización de datos. Estos servicios de importación y exportación permiten simplificar los procesos de creación de nuevos repositorios y la incorporación de nuevos objetos de manera masiva. Estas herramientas son muy útiles al poder aprovechar los objetos y su descripción expuesta en otros repositorios. También permiten dar visibilidad a los objetos digitales al poder exportarlos a otros sistemas de visualización unificada de repositorios sobre diferentes plataformas. Debido a la necesidad de interconexión de repositorios (Bainbridge, Ke, & Witten, 2006; Caplan et al., 2010; Jones, 2007; Lagoze, Lynch, Waters, Van De Sompel, & Hey, 2006; Lynch, Parastatidis, Jacobs, Van de Sompel, & Lagoze, 2007; Tarrant et al., 2009; Ternier et al., 2008; Van de Sompel et al., 2006), actualmente existen múltiples modelos estándar de interconexión, tanto generalistas como específicos. Uno de los más importantes es el protocolo OAI-PMH, sistema desarrollado por OAI[28] donde uno de los modelos de catalogación de la información que soporta es Dublin Core (Nilsson et al., 2008; Suleman & Edward, 2002).

A continuación, se revisan brevemente alguno de los sistemas de gestión de repositorios digitales más empleados en la actualidad.

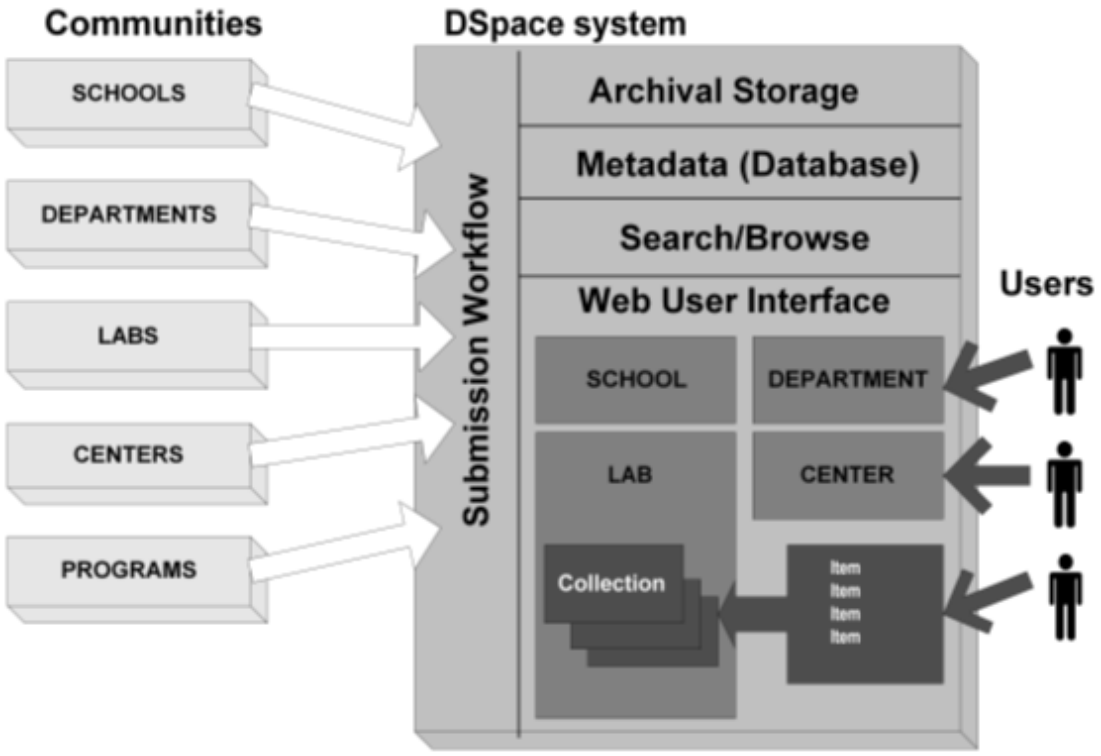
[28] Open Archives Initiative Protocol for Metadata Harvesting (https://www.openarchives.org/pmh/)

2.1.4.1 Dspace

El repositorio *DSpace*[29] (M. Smith et al., 2003) ha sido desarrollado por la empresa Duraspace[30] . Según la web OpenDOAR, actualmente es el sistema de almacenamiento de repositorios digitales más usado en el mercado[31].

DSpace permite la existencia de múltiples modelos de almacenamiento de objetos digitales y ofrece diferentes capas de servicios que dan soporte a los procesos de recuperación tales como la búsqueda y filtrado. De esta forma, se pueden mantener en un mismo repositorio colecciones heterogéneas de objetos digitales En la Figura 5 se muestra esquemáticamente las capas de servicios y el esquema general del sistema.

a)



[29] DSpace homepage (http://www.dspace.org)

[30] DuraSpace homepage (http://www.duraspace.org)

[31] Opendoar - Usage of Open Access Repository Software – Worldwide (http://www.opendoar.org/onechart.php?cID=&ctID=&rtID=&clID=&lID=&potID=&rSoftWareName=&search=&groupby=r.rSoftWareName&orderby=Tally%20DESC&charttype=pie&width=600&height=300&caption=Usage%20of%20Open%20Access%20Repository%20Software%20-%20Worldwide)

b)

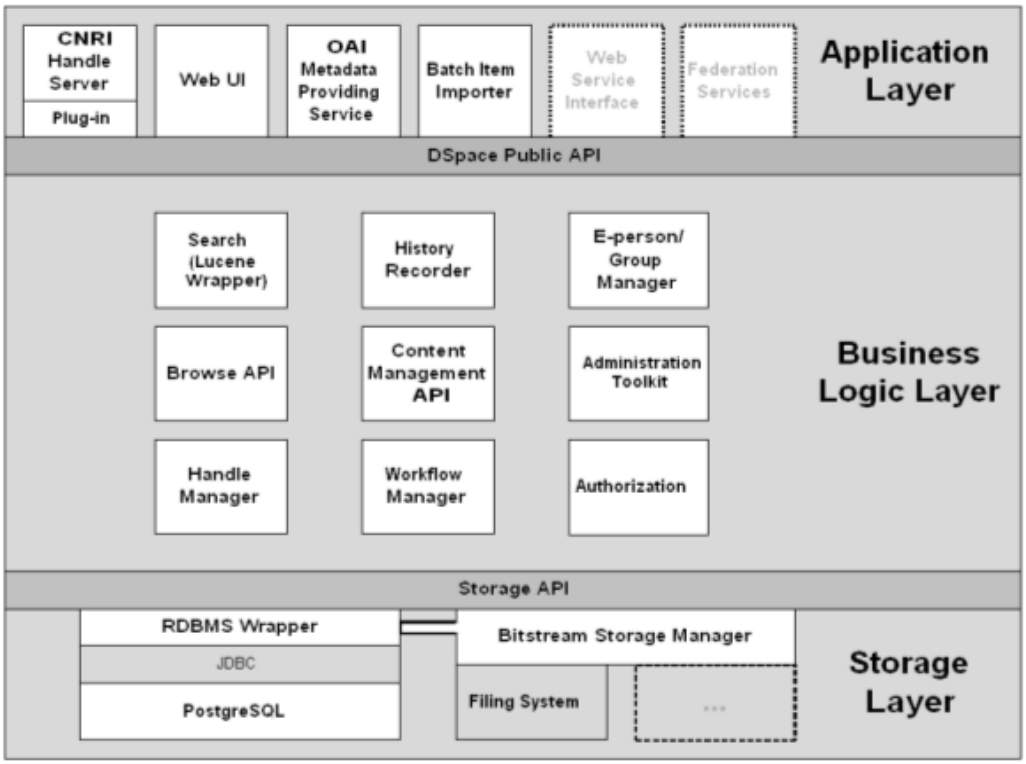


Figura 5. a) Representa el esquema general de DSpace, b) Modelo por capas de DSpace[32]

Por defecto, *DSpace* permite utilizar los esquemas de metadatos estándares de Dublin Core y IEEE-LOM. Así mismo, también permite definir esquemas de metadatos propios expandiendo un esquema base que proporciona *DSpace* con ciertos campos obligatorios. Los vocabularios controlados utilizados en los atributos de los esquemas de metadatos están definidos en el caso de los esquemas de metadatos estándar, y son descritos en un archivo de configuración en el caso de los esquemas de metadatos definidos por el usuario.

2.1.4.2 Fedora

Fedora[33] (Staples, Wayland, & Payette, 2003) es una evolución de *DSpace*, desarrollada por *Duraspace*. La principal diferencia entre *DSpace y Fedora* es la utilización de tripletas relacionales (al estilo RDF) para definir los objetos digitales, donde las descripciones son relaciones con otros objetos más sencillos. Al igual que *DSpace*, permite la definición de

[32] Fuente: (M. Smith et al., 2003)

[33] Fedora homepage (http://fedorarepository.org/)

colecciones sobre múltiples esquemas, y facilita servicios estándar de recuperación y publicación de información. En la Figura 6 se muestra un esquema de la arquitectura de *Fedora*.

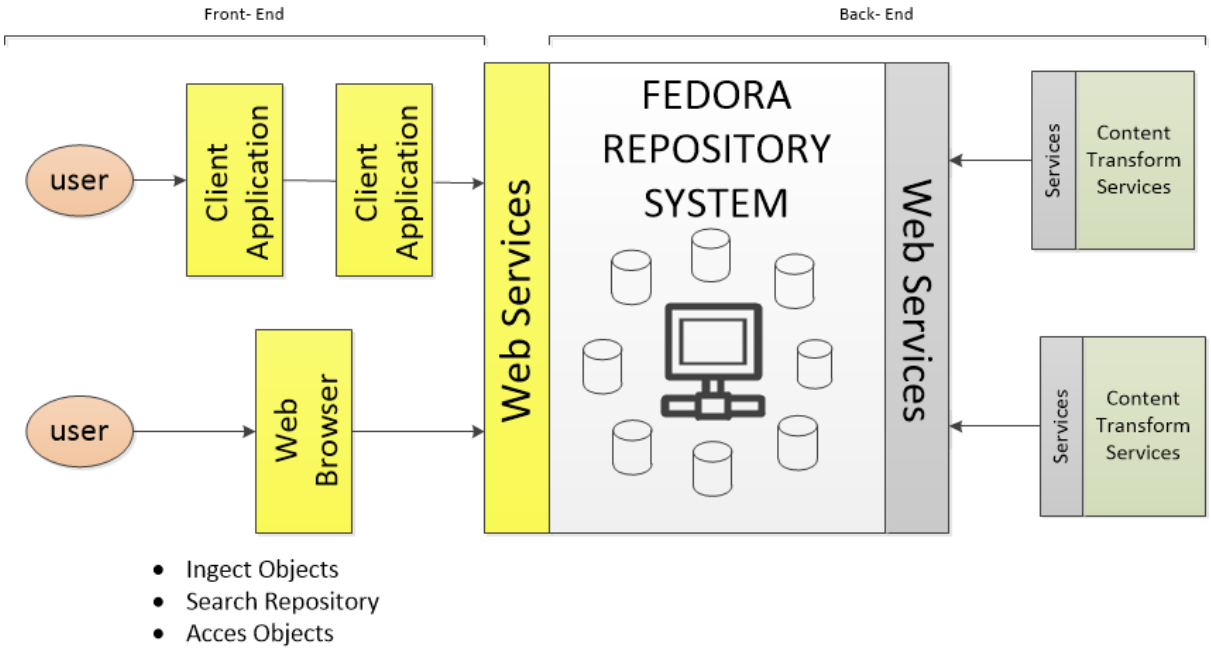


Figura 6. Diagrama simplificado de interconexiones y servicios del repositorio Fedora[34]

En *Fedora*, al principio del proceso de creación de una colección, se debe definir el modelo del esquema de catalogación para la colección. Posteriormente, se pueden crear nuevos modelos diferentes sobre la colección como si se crearan de manera independiente quedando vinculados a la colección inicial.

Cuando se importan colecciones, éstas se vinculan a una colección existente en caso de que el esquema de metadatos sea diferente a los registrados, pero si el sistema detecta que el esquema de la colección importada es compatible con alguno de los registrados, entonces se tratará de realizar una transformación automática (de tipo traducción) para no duplicar colecciones en el sistema. *Fedora* permite la importación de colecciones con esquemas estándar (Dublin Core e IEEE-LOM fundamentalmente), utilizando el protocolo estándar de interconexión OAI-PMH.

Uno de los módulos de interconexión más útiles es el módulo *Islandora* que permite la conexión entre *Fedora* y el CMS (*Content Manager System*) Drupal[35] para editar objetos

[34] Fuente: (DuraSpace, 2008)

[35] Drupal CMS (https://www.drupal.org/)

digitales de y para *Fedora* siempre que utilicen el estándar Dublin Core (el estándar MODS está actualmente en desarrollo). Un ejemplo de la ejecución de *Fedora* utilizando el módulo *Islandora* puede verse en la Figura 7.

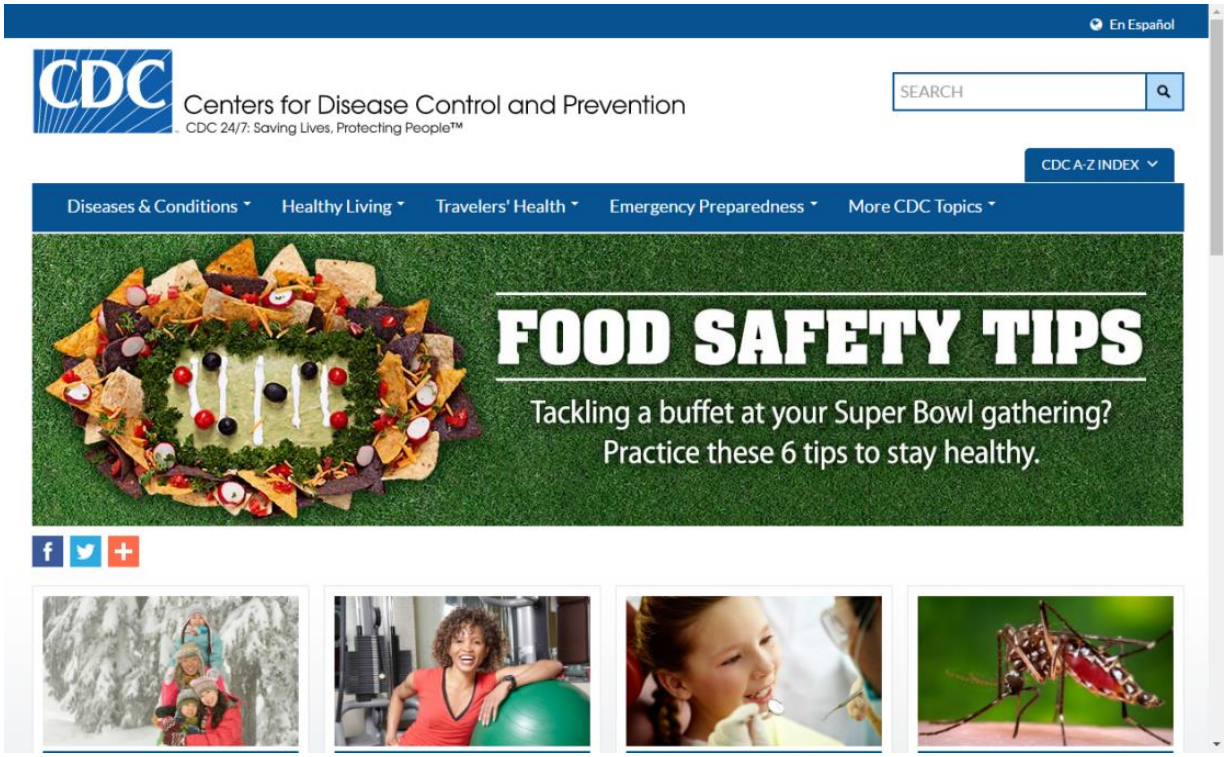


Figura 7. Ejemplo de repositorio sobre Drupal con información gestionada en Fedora usando para la conexión el módulo Islandora[36].

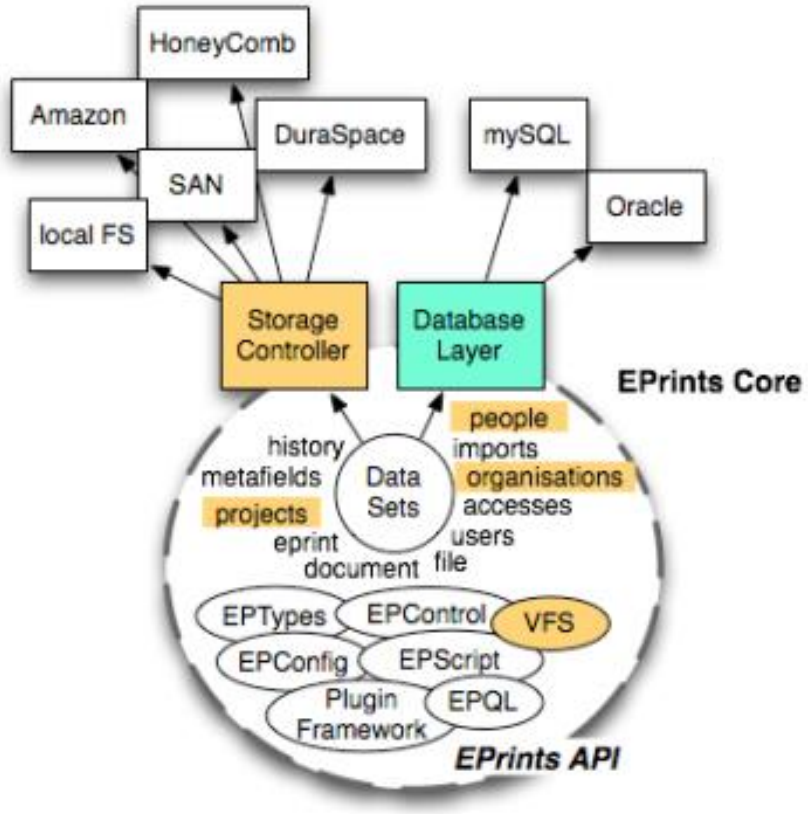
2.1.4.3 Eprints

El repositoiro *EPrints*[37] es una iniciativa gratuita que se desarrolló a través del consorcio OAI-PMH y que permite la creación de repositorios digitales en la web. Según la web *OpenDOAR*, este sistema ocupa la segunda posición entre los más elegidos a la hora de la creación y gestión de repositorios de objetos digitales principalmente para documentos. Uno de los aspectos más característicos de *EPrints* es el módulo de estadísticas que implementa y que proporciona datos de uso de los diferentes objetos que contiene la colección. Durante la instalación del repositorio *EPrints*, se puede especificar un modelo de colección de objetos digitales a través de una plantilla ofrecida por el sistema. A partir de esta plantilla se generará una colección con un sistema de almacenamiento optimizado sobre el modelo definido en la

[36] Fuente: https://www.cdc.gov/

[37] EPrints Homepage (http://www.eprints.org)

plantilla y que servirá de base a los objetos de la colección. Además de la funcionalidad básica del sistema, existen diferentes módulos que ofrecen funcionalidades de gestión, creación, edición y visualización de la información de la colección. El sistema también ofrece la posibilidad de integrar nuevos módulos programados externamente que posibilitan la importación y exportación de las colecciones dando soporte a modelos específicos de dominios o bien modelos personalizados. La eficiencia y la perdida de datos en los objetos (que se importan o exportan) dependerá de si la estructura de la colección es en mayor o menor medida compatible con el módulo (Proudfoot, 2005). La arquitectura de *EPrints* se muestra en la Figura 8 donde se pueden distinguir los diferentes módulos de servicios sobre el núcleo del sistema.



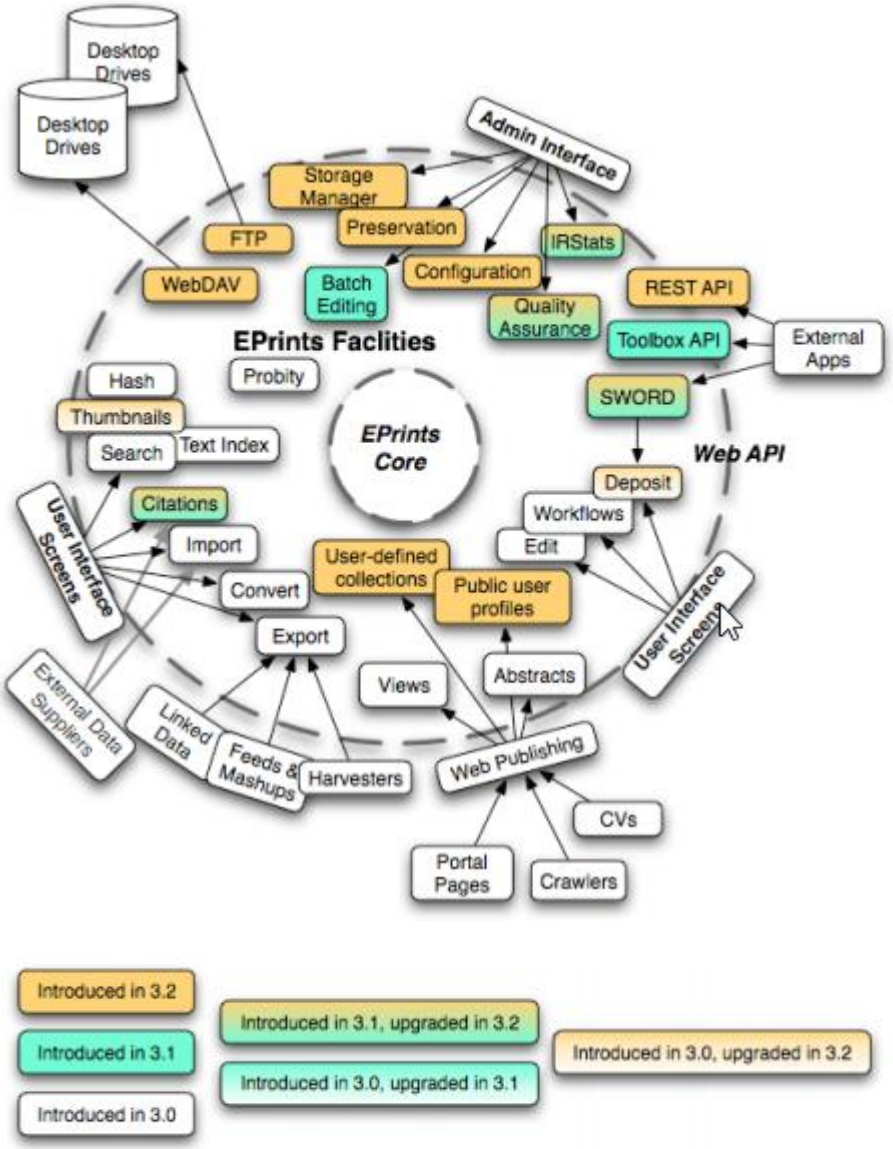


Figura 8. Vista general de EPrints[38]

2.2 La catalogación de objetos digitales mediante metadatos

2.2.1 *Esquemas de metadatos*

Un esquema de metadatos consiste en un conjunto de metadatos estructurados que permiten catalogar diferentes aspectos de un objeto digital. En general, un esquema de metadatos está formado por: (i) una *sintaxis* que establece cómo representar los metadatos usando algún tipo de formato de datos, (ii) una *semántica* que define el significado de cada

[38] Fuente: (Carr, 2009)

metadato, su obligatoriedad o su cardinalidad, y, (iii) unas *reglas de uso* que indican cómo utilizar los metadatos. Los esquemas de metadatos pueden ser de *propósito general* cuando pueden ser usados para catalogar cualquier objeto digital, como por ejemplo Dublin Core *o* METS, o bien de *propósito específico* cuando están orientados para catalogar objetos digitales de un ámbito concreto, como por ejemplo LOM para describir objetos de aprendizaje o MARC-21 para describir datos bibliográficos.

Las siguientes secciones revisan con más detalle los esquemas de metadatos más utilizados para la catalogación de objetos digitales (Park & Tosaka, 2010)

2.2.2 *LOM*

LOM *(Learning Object Metadata)* es una especificación realizada por el grupo de trabajo número 12 del IEEE Learning Technology Standards Committee (IEEE, 2002; Neven & Duval, 2002) que tiene como objetivo definir un conjunto de metadatos para describir objetos de aprendizaje (objetos digitales en el dominio educativo) con la finalidad de facilitar la recuperación de los mismos. Los metadatos se estructuran jerárquicamente en base a nueve categorías principales:

- La categoría *General* agrupa la información general que describe un objeto de aprendizaje en su conjunto.
- La categoría *Ciclo de Vida* describe la historia de un objeto de aprendizaje, qué entidades han intervenido en su creación, etc.
- La categoría *Meta-metadatos* describe metadatos acerca de la propia catalogación tales como la manera en la que la catalogación puede ser identificada, quién la creó, cuándo, etc.
- La categoría *Técnica* describe las características técnicas del objeto de aprendizaje.
- La categoría *Uso Educativo* describe las características educativas y pedagógicas fundamentales del objeto de aprendizaje.
- La categoría *Derechos* describe los derechos de propiedad intelectual y las condiciones de uso de un objeto de aprendizaje.
- La categoría *Relación* describe las relaciones existentes, si las hubiese, entre un objeto de aprendizaje y otros.
- La categoría *Anotación* proporciona comentarios sobre la utilización del objeto de aprendizaje.

- La categoría *Clasificación* describe la clasificación del objeto de aprendizaje respecto a un sistema de clasificación concreto.

Cada categoría está formada por una estructura jerárquica de metadatos, que pueden ser agregados o simples. Los metadatos agregados están formados por otros metadatos, que, a su vez, pueden ser agregados, simples o de ambos tipos. Los metadatos simples contienen directamente valores y constituyen las hojas de la estructura jerárquica. La información está formada por los valores asignados a los metadatos, los cuales pueden proceder de vocabularios controlados o libres. La especificación establece que todos los metadatos son opcionales, y que las instituciones que usen la especificación deberán establecer restricciones de obligatoriedad particulares. En la Figura 9 se muestra una representación de la estructura de LOM.

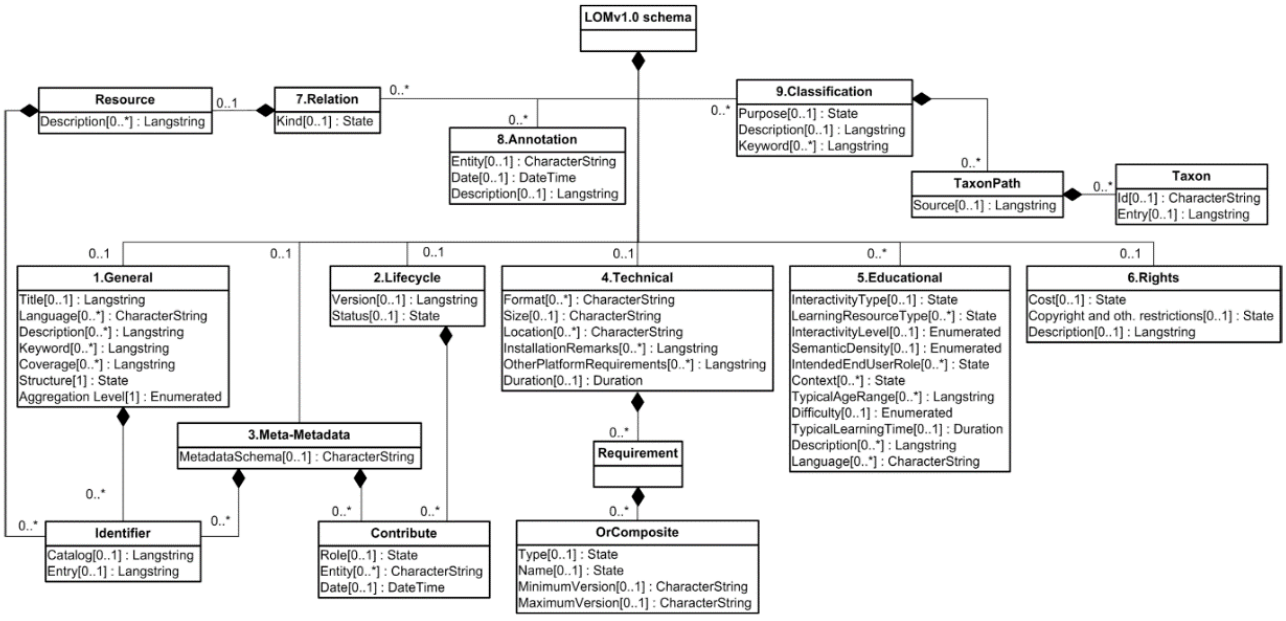


Figura 9. Representación de la estructura de LOM[39]

2.2.3 *Dublin Core*

El DCMI *(Dublin Core Metadata Initiative)* es una especificación que define un conjunto de metadatos plana (D. C. M. I. DCMI, 2012; Méndez, 2006) que tiene como objetivo catalogar objetos digitales de cualquier tipo. Consta de 15 categorías que pueden agruparse, según el tipo de información que almacenan, en:

[39] Fuente: https://es.wikipedia.org/wiki/Learning_Object_Metadata

- Metadatos que describen el contenido del recurso: *Título, Claves, Descripción, Fuente, Tipo de Recurso, Relación*, y *Cobertura*.
- Metadatos que describen las características del recurso en cuanto a propiedad intelectual: *Autor o Creador, Editor, Otros Colaboradores y Derechos.*
- Metadatos referidos a la instanciación del recurso: *Fecha, Formato, Identificador del Recurso y Lengua.*

La especificación promueve el uso de vocabularios controlados para los valores que toman los metadatos. Por otro lado, todos los metadatos son optativos, pueden repetirse y pueden aparecer en cualquier orden. En la Figura 10 se muestra una representación de la estructura de Dublin Core.

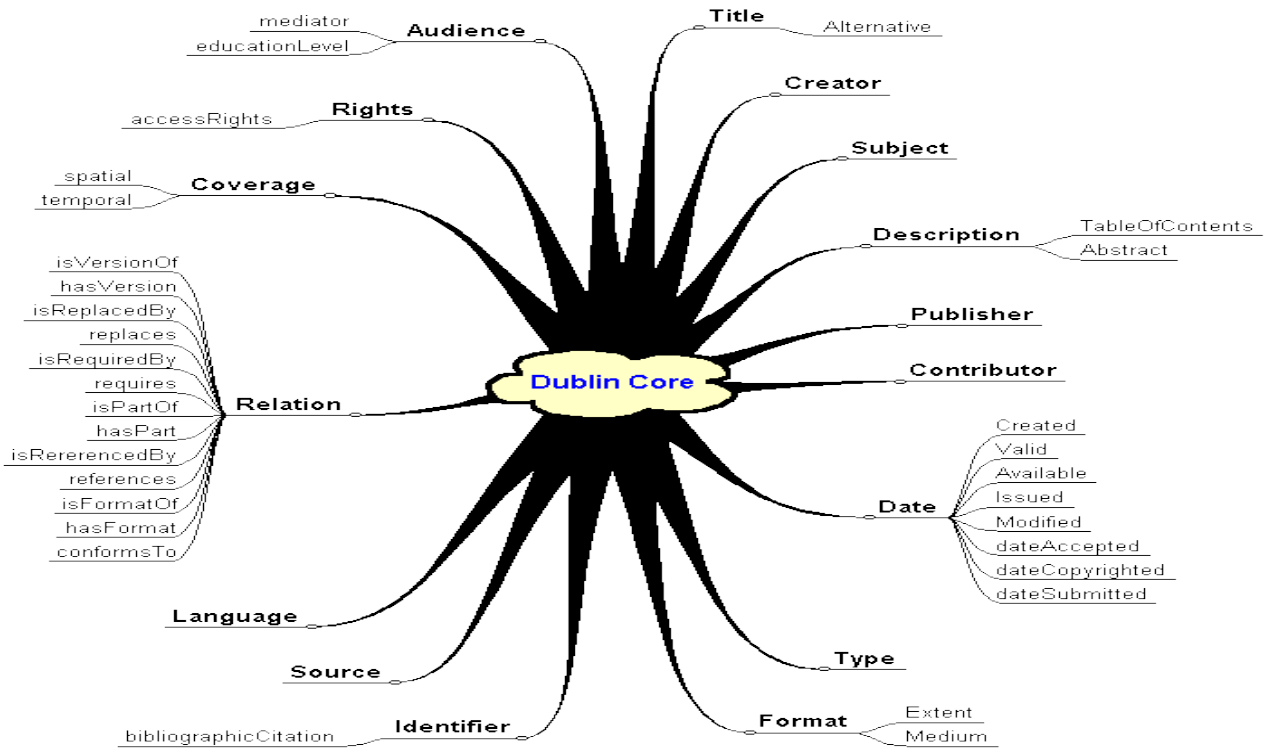


Figura 10. Representación de la estructura de Dublin Core[40]

[40] Fuente: http://ganesha.fr/index.php?post/2008/03/31/Dublin-Core

2.2.4 *MARC*

El estándar MARC-21, *Machine Readable Cataloging 21*, es uno de los modelos estándar más utilizados para catalogación en el ámbito de la bibliografía y la literatura (Fritz & Fritz, 2003; LC & NDMSO, 1999), aunque su estructura soporta otro tipo de obras culturales (p.e., música, el arte, teatro …).

```
<?xml version="1.0" encoding="UTF-8"?>
- <marc:collection xsi:schemaLocation="http://www.loc.gov/MARC21/slim
  http://www.loc.gov/standards/marcxml/schema/MARC21slim.xsd"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xmlns:marc="http://www.loc.gov/MARC21/slim">
    - <marc:record>
        <marc:leader>02631cam a2200433 4500</marc:leader>
        <marc:controlfield tag="001">158709</marc:controlfield>
        <marc:controlfield tag="005">20110119173040.0</marc:controlfield>
        <marc:controlfield tag="008">900925s1619 enk|||| 00| ||eng
            d</marc:controlfield>
      - <marc:datafield tag="035" ind2=" " ind1=" ">
            <marc:subfield code="9">ESTCS101130</marc:subfield>
        </marc:datafield>
      - <marc:datafield tag="035" ind2=" " ind1=" ">
            <marc:subfield code="a">158709</marc:subfield>
        </marc:datafield>
      - <marc:datafield tag="040" ind2=" " ind1=" ">
            <marc:subfield code="a">Cu-RivES</marc:subfield>
            <marc:subfield code="c">Cu-RivES</marc:subfield>
            <marc:subfield code="d">CStRLIN</marc:subfield>
            <marc:subfield code="e">dcrb</marc:subfield>
            <marc:subfield code="d">UtOrBLW</marc:subfield>
        </marc:datafield>
      - <marc:datafield tag="245" ind2="4" ind1="0">
            <marc:subfield code="a">The actes of the ambassage, passed at the
                meating of the lordes and princes of Germany at Naumburg in Thuring,
                concerning the matters there moued by Pope Pius the .iiii. in the yeare
                of our Lorde. 1561. and the fifth daie of February.</marc:subfield>
            <marc:subfield code="b">Item, the aunswere of the same lordes & princes,
                geuen to the Popes Nuntio vpon the eight daye of February. Translated
                out of Dutche into Englishe by R.W.</marc:subfield>
        </marc:datafield>
```

Figura 11. Ejemplo de registro MARC/XML[41]

El modelo MARC-21 tiene una estructura que se divide en cuatro bloques básicos. En conjunto, en todos los bloques se pueden completar más de 800 campos que definirán la descripción del objeto digital de la colección. Los cuatro bloques básicos que componen MARC son:

- *Cabecera*: Conjunto de 24 atributos del registro que definen aspectos tales como la fecha de edición, la fecha de creación, el estado del registro, codificación, la longitud…

[41] Fuente: http://collation.folger.edu/2014/09/folger-tooltips-getting-raw-hamnet-data/

- *Directorio*: Bloque compuesto por varias secciones que definen propiedades referibles desde el resto de los bloques que lo suceden. Normalmente se genera automáticamente al editar o crear el registro por el sistema.
- *Campos variables de control*: Grupo de 8 atributos que identifican el registro MARC-21 y que gestionan características generales de este.
- *Campos variables de datos*: Conjunto mayoritario de datos. Cada dato está dividido en tres partes que definen los atributos en tres niveles de estructura. El primer nivel del dato son 3 dígitos que definen el grupo al que pertenece el mismo, el segundo dos caracteres de control para especificar opciones, y el tercero son pares atributo-valor que definen el objeto dentro del citado grupo.

El modelo MARC-21 puede usarse como medio de entrada y salida de datos. Para ello puede codificarse sobre el formato de archivo XML, dando lugar a la propuesta MARC/XML (Gigee, 2006; Guenther & Radebaugh, 2006; Mahdi Taheri & Hariri, 2012). La Figura 11 proporciona un ejemplo de codificación MARC/XML.

2.3 La catalogación de objetos digitales mediante vocabularios

2.3.1 *Vocabularios*

Los vocabularios son recursos lingüísticos que permiten acceder a los objetos por medio de la palabra o palabras que describen su contenido en los procesos de catalogación temática, tal y como se ha discutido en la sección 2.1 .

En los repositorios digitales, la catalogación de los objetos digitales se realiza de forma semejante a la catalogación bibliográfica, combinando la catalogación descriptiva mediante metadatos con la catalogación temática mediante vocabularios. La idea es describir los objetos digitales utilizando esquemas de metadatos en los que ciertos elementos (i.e. propiedades del objeto digital) toman como valores las palabras de uno o varios vocabularios fijados o recomendados en el esquema de metadatos. Los vocabularios, en este sentido, continúan siendo los catálogos temáticos de representación, organización y recuperación de las colecciones de objetos digitales (Huynh, Mazzocchi, & Karger, 2005). Un ejemplo muy simple es el uso del vocabulario estándar ISO 3166 (ISO 3166/MA, 1997, p. 31) para dar valores al elemento *language* en el esquema de metadatos Dublin Core (Figura 12)

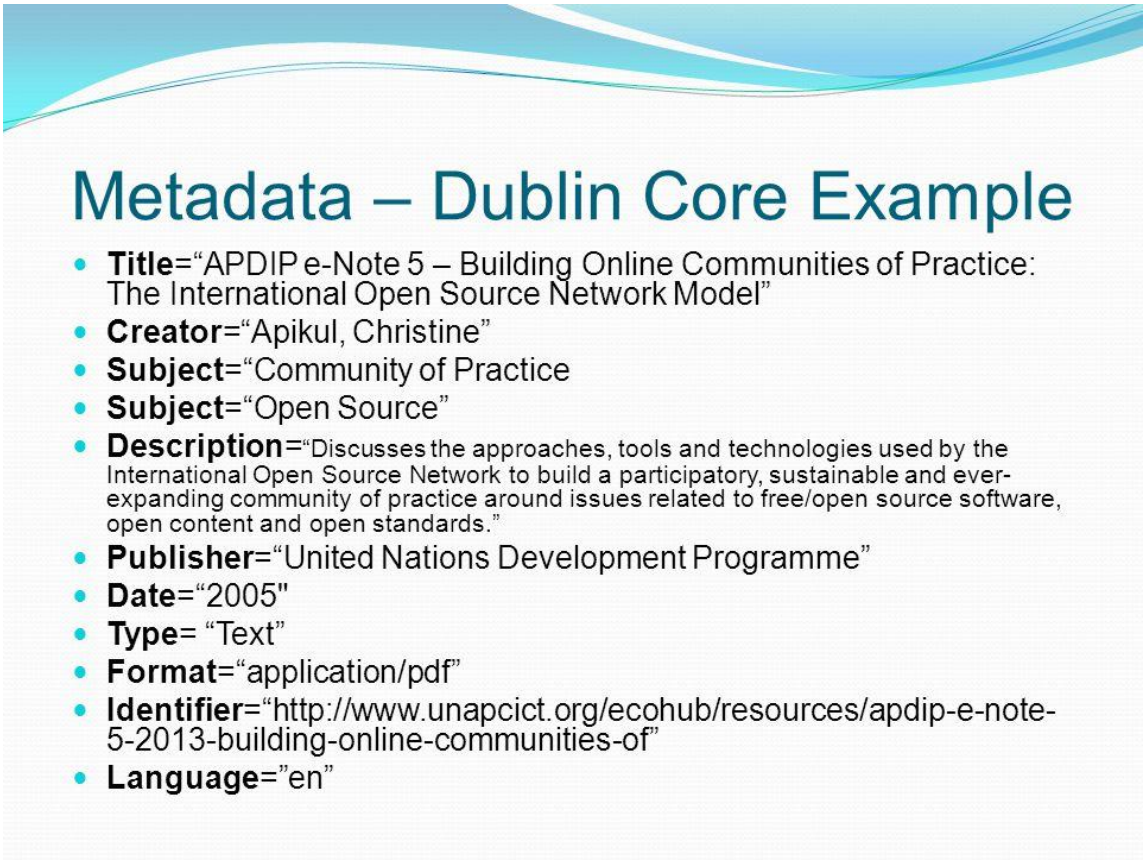


Figura 12. Ejemplo de metadatos Dublin Core en el que se usa el estándar ISO 3166 para dar valor al campo language.[42]

Otro ejemplo, más cercano a la catalogación temática, lo encontramos en el esquema de metadatos de objetos digitales educativos LOM. En este esquema el elemento noveno, "*Clasification*" sirve para catalogar los objetos por su contenido representado mediante un vocabulario de tipo taxonómico que debe identificarse en el subelemento 9.2.1."*Source*" (Figura 9).

[42] Fuente: Christine Apicul http://slideplayer.com/slide/1515416/

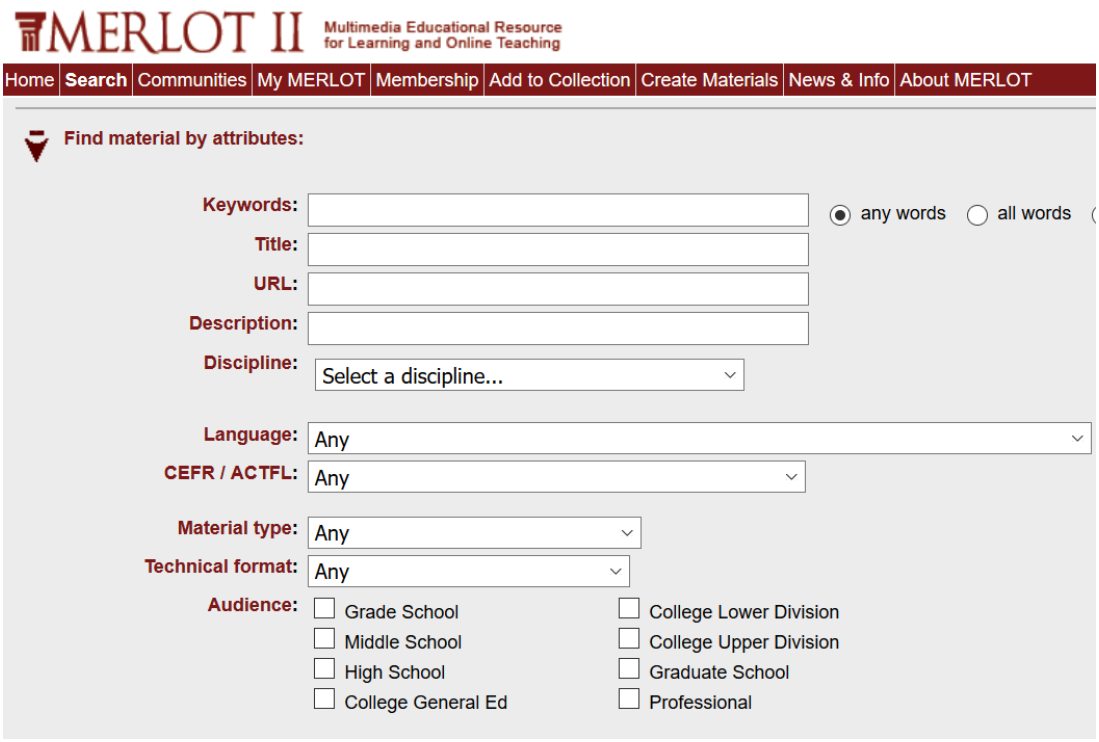


Figura 13. Formulario de búsqueda de recursos basada en las propiedades de LOM del repositorio educativo Merlot[43]

La recuperación de los objetos de un repositorio se lleva a cabo a partir de los elementos y valores de sus metadatos, incluyendo los elementos que sirven para describir el contenido del objeto digital (como, por ejemplo, el elemento *Clasificación* de IEEE LOM). Para ello, lo habitual es utilizar formularios de búsqueda (Figura 13) o menús de navegación (Figura 14). En el primer caso, los formularios muestran los elementos de los metadatos para que los usuarios introduzcan o seleccionen los valores de búsqueda que conozcan. Cuando el formulario permite seleccionar los valores es porque éstos son términos de vocabularios. Como se verá un poco más adelante, los vocabularios pueden ser tan simples como una lista de palabras (el equivalente a un tipo de datos enumerado) o tan complejos como los tesauros y ontologías.

[43] Fuente: https://www.merlot.org/merlot/advSearchMaterials.htm

Figura 14. Vocabulario Materia del Repositorio redined[44]

Los menús de navegación muestran el vocabulario, las palabras y sus relaciones, enlazado con los objetos digitales correspondientes. Las relaciones entre palabras que muestra el vocabulario son normalmente relaciones semánticas, como las de especialización/generalización y parte/todo. En la Figura 15, por ejemplo, se muestra una parte del vocabulario "*Disciplina*" del repositorio de objetos digitales de aprendizaje Merlot, en el que se puede ver cómo se visualiza mediante un sangrado la relación parte/todo entre las palabras "*Services*" y "*Academic suport*". En la Figura 15 se puede ver, también, cómo se visualiza entre paréntesis el número de objetos asociados a cada palabra del vocabulario.

[44] Fuente: http://redined.mecd.gob.es/xmlui/browse?type=subject

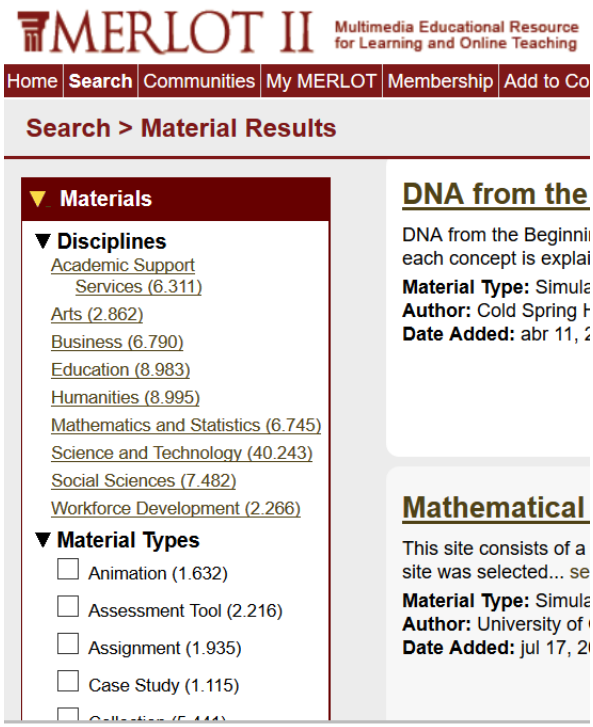


Figura 15. Menú de navegación y búsqueda por metadatos del repositorio internacional Merlot[45]

En lo que resta de esta sección, se revisará cómo son los vocabularios de catalogación de los repositorios digitales y cómo se utilizan no sólo para facilitar el acceso de las personas a los objetos digitales, sino también para facilitar la interoperabilidad entre repositorios.

2.3.2 *Los vocabularios en los repositorios digitales*

Un vocabulario es un conjunto de términos o procedimientos sintácticos convencionales que se utilizan para representar el contenido de un recurso con el fin de permitir su recuperación (Slype, Hípola, & Moya Anegón, 1991). Tal y como ya se ha indicado, en Bibliografía y Documentación se denominan también lenguajes documentales y aportan un sistema común y universal de clasificación de las obras bibliográficas y de los documentos (Lewis & Spärck Jones, 1996). Este concepto de vocabulario también es utilizado en la Recuperación de Información (en adelante RI). En la RI, los vocabularios se utilizan como componentes de los sistemas software de RI para evitar la ambigüedad y polisemia del lenguaje, mejorando así la exhaustividad y la precisión en la indexación y recuperación de los recursos, y proporcionando un mecanismo de navegación temática o conceptual para la

[45] Fuente: https://www.merlot.org/merlot/materials.htm

colección de recursos que es comprensible para las personas que buscan recursos en una colección digital (Lancaster, 1972).

En la medida en la que un repositorio digital es un caso particular de un sistema de RI, los repositorios reproducen el concepto y uso de los vocabularios como componentes del repositorio para la catalogación y recuperación de objetos digitales.

Figura 16. Interfaz de consulta o exploración mediante un vocabulario (el tesauro ETB) del repositorio Educativo Español AGREGA[46]

En este sentido, los vocabularios se usan tanto en el proceso de indexación de objetos digitales como en el de búsqueda. En el proceso de indexación, los objetos digitales, al igual que en la catalogación temática, se representan mediante un grupo restringido de palabras que describen su contenido, y mediante índices del sistema de RI se enlazan las palabras y los objetos (Baeza-Yates et al., 1999). En el proceso de búsqueda de un objeto digital a partir de una consulta del usuario, la exhaustividad de los resultados se puede mejorar cuando se utiliza el vocabulario para ampliar las palabras de la consulta del usuario (Figura 16) mediante sinónimos y cuasinónimos, variantes ortográficas, hipónimos (palabras con significado más específico) e hiperónimos (más general) u otras palabras relacionadas (Lancaster, 1972). La precisión de los resultados de una búsqueda también puede mejorarse si se utiliza el

[46] Fuente: http://www.proyectoagrega.es/default/home.php

vocabulario para coordinar (i.e. combinar) las palabras de la consulta con otras que especifiquen más la consulta. De esta forma es posible, por ejemplo, distinguir entre homógrafos y desambiguar la consulta usando las definiciones, relaciones semánticas o frecuencia de uso que contiene el vocabulario. Por último, el vocabulario y los índices creados entre palabras y objetos digitales pueden utilizarse para proporcionar al usuario un mecanismo más de acceso al repositorio de objetos digitales: un vocabulario de navegación en la colección para explorar, seleccionar y localizar los objetos digitales (Figura 17).

Figura 17. Vocabulario para la navegación en el repositorio digital arqueológico Coclé de Panamá[47]

Los vocabularios, sin embargo, no están exentos de problemas que no sólo no mejoran la exhaustividad y precisión de la recuperación sino que, incluso, pueden empeorarlas. Las dos fuentes principales de fallos atribuibles a los vocabularios provienen de la excesiva o poca especificidad del vocabulario y de las relaciones ambiguas o no controladas, siendo la falta de equilibro en la especificidad el factor, probablemente, más crítico para la eficacia de la búsqueda (Lancaster, 1972). Básicamente, cuando el vocabulario es muy específico los objetos necesitan describirse con muchas palabras de significado muy preciso. Esto proporciona mayor precisión en la indexación pero, al mismo tiempo, dificulta la localización de los recursos, porque el usuario tiene que acertar en usar en su consulta las palabras específicas con las que ha sido catalogado el objeto. Esto exige del usuario un conocimiento profundo del vocabulario

[47] Fuente: http://oda-fec.org/cocle/

y de los objetos digitales, necesario para saber expresar la consulta con precisión. Por el contrario, cuanto más general sea el vocabulario, más probabilidades tiene el usuario de encontrar los objetos que busca, utilizando conceptos de significado amplio (mejora de la exhaustividad), pero también es probable que los resultados obtenidos sean, en un alto porcentaje, irrelevantes. Se trata, por lo tanto, de lograr un equilibrio entre generalidad y especificidad del vocabulario que debe contener palabras suficientemente específicas como para permitir recuperar los objetos digitales de forma precisa y, al mismo tiempo, palabras suficientemente generales como para recoger las consultas de usuarios menos expertos en el dominio de conocimiento y en el vocabulario del repositorio.

Respecto a la segunda fuente de fallos en la RI debida a los vocabularios, ésta proviene de las relaciones ambiguas o no controladas entre palabras del vocabulario. Estas relaciones pueden generar coordinaciones falsas de palabras durante el proceso de búsqueda cuando se amplía la consulta. Por ejemplo, si se busca un objeto sobre "*lenguaje de programación*" y existe una sinonimia no controlada (p.e., no contextualizada) en el vocabulario entre "*lenguaje*" e "*idioma*" se podría generar la coordinación falsa "*idioma*" y "*programación*" que seguramente obtendría resultados erróneos.

Una tercera fuente de fallos, que no es directamente atribuible al vocabulario, sino a una catalogación subjetiva, es el uso de palabras no adecuadas para representar a los objetos. Merece la pena tenerlo en cuenta porque constituye uno de los problemas más habituales cuando los vocabularios se generan de forma libre y abierta sin un estudio y consenso previo. En estos casos, los objetos pueden ser irrecuperables para los usuarios que no han participado en la catalogación y que no tienen el mismo criterio para denotar a los objetos que los catalogadores.

En definitiva, los vocabularios son el núcleo del subsistema del repositorio digital para la catalogación y recuperación de los objetos digitales basados en el significado y contenido de dichos objetos. Este subsistema facilita el acceso a los objetos mejorando la exhaustividad y precisión de las consultas de los usuarios y, también, sirviendo de mapa conceptual de navegación en la colección de objetos. Sin embargo, debe tenerse en cuenta que cuestiones como el desequilibrio entre palabras generales y específicas, la existencia de palabras y relaciones entre palabras ambiguas y no controladas y una catalogación errónea pueden afectar de forma muy negativa a la eficiencia del repositorio respecto a la recuperación de los objetos digitales.

2.3.3 *Los vocabularios controlados y libres*

Los vocabularios utilizados en la catalogación de objetos digitales son, básicamente, de dos tipos: vocabularios controlados y vocabularios libres. Un vocabulario controlado es una colección de términos en una o varias lenguas relativos a un área de conocimiento específica. También se denomina *terminología* (p.e., *vocabulario del vino* o *terminología del vino*) y, normalmente se crean de forma consensuada por un comité de especialistas en terminología y en el dominio de conocimiento del vocabulario. Un *término* es una palabra o grupo de palabras del lenguaje natural que se utiliza para designar a un único concepto, de forma que a diferencia de un vocabulario del lenguaje natural, en los vocabularios controlados no existe ambigüedad (Arnold, Balkan, Humphreys, Meijer, & Sadler, 1994; ISO/TC 37/SC 1, 2014). Los términos pueden estar agrupados en categorías o en facetas. Por ejemplo, en la Figura 20 se muestra la categoría 8 "*Language, Linguistics, Literature*", dentro de la cual está la subcategoría 8.1. "*Linguistics and language*" y los términos de esta subcategoría: "*Facets of linguistics*", "*General linguistics*", etc. La diferencia entre categorías y facetas es que las categorías pueden solaparse, es decir, pueden existir términos que pertenezcan a más de una categoría, mientras que las facetas son categorías disjuntas (Aitchison, Gilchrist, & Bawden, 2000). Los términos, además, pueden tener información asociada como su significado, información gramatical, información de uso, ortográfica, etc., en lo que se denomina la *entrada terminológica* (Figura 18). Los términos, además, pueden estar relacionados con otros términos del vocabulario, como puede verse en la Figura 19.

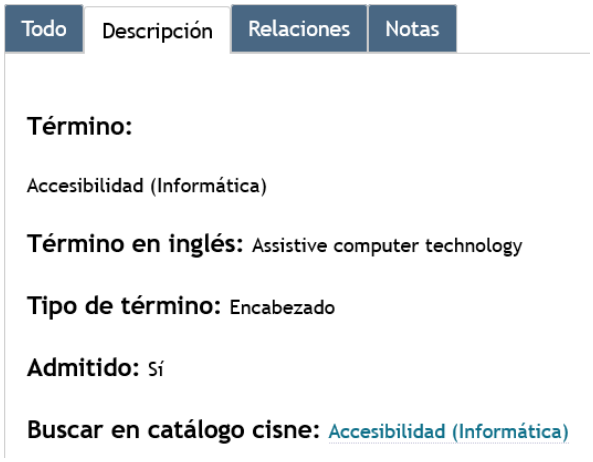


Figura 18. Entrada terminológica para el término "Accesibilidad (Informática)" del tesauro de la Biblioteca de la UCM[48]

[48] Fuente: http://alfama.sim.ucm.es/tesauro/

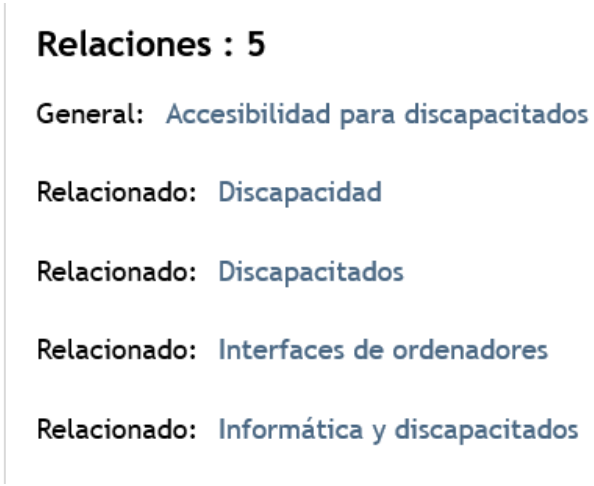


Figura 19. Relaciones con otros términos del término "Accesibilidad (Informática)"[49]

Los vocabularios controlados pueden ser generales o específicos de un dominio de conocimiento (Hirst, 2009). Un ejemplo de vocabulario general controlado es cualquiera de los vocabularios para la catalogación de recursos bibliográficos, como los ya mencionados LCC, DDC y la CDU (Figura 20).

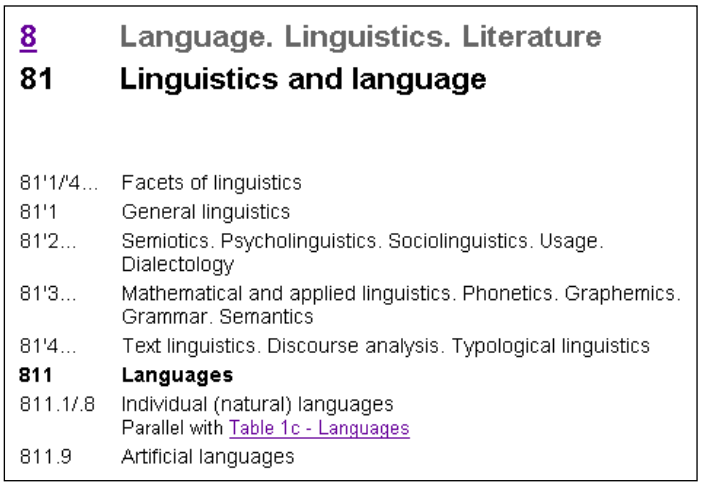


Figura 20. Muestra de la Clasificación Universal Decimal (CDU)[50]

Un ejemplo de un vocabulario controlado de especialidad es el sistema de clasificación de la ACM (Figura 21)

[49] Fuente: http://alfama.sim.ucm.es/tesauro/

[50] Fuente: http://www.udcc.org/

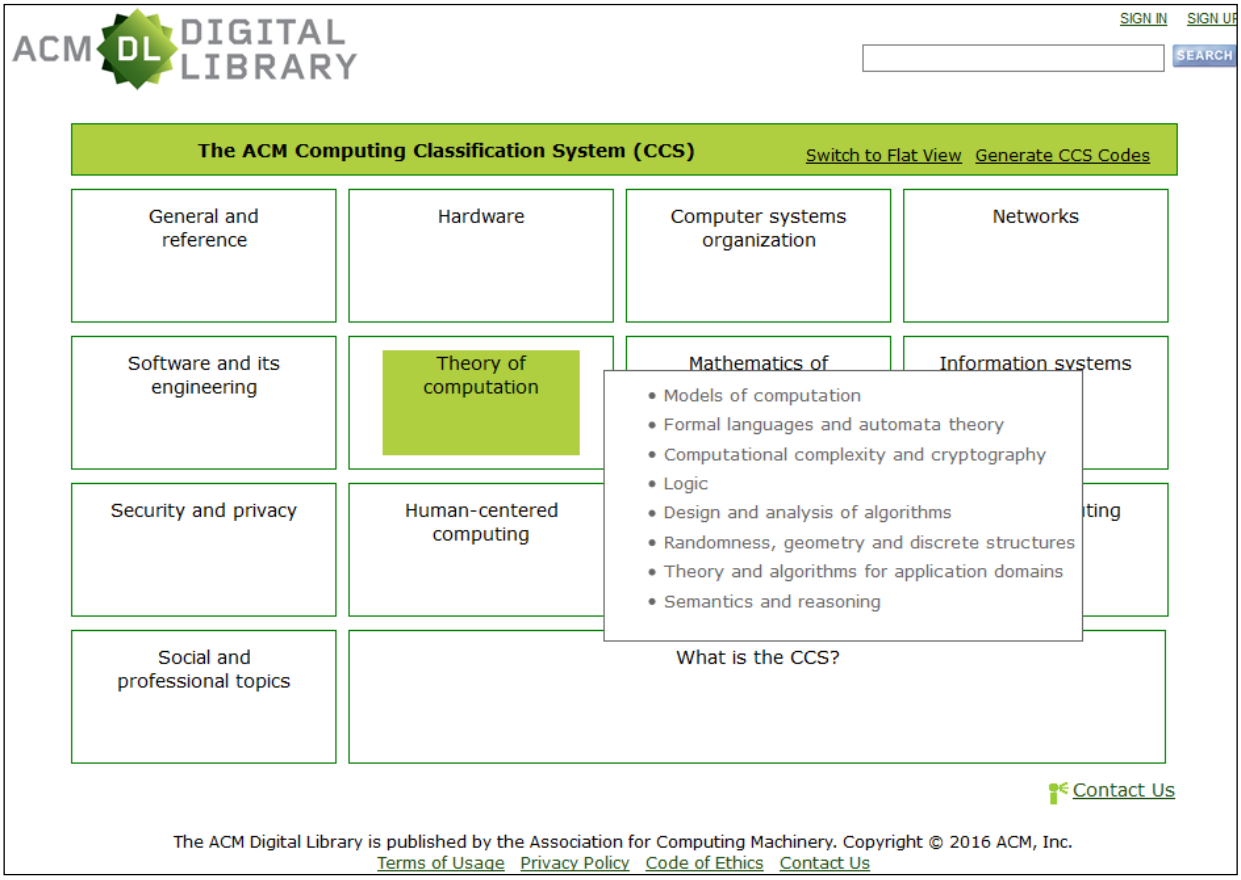


Figura 21. Muestra del Sistema de Clasificación de la ACM [51]

Los vocabularios controlados mejoran la eficiencia en la recuperación de objetos digitales porque minimizan al máximo los fallos debidos al propio vocabulario, pero tienen dos inconvenientes: (i) no garantizan que la catalogación sea correcta, no sólo por la subjetividad inherente a la asignación de términos, sino también porque no está garantizado que el vocabulario sea capaz de describir con precisión todos los objetos digitales de una colección (Friesen, 2004; Heath, McArthur, McClelland, & Vetter, 2005; Hepp, 2007), especialmente en contextos altamente especializados (Lee & Sugimoto, 2006); y, (ii) no está garantizado que sean fáciles de manejar por los usuarios. Saber buscar en un vocabulario que no es familiar o que cataloga los objetos con un sentido diferente al que tiene el usuario que busca, constituye una de las principales desventajas de estos vocabularios. Esta disociación entre el vocabulario de catalogación y el vocabulario de los usuarios es, probablemente, la principal causa de creación y uso de múltiples vocabularios para un mismo dominio temático y, en consecuencia, una de las principales dificultades para la reutilización de colecciones de objetos digitales y la

[51] Fuente: http://dl.acm.org/ccs/ccs.cfm

interoperabilidad entre repositorios (Friesen, 2004; Lee & Sugimoto, 2006; Van Assche, Anido-Rifon, Campbell, & Willem, 2003).

Figura 22. Uso integrado del sistema UDC en el tesauro de la Biblioteca de la UCM[52]

En este sentido, desde el punto de vista de la interoperabilidad, los vocabularios pueden utilizarse como, (i) vocabularios de referencia cuando son utilizados por los repositorios como sistema común de catalogación para intercambiar información entre repositorios, como por ejemplo el uso del sistema UDC en la Biblioteca Digital de la UCM (Figura 22) y, (ii) los vocabularios *ad hoc*, creados específicamente para un grupo de usuarios con un propósito y dominio de conocimiento concretos (p.e., Octeto[53]). En este último caso, sin embargo, debe considerarse el uso de directrices estándares para la construcción de vocabularios de forma que se pueda facilitar, en todo caso, la interoperabilidad entre vocabularios y repositorios (ANSI/NISO, 2005; DRI, 2003; LMF, 2008; Nilsson et al., 2008; Van Assche et al., 2003).

Los vocabularios libres, por su parte, son vocabularios creados, normalmente, de forma inductiva y colaborativa por los usuarios de una colección, según se van añadiendo objetos digitales a la colección para describir los contenidos de dichos objetos. Se denominan de forma general *folksonomías* (Mathes, 2004). Pueden ser creados por comunidades de especialistas en una materia para categorizar los objetos de una colección particular (véase la Figura 23 para un ejemplo), o bien pueden ser creados de forma más abierta por comunidades de usuarios en la web, no necesariamente especialistas. Son ejemplos de *folksonomías* abiertas las categorizaciones de los sitios web *del.icio.us* que categoriza enlaces a sitios favoritos en la web o *Flickr* un repositorio de fotos (Figura 24).

[52] Fuente: http://alfama.sim.ucm.es/tesauro/

[53] Aplicación Octeto de la universidad Universitat Jaume I (http://cent.uji.es/octeto/)

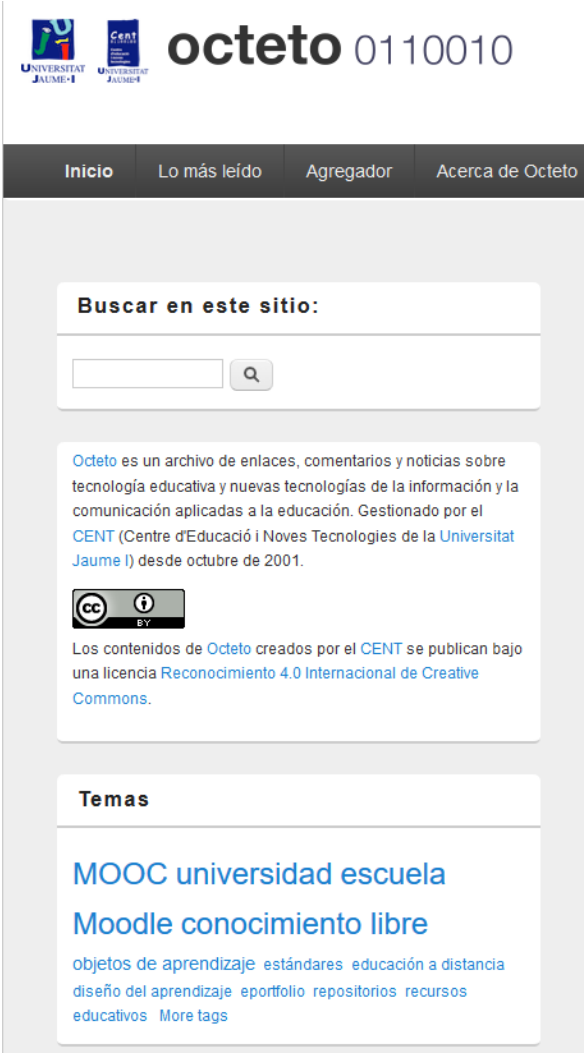


Figura 23. Vocabulario de tópicos del sistema Octeto de noticias de Tecnología Educativa de la Universidad Jaume I[54]

En las *folksonomías* los términos son etiquetas creadas colaborativamente por una comunidad de usuarios, no existe un conjunto de etiquetas fijas, ni un vocabulario controlado predeterminado y son los usuarios los que deciden usar unas determinadas palabras para designar los objetos digitales según sus puntos de vista, necesidades e intereses (Rodríguez Bravo, 2011). Esto hace que las *folksonomías* sean sistemas rápidos y baratos de construir, flexibles y adaptados a los usuarios que las utilizan. Sin embargo, la falta de control genera problemas que afectan de forma significativa a la precisión en la recuperación de la información: la polisemia, la sinonimia, las variantes morfológicas (como plurales), errores ortográficos, y la profundidad o la especificidad del vocabulario (Noruzi, 2006). Además, la

[54] Fuente: http://cent.uji.es/octeto/

ausencia de normas para la construcción de términos compuestos y las etiquetas subjetivas y no pertinentes complican el mantenimiento de la coherencia del vocabulario.

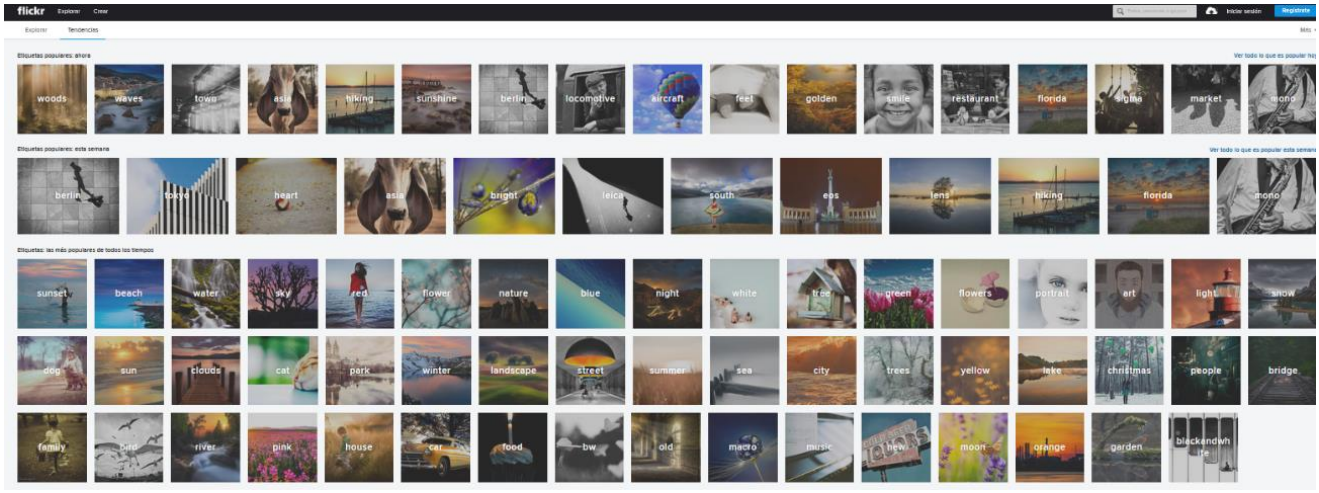


Figura 24. Muestra de la folksonomía de flickr[55]

2.3.4 *Los vocabularios controlados: tipología*

La descripción de los distintos tipos de vocabularios controlados merece una sección aparte, por su influencia en la catalogación y recuperación de los objetos digitales. Basándose en las directrices del estándar de construcción de vocabularios monolingües (ANSI/NISO, 2005) y el análisis de la Comisión Europea para la Normalización en su recomendación (Van Assche et al., 2003) sobre el uso de los vocabularios para la descripción de objetos digitales educativos (i.e. objetos de aprendizaje), la selección de un tipo u otro para la catalogación de los objetos digitales en un repositorio depende fundamentalmente de cuáles representan mejor: (i) la naturaleza de los objetos digitales y, en consecuencia, los requisitos de descripción y organización de las colecciones, (ii), el tipo de consultas, búsquedas y perfiles de usuario, (iii), la compatibilidad con el repositorio donde se integra, y, (iv), los recursos (económicos, personal y temporales) que se pueden invertir para su construcción (si es necesario) y, en todo caso, mantenimiento. A este respecto, se pueden distinguir cinco tipos principales de vocabularios controlados, que se revisan brevemente a continuación.

[55] Fuente: https://www.flickr.com/

2.3.4.1 Listas de términos.

Las listas de términos son un tipo de vocabulario terminológico no estructurado que se encuentra formado por un conjunto de términos prefijados en los que no existen relaciones entre los mismos más allá de representar a un dominio de conocimiento concreto (Hedden, 2008). En la Figura 25 se muestra un ejemplo de una lista de palabras que representa los códigos de referencia de los países según la norma ISO 3166. En la Figura 12 se mostró un ejemplo de cómo se usa este estándar para dar valores al elemento *language* en el esquema de metadatos Dublin Core resumido en la Figura 10.

Figura 25. Tabla de códigos de referencia de los países según la norma ISO 3166[56]

2.3.4.2 Taxonomías.

Las taxonomías son un tipo de vocabulario terminológico estructurado donde los términos presentan relaciones de dependencia u orden dando lugar a una estructura jerárquica

[56] Fuente: http://www.monografias.com/trabajos71/norma-iso-web-internet/norma-iso-web-internet2.shtml

en forma de árbol (Hedden, 2008). El proceso de catalogación consiste en encontrar las secuencias de términos dentro del árbol, denominados *caminos taxonómicos*, que mejor caracterizan al objeto digital que está siendo catalogado. Cuanto más largo sea el camino taxonómico, más específica será la información representada. La Figura 26 muestra un ejemplo de taxonomía. Merece también destacarse que en algunos casos las taxonomías se integran en vocabularios de ámbito general para mejorar la eficacia de la búsqueda evitando, así, los fallos por excesiva especialización. Un ejemplo es la integración de las taxonomías de Medicina *Mesh* (Lipscomb, 2000) y la taxonomía de tópicos de la web *Open Directory* en la ontología *Cyc* (Reed & Lenat, 2002).

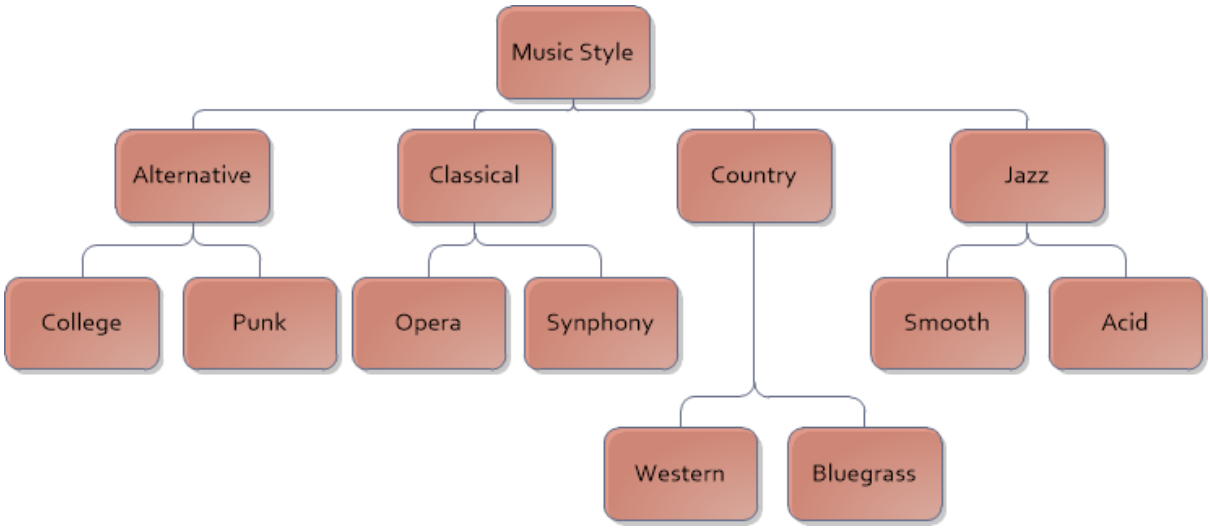


Figura 26 Taxonomía de los estilos musicales[57]

2.3.4.3 Tesauros.

Los tesauros son redes de términos y relaciones semánticas entre término. Las relaciones semánticas estándares definidas en el estándar de construcción de tesauros (ANSI/NISO, 2005) son "las relaciones de equivalencia, homógrafos, jerárquicas y asociativas y deben visualizarse claramente mediante marcadores estándares y recíprocos". Los tesauros pueden tener organizados los términos en categorías o facetas. En particular, la organización de los términos en facetas conduce a los denominados *tesauros facetados* (Hedden, 2008; Nasir Uddin & Janecek, 2007; Yee, Swearingen, Li, & Hearst, 2003). Los términos pueden tener información adicional (denominadas notas de ámbito) para precisar la definición del término.

[57] Fuente: http://www.hipertexto.info/documentos/indizacion.htm

En la Figura 27 se muestra el término "*Language instruction*" del tesauro de la UNESCO utilizado en sus repositorios digitales. El marcador MT indican la categoría a la que pertenece el término, UF (*used for*) indica los sinónimos (términos no preferidos) y el marcador NT (*narrower term*) indica los términos más específicos. Obsérvese a la derecha de los términos la indicación del número de objetos asociados al término y el enlace a los mismos.

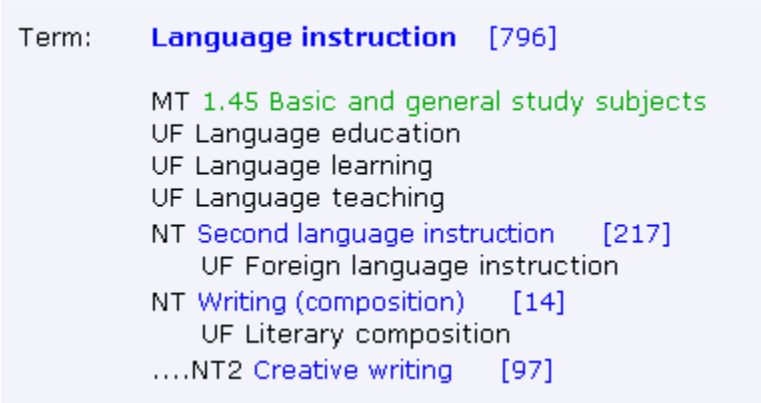


Figura 27. Término "Language instruction" del tesauro del repositorio de recursos documentales de la UNESCO.[58]

2.3.4.4 Ontologías.

A diferencia de los otros tipos de vocabularios controlados, las ontologías están formadas por conceptos que pueden denotarse o no con términos, interconectados mediante relaciones semánticas diversas de algún dominio (Horrocks, 2008; Tello, 2001). Estos conceptos y relaciones han sido consensuados por los miembros de un dominio de conocimiento determinado y a través de la ontología es posible reflejar dicho conocimiento. Los principales elementos de una ontología son los conceptos (ideas que se quieren formalizar), relaciones (representan un enlace que existe entre los conceptos), funciones (un tipo de relación donde se identifica un elemento mediante el cálculo de una función sobre varios elementos de la ontología), instancias (objetos concretos de un concepto) y axiomas (teoremas definidos sobre relaciones que deben cumplir los elementos de una ontología). La catalogación usando una ontología consiste en describir un objeto digital usando las instancias de los conceptos definidos en la ontología. De esta forma el objeto queda descrito con respecto a un dominio de conocimiento, y es posible aprovechar las relaciones definidas en la ontología entre sus

[58] Fuente: http://vocabularies.unesco.org/browser/thesaurus/en/

términos para realizar consultas. En la Figura 28 se muestra un ejemplo de ontología sobre el dominio de los vinos y las comidas (Dean et al., 2003).

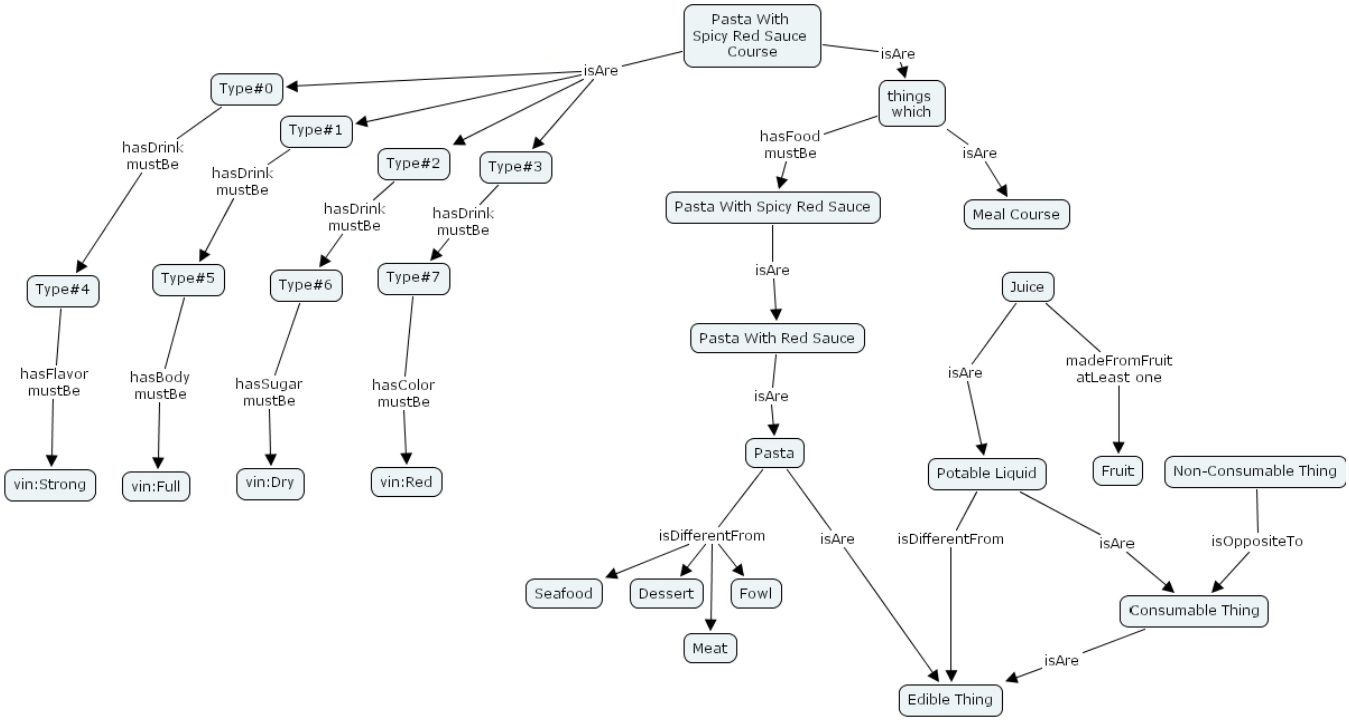


Figura 28. Ejemplo de ontología sobre el dominio de los vinos[59]

2.3.4.5 Glosarios.

Los glosarios son "catálogos de palabras de una misma disciplina, de un mismo campo de estudio, de una misma obra, etc., definidas o comentadas" (RAE, 2013). En este sentido están orientados, fundamentalmente, para uso humano. Su uso directo como catálogos temáticos en los repositorios digitales es escaso, pero tienen la ventaja de que las definiciones que contienen ayudan a precisar el contenido de los objetos digitales que describen (Tabata & Mitsumori, 2002).

2.4 Tecnologías utilizadas para representar los esquemas de catalogación

Para facilitar el uso de los vocabularios y los esquemas de metadatos, se realizan representaciones equivalentes de los mismos en formatos de datos procesables por un sistema

[59] Fuente: http://cmapsinternal.ihmc.us/viewer/cmap/1084988313677_1007478705_1904

informático. IMS VDEX, XML, RDF y OWL son los formatos que se revisan a continuación por estar entre los más utilizados en los repositorios digitales.

```
<?xml version="1.0" encoding="UTF-8"?>
<vdex orderSignificant="false" profileType="hierarchicalTokenTerms" language="en"
xsi:schemaLocation="http://www.imsglobal.org/xsd/imsvdex_v1p0 imsvdex_v1p0.xsd"
xmlns="http://www.imsglobal.org/xsd/imsvdex_v1p0" xmlns:xsi=
"http://www.w3.org/2001/XMLSchema-instance">
   <vocabName>
      <langstring>MeSH (National Institute of Health Medical Subject Headings)
      </langstring>
   </vocabName>
   <vocabIdentifier>http://www.fdgroup.com/~ftpkod/kmap/mesh_v1p0.xml
   </vocabIdentifier>
   <term>
      <termIdentifier>L01</termIdentifier>
      <caption>
         <langstring>Information Science</langstring>
      </caption>
      <term>
         <termIdentifier>L01.040</termIdentifier>
         <caption>
            <langstring>Book Collecting</langstring>
         </caption>
      </term>
      <term>
         <termIdentifier>L01.080</termIdentifier>
         <caption>
            <langstring>Chronology</langstring>
         </caption>
      </term>
```

Figura 29. Fragmento de Taxonomía para IMS VDEX[60]

2.4.1 *IMS VDEX*

El IMS Vocabulary Definition Exchange (VDEX)[61] es una especificación de IMS Global Learning Consortium utilizada para representar vocabularios terminológicos (Sarasa, Canabal, Sacristán, & Jiménez, 2008). En este sentido, la especificación soporta la representación de vocabularios controlados y de vocabularios jerárquicos o taxonomías. También soporta la representación de tesauros, aunque el modelo de información no está preparado explícitamente para representar este tipo de relaciones. Para expresar estas relaciones, IMS VDEX, proporciona términos que representan las posibles relaciones definidas entre otros pares de términos. Para poder describir cada tipo de vocabulario existe un perfil de

[60] Fuente: https://www.imsglobal.org/vdex/vdexv1p0/imsvdex_bestv1p0.html

[61] IMS Vocabulary Definition Exchange (https://www.imsglobal.org/vdex/index.html)

IMS VDEX concreto. En la Figura 29 se muestra un ejemplo de un fragmento de una taxonomía descrita en IMS VDEX.

```
<?xml version="1.0" encoding="UTF-8"?>
<lom xmlns=http://ltsc.ieee.org/xsd/LOM xsi:schemaLocation=
"http://ltsc.ieee.org/xsd/LOM
  lomCustom.xsd">
   <lom:general>
     <lom:identifier>
       <lom:catalog>Catálogo unificado mec-red.es-ccaa de identificación de ODE
       </lom:catalog>
       <lom:entry>es_20070518_3_0030500</lom:entry>
     </lom:identifier>
     <lom:title>
       <lom:string>La energía externa del Planeta</lom:string>
     </lom:title>
     <lom:language>es</lom:language>
     <lom:description>
       <lom:string>
         Explicación del origen de la energía que proviene del Sol, la
         composición y
         funciones de la Atmósfera y de la Hidrosfera.
       </lom:string>
     </lom:description>
     <lom:keyword>
       <lom:string>geología </lom:string>
     </lom:keyword>
    <lom:keyword>
       <lom:string> biosfera</lom:string>
     </lom:keyword>
     <lom:coverage>
```

Figura 30. Fragmento de documento XML con metadatos LOM[62]

2.4.2 *XML*

Otra tecnología ampliamente utilizada para codificar esquemas de catalogación es XML (eXtensible Markup Language)[63] (Bradley, 2001; Evjen et al., 2007), un metalenguaje que permite describir lenguajes de marcado que representan tipos de documentos electrónicos. Un *lenguaje de marcado* se caracteriza porque permite estructurar la información mediante conjuntos de marcas que delimitan *elementos* (fragmentos de contenidos de la información). Las marcas aparecen en forma de pares, una marca de apertura y otra de cierre (también denominadas etiquetas). Así mismo una marca de apertura puede tener asociados uno o más atributos que refinan el contenido delimitado por una marca. Los elementos pueden estar formados por otros elementos, por contenido o por ambos. La estructura y relaciones que

[62] Fuente: http://agrega.educacion.es/wiki/index.php?title=Creaci%C3%B3n_de_Binding_XML_de_LOM-ES

[63] https://www.w3.org/TR/REC-xml/

existen entre las marcas es lo que constituye la definición del lenguaje de marcado, el cual se representa en lo que se denomina un *esquema*. Una instancia de un lenguaje de marcado es un documento XML en el que aparecen las marcas delimitando y describiendo porciones de información. Se utiliza para facilitar el uso de los esquemas de metadatos, para lo cual se define una correspondencia del modelo de metadatos a un lenguaje XML, de forma que los metadatos se corresponden con un lenguaje de marcas, y las instancias de los metadatos asociadas a un objeto digital se pueden representar como un documento XML. En la Figura 30 se muestra un fragmento de un documento XML que codifica unos metadatos LOM.

2.4.3 *RDF*

RDF (Resource Description Framework) (Hitzler, Krotzsch, & Rudolph, 2009) es una especificación que permite describir propiedades sobre recursos web en general mediante tripletas sujeto-predicado-objeto, donde el sujeto es el recurso que se está describiendo, el predicado es la propiedad o relación que se desea establecer acerca del recurso, y por último el objeto es el valor de la propiedad o el otro recurso con el que se establece la relación. Cada objeto a su vez puede ser otra tripleta. Esta forma de representación dota a la vez de estructura y semántica a la información. Permite representar vocabularios y sistemas de catalogación, así como instancias de los mismos. En la Figura 31 se muestra un fragmento de un documento RDF.

2.4.4 *OWL*

Por último, *OWL (Ontology WebLanguage)* (Bechhofer, 2009) es una especificación que extiende a RDF para poder representar ontologías. Además de poder describir propiedades, permite representar información semántica sobre las relaciones. OWL proporciona tres lenguajes, cada uno con nivel de expresividad mayor que el anterior. En el contexto de los tesauros y taxonomías, el lenguaje más adecuado es el OWL Lite, dado que permite la introducción de clasificaciones jerárquicas y restricciones simples como las referidas a la cardinalidad. En la Figura 32 se muestra un ejemplo de fragmento de ontología en OWL.

```
- <rdf:RDF xml:base="http://chroniclingamerica.loc.gov/lccn/sn85042071">
  - <rdf:Description rdf:about="/lccn/sn85042071.rdf">
      <dcterms:modified rdf:datatype="http://www.w3.org/2001/XMLSchema#dateTime">2010-05-04T22:32:29-04:00</dcterms:modified>
      <dcterms:creator rdf:resource="http://chroniclingamerica.loc.gov/awardees/dlc#awardee"/>
      <ore:describes rdf:resource="/lccn/sn85042071#title"/>
      <rdf:type rdf:resource="http://www.openarchives.org/ore/terms/ResourceMap"/>
      <dcterms:created rdf:datatype="http://www.w3.org/2001/XMLSchema#dateTime">2010-05-04T22:32:29-04:00</dcterms:created>
    </rdf:Description>
  - <rdf:Description rdf:about="/lccn/sn85042071#title">
      <dcterms:title>Black Rock gazette.</dcterms:title>
      <dcterms:hasFormat rdf:resource="/lccn/sn85042071/marc.xml"/>
      <frbr:successor rdf:resource="/lccn/sn94057511#title"/>
      <owl:sameAs rdf:resource="info:oclcnum/11718801"/>
      <owl:sameAs rdf:resource="info:lccn/sn85042071"/>
      <dc:publisher>B. Ferguson</dc:publisher>
      <rda:placeOfPublication>Black Rock, N.Y.</rda:placeOfPublication>
      <dcterms:language rdf:resource="http://www.lingvoj.org/lang/en"/>
      <rdfs:seeAlso rdf:resource="http://lccn.loc.gov/sn85042071"/>
      <rdfs:seeAlso rdf:resource="http://www.worldcat.org/oclc/11718801"/>
      <ore:isDescribedBy rdf:resource="/lccn/sn85042071.rdf"/>
      <dcterms:coverage rdf:resource="http://sws.geonames.org/5110629/"/>
      <dcterms:coverage rdf:resource="http://dbpedia.org/resource/Buffalo%2C_New_York"/>
      <dcterms:date rdf:datatype="http://www.loc.gov/standards/datetime#edt">1824/1827</dcterms:date>
      <dc:subject>Erie County (N.Y.)--Newspapers.</dc:subject>
      <dc:subject>Black Rock (Buffalo, N.Y.)--Newspapers.</dc:subject>
      <dc:subject>Buffalo (N.Y.)--Newspapers.</dc:subject>
      <rdf:type rdf:resource="http://purl.org/ontology/bibo/Newspaper"/>
    </rdf:Description>
  </rdf:RDF>
```

Figura 31. Fragmento de un documento en formato RDF[64]

```
- <owl:Ontology rdf:about="">
  - <rdfs:comment xml:lang="en">
      An example ontology that contains all constructs required for the various versions of the Pizza Tutorial run by Manchester University
    </rdfs:comment>
  - <owl:versionInfo xml:lang="en">
      v.1.5. Removed protege.owl import and references. Made ontology URI date-independent
    </owl:versionInfo>
    <owl:versionInfo rdf:datatype="http://www.w3.org/2001/XMLSchema#string">version 1.5</owl:versionInfo>
  - <owl:versionInfo xml:lang="en">
      v.1.4. Added Food class (used in domain/range of hasIngredient), Added several hasCountryOfOrigin restrictions on pizzas, Made hasTopping invers functional
    </owl:versionInfo>
  </owl:Ontology>

- <owl:ObjectProperty rdf:about="#hasBase">
    <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#InverseFunctionalProperty"/>
    <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#FunctionalProperty"/>
    <rdfs:subPropertyOf rdf:resource="#hasIngredient"/>
    <rdfs:range rdf:resource="#PizzaBase"/>
    <rdfs:domain rdf:resource="#Pizza"/>
    <owl:inverseOf rdf:resource="#isBaseOf"/>
  </owl:ObjectProperty>
```

Figura 32. Fragmento de una ontología en formato OWL[65]

[64] Fuente: http://www.ourontario.ca/ppt/linkeddata/

[65] Fuente: http://protege.stanford.edu/ontologies/pizza/pizza.owl

2.5 Reconfiguración de las colecciones digitales

A lo largo de la revisión de los esquemas estándar de catalogación para las colecciones de objetos digitales se han ido presentando los problemas que limitan su eficacia. Estos problemas se generan por la diversidad de esquemas de catalogación, incluso para colecciones de un mismo dominio de conocimiento. La rápida evolución de la tecnología permite crear cada vez esquemas de colección más completos y más extensos descriptivamente sin que la búsqueda y recuperación se vean afectadas. Este factor hace que haya todavía mayor diversidad no sólo en los esquemas de catalogación sino también en las aplicaciones que los gestionan.

En este sentido, el primero de los problemas es la dificultad para conciliar la diversidad de esquemas de catalogación, tanto de modelos de metadatos como de vocabularios. Esta diversidad está justificada, como ya se ha visto por: (i) las diferencias en la naturaleza de las colecciones de objetos digitales que proceden de dominios de conocimiento diferentes, (ii), las diferencias en los criterios de catalogación entre comunidades de especialistas de un mismo dominio, y, (iii), la falta de precisión descriptiva de los estándares de esquemas, lo que produce una dificultad para entender y usar esquemas de catalogación complejos (p.e., MARC o LCC) que deriva en la simplificación de los mismos o en la creación de esquemas propios (p.e., *folksonomías*). Esta diversidad de esquemas de catalogación limita la interoperabilidad entre los repositorios y, en consecuencia, dificulta la migración y la reutilización de los objetos de las colecciones incrementando significativamente el coste del mantenimiento de:

- Las colecciones de objetos digitales que a lo largo de su ciclo de vida necesitan actualizar los esquemas de catalogación, metadatos o vocabularios debido a la actualización de los modelos estándares en los que se basan o bien debido a la aparición de nuevos modelos de catalogación (Caplan, 2008; Caplan et al., 2010). Tal es el caso de la BNE que, actualmente, necesita incorporar el nuevo estándar de catalogación RDA (Sección 2.1).
- Las colecciones que van incorporando nuevos objetos con contenidos, propiedades y valores no contemplados en el diseño inicial de la colección (Fernández-Valmayor Crespo, Guinea Bueno, Navarro Martín, & Sierra Rodríguez, 2005).
- Las colecciones creadas sobre la base de otras colecciones o porciones de colecciones existentes (Romero López, 2013, 2014; Bekaert et al., 2006; Rani et al., 2006; Salem, 2009; Schwertner & Chavez, 2005).

- Los sistemas que centralizan múltiples colecciones para promover su difusión y uso (p.e., Europeana para la difusión del patrimonio cultural de los países europeos). Estos sistemas importan y unifican los esquemas de catalogación de diversos repositorios, cada uno con un esquema de catalogación diferente (Barroso, Azevedo, & Ribeiro, 2009; Benedetti & Masci, 2005; Caplan, 2010; Goldsmith & Knudson, 2006)

El segundo de los problemas encontrados es el de la adaptación de los esquemas de catalogación a los continuos y rápidos cambios tecnológicos que generan nuevas formas de codificación de los esquemas de catalogación, nuevos estándares de catalogación, nuevos formatos de contenidos de los objetos digitales que no han sido tenidos en cuenta o nuevos sistemas de almacenamiento y gestión de los esquemas de catalogación y de las colecciones de objetos digitales. Este problema afecta de forma seria a la perdurabilidad de los esquemas de catalogación y, en consecuencia, a la perdurabilidad de las colecciones.

Una manera sencilla de resolver tanto el problema de la conciliación entre esquemas de catalogación diversos como el de la adaptación permanente a nuevas tecnologías para la catalogación de colecciones de objetos digitales sería poder reconfigurar los esquemas de catalogación en cualquier momento del ciclo de vida de una colección sin que esta reconfiguración afectase a las estructuras de almacenamiento interno de dicha colección y sin que afectase de forma significativa al servicio que presta el repositorio durante la reconfiguración.

De esta forma, la reconfiguración de los esquemas de catalogación es una cuestión crítica en el mantenimiento de las colecciones de objetos digitales y para asegurar la durabilidad de las colecciones, por cuanto que afecta al modo en que se organiza la colección, se recuperan los objetos y se navega por dicha colección. No obstante, y a pesar de la enorme importancia que tiene la reconfiguración de colecciones digitales de cara al mantenimiento y la preservación de las mismas, en la bibliografía revisada, el número de soluciones genéricas encontradas para la reconfiguración de esquemas en general es escaso. A continuación, se revisan las principales. En (Joaquín Gayoso-Cabada et al., 2016a, 2016c) se discuten otros enfoques más específicos, en el contexto de la reconfiguración de vocabularios controlados, más concretamente, la reconfiguración de tesauros facetados.

2.5.1 *Enfoques basados en traducción*

Estos enfoques se basan en "traductores" *ad hoc* para esquemas de metadatos prefijados (p.e., Mattso et al., 2002; Cañizares-González, 2013) que utilizan tablas de correspondencia

entre los elementos de los esquemas de transformación (Wang, Isenor, & Graybeal, 2011). En este caso es frecuente el uso de lenguajes de programación declarativos que facilitan la localización y el control de los cambios para la construcción de los traductores, como XSLT[66], utilizado de forma habitual para la transformación de esquemas de catalogación implementados (o serializados) con XML (Peltier, Bézivin, & Guillaume, s.f.).

2.5.2 *Enfoques basados en guías y especificaciones de transformación*

Estos enfoques se basan en el uso de guías y especificaciones de transformación (o armonización), estándares que pueden ser utilizadas en los repositorios como interfaz de conversiones entre los esquemas de catalogación de los objetos digitales y los esquemas internos del repositorio y en el proceso inverso de exportación (p.e., IEEE P1484.12.4™/D1, Guidelines for Using the IMS LRM to IEEE LOM 1.0 Transform)

2.5.3 *Enfoques basados en "interliguas"*

Estos enfoques promueven la utilización de un "interlingua" o descripción terminológica-conceptual de referencia respecto al cual referir los esquemas a transformar.

Esta aproximación se utiliza, por ejemplo, en (Nilsson et al., 2009) que propone utilizar los quince elementos básicos de Dublin Core como "interligua". Esto permite una transformación automática, pero parcial -la referida a los quince elementos básicos Dublin Core- de los elementos de un esquema de metadatos en otro. En esta propuesta, además, se describen cuatro niveles de correspondencia referida a los elementos Dublin Core: (i) correspondencia a nivel de términos, (ii) correspondencia a nivel semántico formal de los términos (basada en RDF para expresar formalmente la semántica), (iii) correspondencia a nivel de registro o conjunto de elementos y, (iv), correspondencia a nivel de registro entre elementos y sus restricciones (conforme al modelo abstracto Dublin Core).

En (Jurkiewicz & Nowiński, 2011) se propone un esquema de metadatos "interlingua", llamado BWMeta, general y capaz de preservar toda la información de catalogación de documentos tanto si está estructurada como metadatos como si es texto libre. Finalmente, en (Nilsson, 2008) se propone el uso de RDF como marco de armonización entre estándares generalizando la propuesta del Dublin Core Initiative de representación formal de la semántica de los elementos Dublin Core.

[66] https://www.w3.org/TR/xslt

2.6 Trabajos previos en el grupo de investigación

2.6.1 *Introducción*

En este apartado se revisan brevemente los trabajos previos desarrollados en el Grupo de Investigación en Ingeniería del Software e Inteligencia Artificial (ILSA) de la Universidad Complutense de Madrid sobre creación y gestión de colecciones de objetos digitales en los que se apoya la investigación llevada a cabo en esta tesis. Estos trabajos han cristalizado en los sistemas *Chasqui*, OdA, @note y *Clavy*. Estos cuatro sistemas presentan, como principal característica, el permitir la *reconfiguración dinámica* de los esquemas de catalogación. Así mismo, tanto @note como *Clavy* incorporan los resultados obtenidos en esta tesis.

2.6.2 *Chasqui*

El sistema *Chasqui* es un sistema desarrollado para la creación de repositorios de colecciones de objetos de digitales que almacena información sobre la catalogación de objetos arqueológicos (Arnaiz Barrero, 2008; Bueno, 2004; Fernández-Valmayor Crespo et al., 2005; J. L. Sierra, Fernandez-Valmayor, Guinea, Hernanz, & Navarro, 2005). El sistema surge de la necesidad de crear una infraestructura para repositorios que, entre sus características, tenga la posibilidad de que el esquema de catalogación de los objetos que se añaden al sistema evolucione dinámicamente, conforme se añaden nuevos objetos al mismo. Efectivamente, en el ámbito arqueológico es imposible prever la naturaleza de los recursos que se encuentran en una excavación, y es necesario modificar las estructuras de las catalogaciones de una colección de manera periódica según aparecen elementos con nuevas características. La Figura 33 muestra una instancia de Chasqui.

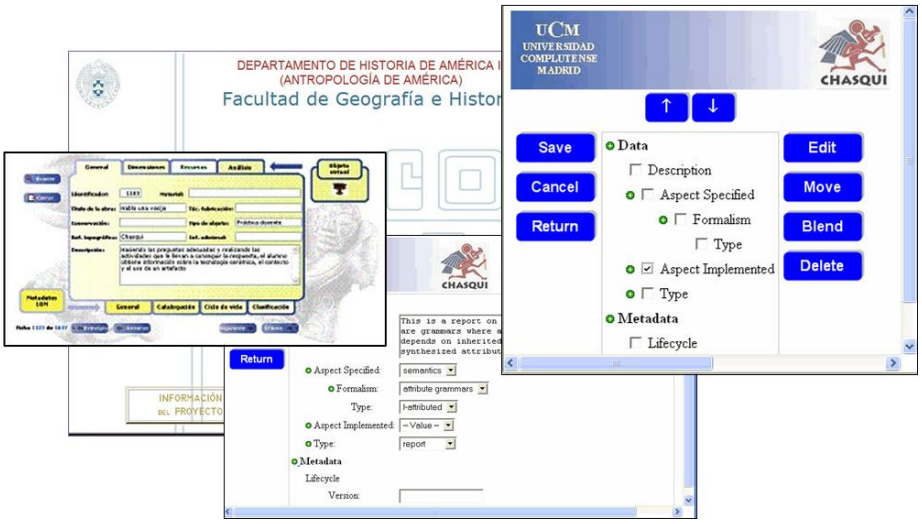


Figura 33. Una instancia del sistema Chasqui

En *Chasqui*, los objetos digitales pueden etiquetarse mediante *caminos* en un tesauro facetado (José Luis Sierra & Fernández-Valmayor, 2006). Dicho tesauro no existe a priori, sino que se crea dinámicamente, como resultado de añadir más y más objetos al sistema. Efectivamente, el añadido de un nuevo objeto supone actualizar el tesauro con los nuevos caminos indicados, si es que estos no existen ya. Por tanto, en un momento dado, el tesauro puede concebirse como la unión de todos los caminos asociados con los objetos almacenados en el repositorio.

De esta forma, *Chasqui* incluye un claro ejemplo de esquema de catalogación reconfigurable, donde la reconfiguración es transparente al usuario. No obstante, el sistema en sí adolece también de algunos problemas:

- Por una parte, *Chasqui* no incluye mecanismos que permitan editar el esquema de catalogación emergente. De esta forma, dicho esquema puede contener problemas tales como términos sinónimos, caminos de clasificación incorrectos, discrepancias entre múltiples expertos, etc.
- Por otra parte, *Chasqui* no aborda mecanismos que permitan adaptar las estructuras internas que soportan la navegación o la búsqueda por el repositorio. En su lugar, incluye un conjunto de tablas de navegación basadas en un estado concreto del tesauro, que deben mantenerse manualmente por un programador si dicho estado cambia.

2.6.3 *Oda*

A raíz del trabajo desarrollado en la aplicación *Chasqui*, se detectaron nuevas necesidades para la gestión de repositorios de objetos digitales en dominios especializados

(como el de la arqueología), de manera que se creó un nuevo sistema denominado ODA que extendía el modelo *Chasqui* para dar una mayor versatilidad a las ediciones, personalizar la visualización de las colecciones y reducir los requisitos de instalación (Fernández-Pampillón Cesteros, 2012; Fernández-Valmayor Crespo et al., 2013). La Figura 34 muestra una instancia de ODA.

Al contrario que en *Chasqui*, en ODA es posible definir y editar explícitamente los esquemas de catalogación. Dichos esquemas consisten en:

- Organizaciones jerárquicas de *atributos*. Cada atributo puede introducir, en los objetos digitales, un tipo de valor, o bien puede servir únicamente a propósitos estructurales (es decir, agrupar otros atributos). Los atributos se organizan en jerarquías, de forma que cada atributo posee, normalmente, un atributo padre.
- Vocabularios *controlados*. Dichos vocabularios se conciben como listas de términos, que pueden asociarse con distintos atributos para controlar sus contenidos en los objetos.

Tal y como se ha indicado, ODA proporciona mecanismos de edición para dichas estructuras, permitiendo definir los esquemas de catalogación más adecuados para cada escenario particular. Así mismo, el sistema implementa parcialmente la propuesta de reconfiguración de dichas estructuras descrita en (J. L. Sierra & Fernández-Valmayor, 2008). En particular, permite modificar libremente la filiación de los atributos en el esquema (cambiar sus padres). El sistema permite también editar las listas de términos de los vocabularios, cambiar el nombre de los atributos, añadir nuevos atributos, reordenar atributos, y eliminar aquellos atributos que no se usan para catalogar objetos.

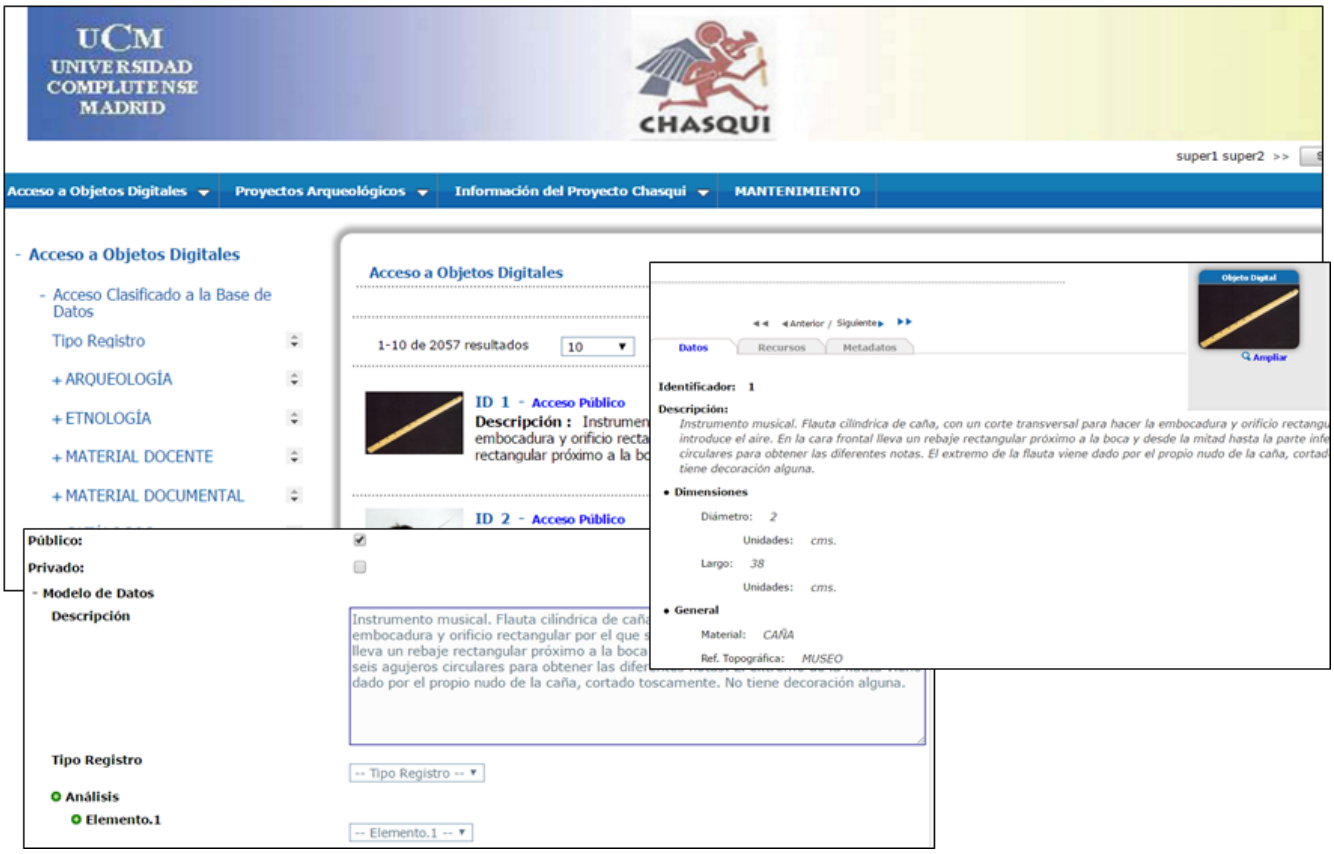


Figura 34. Una instancia del sistema ODA.

Al contrario que en *Chasqui*, los cambios en los esquemas no se reflejan únicamente en la manera en la que se presentan los objetos, sino que también se reflejan en las estructuras que soportan la navegación y las consultas por el repositorio. Para ello se utiliza un mecanismo basado en índices invertidos en RI (Chowdhury, 2010; Culpepper & Moffat, 2010; Zobel & Moffat, 2006), implementado mediante tablas relacionales. Dicho mecanismo funciona bien para colecciones de tamaño moderado, aunque conforme el tamaño de la colección aumenta, el rendimiento también se degrada notablemente.

2.6.4 *@Note*

La aplicación *@note*[67] (Gayoso Cabada, 2012; Joaquin Gayoso-Cabada et al., 2012; Ruiz et al., 2012) permite la creación de colecciones de anotaciones críticas sobre libros digitalizados. Una de sus principales características es que permite llevar a cabo actividades de anotación crítica de forma compartida. Las anotaciones se catalogan con una taxonomía que

[67] Página Web de @note (http://a-note.fdi.ucm.es)

sirve de base para de navegar y buscar sobre la colección de anotaciones. La Figura 35 muestra una instancia de la aplicación @note.

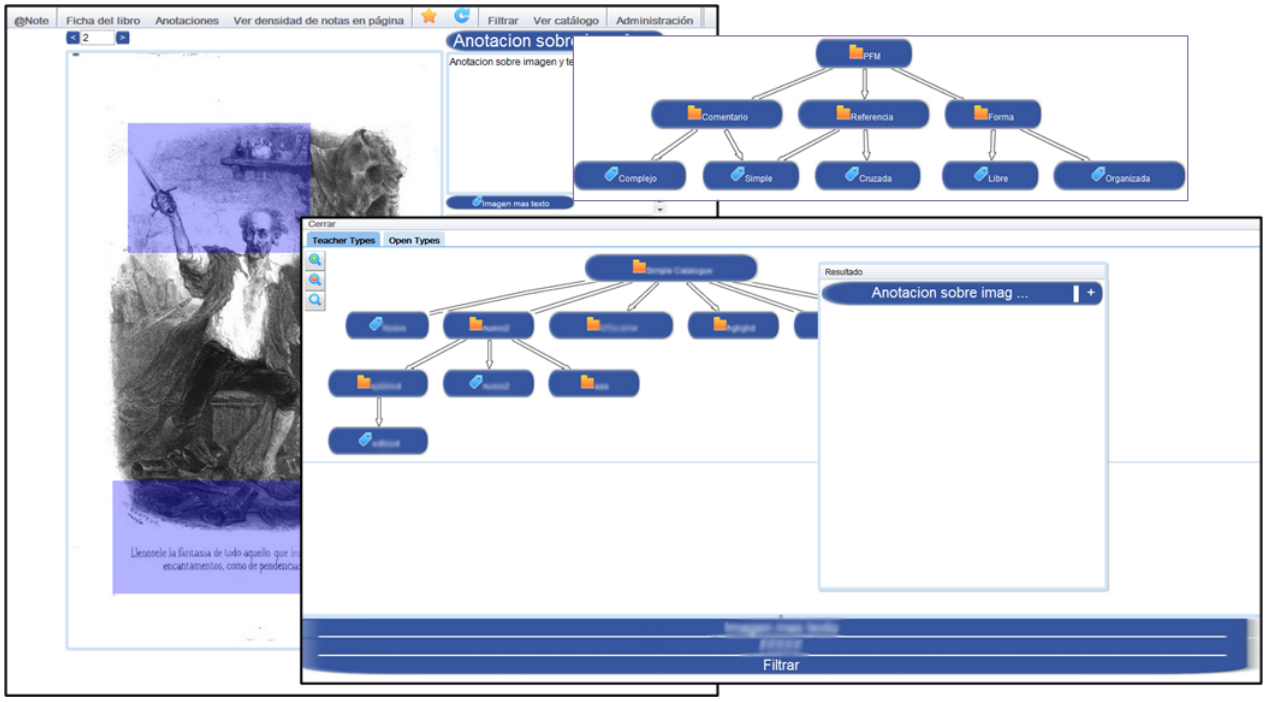


Figura 35 Instancia de la aplicación @note

En @note las taxonomías son creadas y gestionadas por expertos en la obra que se anota, y pueden ser editadas dinámicamente durante la actividad para incluir, eliminar o modificar los términos y su estructura. El conjunto de operaciones de reconfiguración es similar al soportado por ODA, con la salvedad de que en @note se permite fusionar conceptos para resolver problemas de sinonimia. El sistema utiliza índices invertidos para soportar las reconfiguraciones dinámicas y colaborativas de las taxonomías. Este enfoque orientado a la definición colaborativa y a la reconfiguración de taxonomías, junto con su explotación en la ayuda a los expertos en el proceso de reconfiguración, constituyen uno de los resultados de la presente tesis, tal y como se expone en el Capítulo 4.

2.6.5 *Clavy*

El sistema *Clavy* es una evolución del sistema ODA, que aborda el problema de la adaptación de las estructuras de información internas que habilitan la navegación eficiente por las colecciones una vez que los esquemas se reconfiguran. La Figura 36 muestra una instancia de *Clavy*.

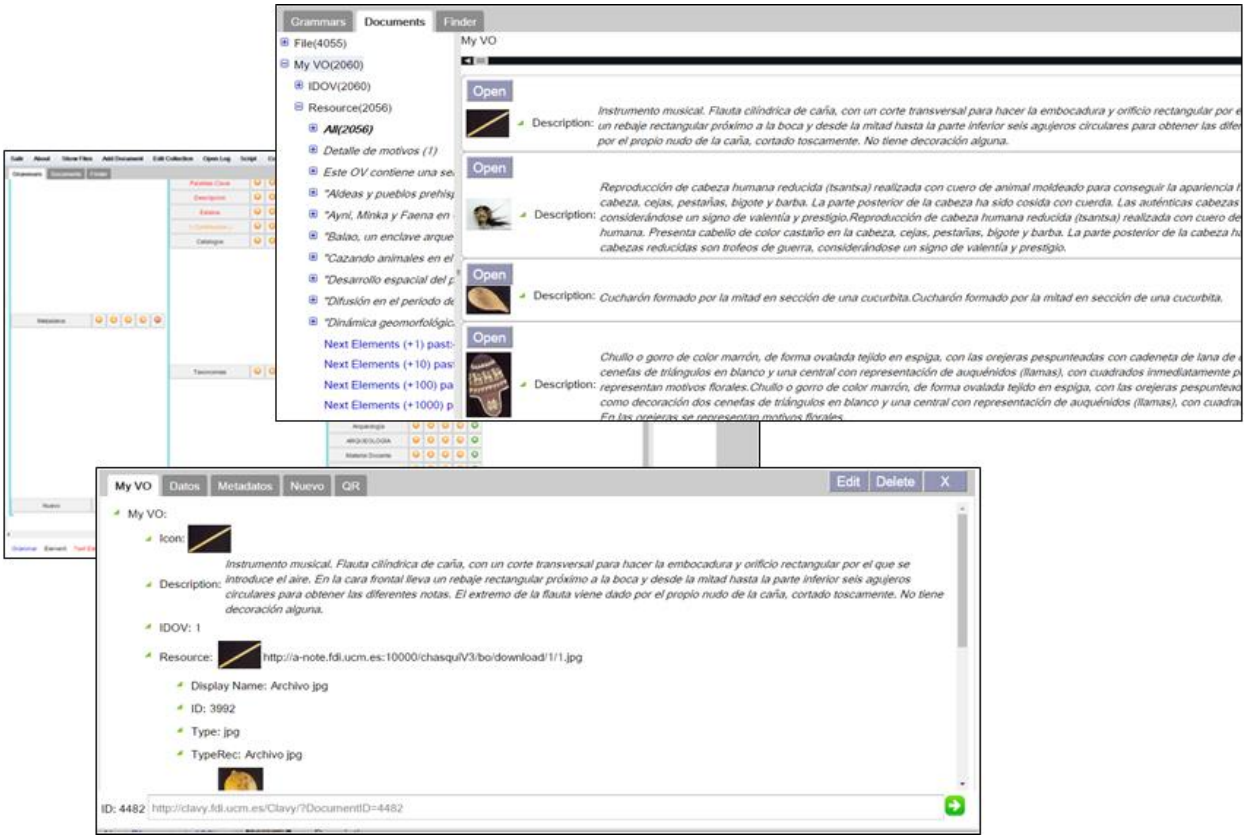


Figura 36. Instancia de Clavy

Clavy incorpora un modelo de catalogación análogo al de ODA, y hereda las operaciones de edición soportadas por dicho sistema. Así mismo, *Clavy* está equipado con una arquitectura de conectores de importación / exportación que facilita la ingesta y exportación de colecciones entre múltiples formatos. Un ejemplo de la funcionalidad brindada por esta arquitectura es la importación / exportación de colecciones ODA a/desde HTML y XLS que se describe en (Gayoso Cabada, Fernández-Pampillón Cesteros, & Sierra Rodríguez, 2015a, 2015b). El sistema *Clavy* también posee un módulo genérico de importación y exportación de objetos y estructura para bases de datos en *MySQL*, un módulo de consulta de colecciones vía servicios a través del estándar de interconexión OAI-PMH, y la definición de un sistema de servicios con lenguaje propio para la consulta y búsqueda sobre la colección de objetos, sus descriptores y la estructura de catalogación de la colección.

Aparte del énfasis en el soporte para la importación / exportación de colecciones, una de las principales aportaciones de *Clavy* frente a ODA estriba en una organización interna más sofisticada que la de los índices invertidos mantenida en ODA, y más orientada a amortiguar eficientemente el impacto de las reconfiguraciones en el rendimiento del sistema. Dicha

organización es otro de los principales resultados de esta tesis, tal y como se describe en los próximos capítulos.

2.7 A modo de conclusión

En este capítulo se ha puesto de manifiesto la enorme importancia que cobran los esquemas de catalogación como ejes básicos de la organización de las colecciones digitales. Al igual que ocurre con los escenarios tradicionales abordados por la disciplina de la Biblioteconomía y la Documentación, la catalogación es esencial para descubrir, acceder a, y, en última instancia, utilizar los recursos integrados en las colecciones. A este respecto, se ha propuesto una multitud de estándares y modelos para esquemas de catalogación, tanto orientados a la catalogación descriptiva (metadatos) como a la catalogación temática (vocabularios) cuyo objetivo último es unificar la catalogación en los repositorios, y posibilitar la interoperabilidad entre los mismos y el resto de sistemas que explotan los objetos digitales. No obstante, también en este capítulo se ha puesto de manifiesto cómo la situación actual en la catalogación de colecciones digitales dista mucho de ser ideal. Efectivamente:

- Por una parte, la existencia de múltiples propuestas, modelos y estándares de catalogación, algunos de ellos genéricos (p.e., Dublin Core), otros muchos de propósito más específico (p.e., LOM, en el dominio educativo), ponen de manifiesto la imposibilidad práctica de alcanzar una solución *universal* a la cuestión de la catalogación de objetos digitales. De esta forma, dotar a las plataformas de ciertas capacidades para la transformación, adaptación y reconfiguración de las catalogaciones es un aspecto inevitable.
- Por otra parte, en muchos dominios es muy complicado adoptar esquemas de catalogación estándar, debido a la alta especificidad de los mismos. En estos dominios es habitual que se formulen mecanismos de catalogación específicamente orientados a cubrir las necesidades concretas que se afrontan durante la producción y preservación de los objetos digitales. Así mismo, en estos dominios es también frecuente que los esquemas de catalogación evolucionen a lo largo del ciclo de vida de las colecciones. De nuevo, por tanto, la necesidad de disponer de mecanismos que permitan abordar la reconfiguración de los esquemas de catalogación se configura como un aspecto crítico.

A este respecto sorprende, por tanto, que el problema de la reconfiguración de colecciones digitales haya recibido una atención marginal en la literatura, frente a otros

aspectos, tales como la formulación y la utilización de estándares para propiciar la interoperabilidad entre sistemas. Para paliar esta carencia, los trabajos llevados a cabo por el Grupo de Investigación en Ingeniería de Lenguajes Software y Aplicaciones (ILSA) han tenido una marcada orientación hacia la reconfiguración de las colecciones en general, y hacia la reconfiguración de los esquemas de catalogación en particular. Estos trabajos ponen de manifiesto que la reconfiguración de colecciones no es un problema trivial. Efectivamente:

- Aparte de contar con mecanismos significativos que permitan reconfigurar los esquemas, también es necesario contemplar en qué forma tales reconfiguraciones afectan al resto de las funcionalidades de los sistemas de gestión de colecciones (p.e., navegación y búsqueda de contenidos en repositorios), y cómo pueden proporcionarse mecanismos que minimicen el impacto en el rendimiento causado por dichas reconfiguraciones. Efectivamente, la operación de los sistemas de gestión de colecciones depende, en gran medida, del modo en que se han creado las estructuras de almacenamiento y recuperación de los objetos a partir de la catalogación de los mismos. Por tanto, si los esquemas de catalogación cambian, será necesario llevar a cabo una reconstrucción, parcial o total, de estos índices, lo que, a su vez, puede repercutir en el rendimiento de los servicios de búsqueda y navegación que dependen de ellos (Ambroziak, 2002; Graefe, 2006; Lim, Wang, Padmanabhan, Vitter, & Agarwal, 2003; Ponnekanti, 2003; Ponnekanti & Kodavalla, 2000) . Si la colección es suficientemente grande este proceso puede resultar en tiempos de respuesta inaceptables para el usuario.
- Por otra parte, otro aspecto especialmente crítico es, una vez aceptado que el proceso de reconfiguración es intrínseco al ciclo de vida de las colecciones, encontrar mecanismos que ayuden a los expertos encargados de crear y preservar las colecciones a explotar de manera efectiva los mecanismos de reconfiguración.

De esta forma, el presente trabajo de tesis se orienta a solventar estos aspectos derivados de la reconfiguración de esquemas de catalogación de colecciones mediante: (i) la búsqueda de mecanismos que permitan orquestar la reconfiguración de distintos tipos de esquemas de catalogación, tanto orientados a metadatos como orientados a vocabularios, (ii) la formulación de enfoques que guíen a los creadores y preservadores de las colecciones en los procesos de reconfiguración, y (iii) la búsqueda de soluciones al impacto que supone en el rendimiento del sistema de gestión de colecciones la reconfiguración de los esquemas de catalogación de las mismas.

Capítulo 3 - Objetivos y Planteamiento del Trabajo

El punto de partida de esta tesis doctoral es los trabajos realizados sobre esquemas de metadatos reconfigurables en la Universidad Complutense de Madrid en el contexto de los sistemas ya descritos en el capítulo anterior. Dicho trabajo, que se inició en 2001, se ha llevado a cabo en el contexto de distintos proyectos del Plan Nacional de I+D+i: SIMBA (Sistema de Información Multimedia Basado en Lenguajes de Marcado y Agentes, TIC2000-0737-C03-01), REI-MLH (Recursos Educativos e Informativos basados en componentes distribuidos: Metodología, Lenguajes y Herramientas, TIC2002-04067-C03-02), OdAVirtual (Objetos de Aprendizaje en el Campus Virtual, TIN2005-08788-C04-01) y GENHOE-VIRTUAL (Un Enfoque Generativo para la Construcción de Herramientas de Producción y Despliegue de Objetos Educativos en el Campus Virtual, TIN2010-21288-C02-01). Dicho trabajo ha desembocado en el modelo de metadatos incluido en el sistema ODA al que se ha hecho referencia en el capítulo anterior, así como en propuestas como las descritas en (J. L. Sierra & Fernández-Valmayor, 2008), también referidas en el capítulo anterior. Estas propuestas propugnan una definición colaborativa de esquemas de metadatos por parte de los expertos en un determinado dominio asociado al problema, explícitamente orientados a satisfacer las necesidades expresivas en dichos dominios. Este enfoque difiere, y en cierta medida complementa, al sostenido por los enfoques basados en esquemas de metadatos más estandarizados, enfoques también descritos en el capítulo anterior. Así mismo, y con el fin de permitir una definición social e iterativa de estos esquemas, es fundamental que el proceso de creación de los mismos permita su reconfiguración, a fin de resolver las posibles discrepancias entre los distintos expertos, o las posibles limitaciones detectadas durante la creación de las colecciones desde cero o mediante la incorporación, o reutilización, de distintas colecciones ya existentes. Tal y como ya se ha indicado en el capítulo previo, las propuestas resultantes de los trabajos realizados contemplan un rico repertorio de operaciones de reconfiguración, tales como el renombrado de los elementos de información de los esquemas, la reorganización de sus estructuras, la fusión de atributos para resolver problemas de sinonimia, la creación de nuevos elementos y la eliminación de elementos no utilizados, etc. No obstante, todos estos trabajos adolecen de las siguientes limitaciones:

- Los trabajos aludidos se centran fundamentalmente en esquemas de metadatos, ignorando otro tipo de esquemas de catalogación, como los basados en vocabularios controlados (taxonomías, tesauros, ontologías, etc.). Si bien la propuesta de ODA contempla el uso de vocabularios controlados, entendidos como conjuntos planos de

términos que pueden ser compartidos por distintos *atributos* (los elementos de información básicos en los esquemas de metadatos ODA), dicha propuesta no se ocupa en modo alguno de la reconfiguración de dichos vocabularios.

- Aunque los modelos propuestos aportan los mecanismos necesarios para llevar a cabo la reconfiguración de los esquemas, estos modelos no proponen mecanismos concretos que ayuden a los expertos a planificar de manera apropiada dichas operaciones de reconfiguración. Efectivamente, sería conveniente disponer de mecanismos que permitieran, por ejemplo, recabar información ejecutiva que ayude a los expertos a decidir cuándo hay elementos de información equivalentes entre sí, elementos de información en desuso, elementos erróneos, organizaciones alternativas de las estructuras de los esquemas, etc. Los trabajos realizados no aportan ningún mecanismo de este tipo, limitándose a proponer modelos básicos de edición colaborativa de los esquemas.
- Por otra parte, en los trabajos realizados no se ha abordado con suficiente profundidad la forma de minimizar el impacto que las reconfiguraciones en los esquemas pueden tener sobre las distintas funcionalidades del sistema de gestión de colecciones, ya que dichas funcionalidades dependerán en mayor o menor medida de indexaciones de los objetos digitales basadas en dichos esquemas.

En esta tesis, por tanto, se abordan estas limitaciones. Para ello se aborda, por una parte, la reconfiguración de los tipos más habituales de esquemas de catalogación, tanto esquemas de metadatos como distintos tipos de vocabularios controlados. Por otra parte, se aborda la definición de mecanismos que ayuden a los expertos en el dominio en el proceso de reconfiguración. Por último, se aborda el diseño de mecanismos que permitan a los sistemas de gestión de colecciones adaptarse mejor a las reconfiguraciones en los esquemas. El trabajo de tesis se ha llevado a cabo en el contexto de los siguientes proyectos de investigación:

- El proyecto Collaborative Annotation of Digitalized Literary Text del Google's Digital Humanities Award Program 2010.
- El proyecto MUGECODER (Modelo Unificado de Gestión de Colecciones Digitales con Estructuras Reconfigurables: Aplicación a la Creación de Bibliotecas Digitales Especializadas para Investigación y Docencia, HUM14_251), del Programa de Ayudas de la Fundación BBVA a Proyectos de Investigación (Convocatoria 2014).

- El proyecto RedR+Human (*Repositorios Educativos Dinámicamente Reconfigurables en Humanidades*, TIN2014-52010-R) del subprograma de Investigación Orientada a los Retos de la Sociedad, convocatoria 2014 del Plan Nacional de I+D+i.

En este capítulo se presentan los objetivos que se pretenden alcanzar en la tesis planteada (apartado 3.1), así como el plan de trabajo seguido para alcanzar dichos objetivos (apartado 3.2). El capítulo termina con un apartado de recapitulación (apartado 3.3).

3.1 Objetivos de la tesis

Tal y como se ha indicado anteriormente, el trabajo de esta tesis toma como punto de partida la investigación previa sobre colecciones digitales con esquemas de metadatos reconfigurables llevada a cabo en la Universidad Complutense de Madrid, que ha cristalizado en los sistemas ya descritos en el capítulo anterior. Así mismo, y tal y como también se ha indicado, básicamente dicha investigación ha convergido en modelos de edición colaborativa de esquemas de metadatos específicamente orientados a cada dominio concreto de aplicación. Este hecho puede constatarse, por ejemplo, en el sistema ODA, en el que, tal y como ya se ha mostrado en el capítulo anterior, se proporciona un editor de esquemas de metadatos que permite editar los distintos atributos ODA, organizar jerárquicamente dichos atributos, y gestionar toda esta información. Dicha gestión contempla, entre otras operaciones, la reorganización de las jerarquías, en el sentido de que, en cualquier momento, es posible cambiar la filiación de los atributos, modificando sus padres a atributos, no únicamente del mismo esquema, sino de cualquiera de los esquemas utilizados en la catalogación de las colecciones. No obstante, y tal y como ya se ha indicado anteriormente, dichas investigaciones no se han centrado en otro tipo de esquemas de catalogación, tales como los distintos tipos de vocabularios controlados analizados en el capítulo anterior. Así mismo, tampoco han propuesto herramientas que faciliten a los expertos decidir qué reconfiguraciones realizar. Por último, tampoco ha ahondado en la forma de minimizar el impacto causado por las reconfiguraciones en las distintas funcionalidades del sistema de gestión, utilizando, en su lugar, soluciones más o menos genéricas (como, por ejemplo, la estrategia de índices invertidos implementada en ODA). Si bien para colecciones de tamaño pequeño dicho impacto puede abordarse de manera aceptable con soluciones genéricas como las ya aludidas, conforme dicho tamaño aumenta se corre el riesgo de que el rendimiento del sistema se degrade hasta límites inaceptables. Es, por tanto, en estos tres aspectos (el abordar otros modelos de catalogación, el asistir a los expertos

en la reconfiguración y el proporcionar infraestructuras eficientes para soportar dicha reconfiguración) en los que se centran los objetivos planteados en esta tesis. Más concretamente, la tesis propone tres objetivos básicos:

1. El primer objetivo aborda la reconfiguración de distintos tipos de esquemas de catalogación, tanto esquemas de metadatos, como vocabularios controlados.
2. El segundo objetivo aborda la concepción de mecanismos que ayuden a los expertos a planificar las reconfiguraciones de los esquemas para adaptar los mismos a las necesidades específicas de catalogación de las colecciones.
3. El tercer objetivo se centra en la concepción de infraestructuras para los sistemas de gestión de colecciones que permitan amortiguar adecuadamente el impacto que las reconfiguraciones en los esquemas de catalogación causan en las diferentes funcionalidades.

Las siguientes secciones profundizan en cada uno de estos aspectos, a fin de sentar las bases necesarias para plantear el trabajo de tesis, planteamiento que se desarrolla en el siguiente apartado.

3.1.1 *Esquemas de catalogación reconfigurables*

La primera limitación detectada en relación con el trabajo realizado hasta el momento en el Grupo de Investigación de la Universidad Complutense de Madrid en el que se enmarca la presente tesis se refiere a que la reconfiguración se circunscribe básicamente a esquemas de metadatos. De esta forma, el propósito del primer objetivo planteado en esta tesis está orientado a paliar dicha limitación, mediante el estudio de la reconfiguración de otro tipo de esquemas de catalogación, tanto otros esquemas de metadatos, como esquemas de catalogación basados en vocabularios controlados.

Para llevar a cabo la reconfiguración se aplicarán, en todos los casos, operaciones básicas de reconfiguración, del estilo de las contempladas en sistemas como ODA. Dado que estas son, fundamentalmente, operaciones de edición de jerarquías, la reconfiguración se abordará a un nivel fundamentalmente *sintáctico*, con el fin de mantener el enfoque lo más independiente posible de los sesgos introducidos por los distintos modelos de catalogación. Desde este punto de vista, el enfoque se orientará principalmente a la reconfiguración de la *estructura* de los esquemas, en lugar de a las *semánticas* específicas de los mismos.

De esta forma, el primer objetivo planteado en esta tesis se concreta como sigue:

Objetivo 1. Abordar la reconfiguración de los principales tipos de esquemas de catalogación, incluyendo esquemas de metadatos y distintos tipos de vocabularios controlados, mediante un repertorio básico de operaciones de reconfiguración.

3.1.2 *Soporte a los expertos en el dominio para la reconfiguración*

Tal y como ya se ha comentado anteriormente, otra limitación palpable del trabajo sobre esquemas de metadatos reconfigurables realizado hasta el momento en el Grupo de Investigación se refiere a que, si bien las propuestas realizadas proporcionan un conjunto completo de herramientas para editar y reconfigurar esquemas, dichas propuestas no abordan en ningún momento la forma en la que se puede ayudar al experto a identificar qué aspectos de los esquemas necesitan realmente ser reconfigurados. Este hecho es especialmente crítico en dominios específicos de naturaleza no necesariamente técnica, en los cuáles los expertos normalmente tienen ideas muy precisas sobre cómo organizar las colecciones de objetos digitales, pero muchas veces no poseen las competencias técnicas necesarias para formalizar, en un primer intento, dichas ideas en esquemas de catalogación. Por tanto, necesitan contar con mecanismos iterativos de refinamiento de tales esquemas, que les permitan mejorar paulatinamente los mismos y que les ayuden a planificar, en cada etapa de mejora, las reconfiguraciones más convenientes.

En esta tesis se considera que la forma de proporcionar información útil a los expertos en el proceso de reconfiguración es mediante un *análisis de la propia actividad de catalogación de la colección*. A este respecto, se plantea abordar este segundo objetivo mediante un análisis de la forma en la que los esquemas se utilizan *realmente* en la organización de las colecciones, analizando los usos concretos que se han hecho de dichos esquemas. El resultado de dicho análisis permitirá inferir estructuras de alto nivel comparables con las utilizadas por el modelo de esquema propuesto. Este hecho permitirá a los expertos comparar la estructura originalmente propuesta en el esquema con la que realmente emerge del uso de dicho esquema en la organización de la colección. Como resultado de dicha comparación, los expertos podrán planificar las reconfiguraciones que estimen más adecuadas. Efectivamente, los expertos podrán identificar discrepancias entre la organización propuesta y la organización inferida de la colección, así como planificar las formas más adecuadas de resolver dichas discrepancias mediante herramientas de edición análogas a las proporcionadas por sistemas tipo ODA. Con el fin de ajustar el esfuerzo requerido por este objetivo a los tiempos de investigación de una

tesis doctoral, la propuesta realizada en este objetivo se restringirá a un tipo particular de esquema de catalogación: *vocabularios taxonómicos*. El motivo de esta elección es doble:

- Por una parte, las taxonomías son uno de los tipos de vocabularios controlados más frecuentes en la catalogación de colecciones.
- Por otra parte, será posible aprovechar el contexto del proyecto *Collaborative Annotation of Digitalized Literary Texts*, donde surgen actividades de catalogación basadas en taxonomías de conceptos muy específicas y muy bien definidas, por lo que el análisis de dichas actividades cobra pleno sentido.

Por tanto, el segundo objetivo de esta tesis se concreta como:

Objetivo 2. Proponer herramientas de análisis automático de las colecciones digitales que permitan inferir organizaciones conceptuales de alto nivel en las que se refleje el uso actual de los esquemas de catalogación, y que proporcionen a los expertos la base necesaria para llevar a cabo el refinamiento de dichos esquemas mediante herramientas de reconfiguración.

3.1.3 *Mecanismos para la adaptación de las plataformas a las reconfiguraciones*

La última limitación detectada se refiere a la adaptación de las plataformas a las reconfiguraciones realizadas por los expertos en los esquemas. Efectivamente, tal y como se ha discutido en el capítulo anterior, las distintas funcionalidades ofrecidas por las plataformas de gestión de colecciones, tales como la navegación o la búsqueda, utilizan normalmente los esquemas como guía de referencia básica. Así, por ejemplo, y como se ha puesto también de manifiesto en el capítulo anterior, la navegación guiada sobre las colecciones se lleva a cabo en términos de las distintas categorías, conceptos y términos integrados en los esquemas, que se utilizan para formular criterios que permitan acotar el espacio de información. Igualmente, las búsquedas utilizan tales elementos como criterios básicos de filtrado. De esta forma, para soportar de manera eficiente dichas funcionalidades, es necesario indexar de manera apropiada los objetos de las colecciones. La solución de indexación óptima se puede llevar a cabo, de este modo, en términos de la estructura del esquema. Así, por ejemplo, si se consideran taxonomías como modelos de esquema, el índice puede consistir en la propia jerarquía de conceptos contemplados en la taxonomía, asociando a cada nodo el conjunto de objetos seleccionados. Otros tipos de modelos de esquema conducirán a otras soluciones de indexado, pero, de cualquier manera, la estructura del esquema jugará un papel primordial en dicho indexado.

El problema surge, entonces, cuando se introduce la posibilidad de reconfigurar los esquemas. Efectivamente, si se realizan cambios en dichos esquemas, estos cambios afectarán dramáticamente a las estructuras de datos internas en las que se sustentan las funcionalidades de las plataformas. Efectivamente, dado que la indexación óptima depende de la estructura del esquema, si dicha estructura se altera *puede invalidarse parcial o completamente los índices internos mantenidos por el gestor de colecciones*. Con ello, en el peor de los casos será necesario reconstruir totalmente dichos índices. Incluso para colecciones de tamaño moderado dicho proceso puede consumir una cantidad apreciable de tiempo (Heinz & Zobel, 2003; Lester, Zobel, & Williams, 2004). Como resultado, el usuario puede enfrentarse a colecciones temporalmente desactualizadas o fuera de servicio, en el supuesto de que la reconfiguración de las estructuras de datos internas que soportan las funcionalidades se lleve a cabo fuera de línea. Por otra parte, el sistema puede arrojar tiempos de respuesta sub-óptimos que pueden impactar negativamente en la experiencia de usuario, en caso de que se empleen soluciones de indexado genéricas (como, por ejemplo, los índices invertidos utilizados en ODA), ya que el rendimiento de dichas soluciones puede irse degradando conforme aumenta el tamaño de las colecciones.

A este respecto, surge la necesidad de buscar estrategias de indexado que, por una parte, sean invariantes a los cambios en los esquemas, y, por otra, proporcionen un rendimiento aceptable (idealmente cercano al óptimo) en lo que se refiere a las funcionalidades básicas. De esta forma, el tercer objetivo planteado en esta tesis se orienta hacia la búsqueda de dichas estrategias. Dado que, también en este caso, la resolución total de esta problemática puede exceder con mucho los tiempos de investigación de una tesis doctoral, el trabajo en este objetivo se centrará en una de las funcionalidades básicas: la navegación guiada por el esquema de catalogación sobre las colecciones. Dicha elección se debe a que la navegación guiada es una de las funcionalidades más básicas del sistema de gestión, el repositorio, y que, por tanto, otras funcionalidades (por ejemplo, las búsquedas) pueden expresarse en términos de los servicios que soportan la misma[68]. De esta forma, se considera que la resolución de los problemas relativos a la navegación guiada puede suponer un factor clave para la resolución del resto de los problemas asociados a funcionalidades de más alto nivel.

[68] Efectivamente, en última instancia una búsqueda basada en una consulta booleana puede descomponerse en una colección de navegaciones normalizando dicha consulta en forma normal conjuntiva, y encontrando tratamientos apropiados para los átomos negados.

Como consecuencia, el tercer objetivo planteado en esta tesis se concreta como:

Objetivo 3. Proponer estrategias de indexación de colecciones de objetos digitales basadas en esquemas de catalogación dinámicamente reconfigurables que produzcan índices que permanezcan invariantes a las distintas operaciones de reconfiguración soportadas por dichos esquemas y que, al mismo tiempo, ofrezcan rendimientos aceptables, idealmente próximos a los óptimos.

3.2 Planteamiento del trabajo

Teniendo en cuenta las consideraciones realizadas en el anterior apartado, así como los objetivos propuestos en el mismo, es posible afirmar que el propósito de esta tesis es realizar avances en la propuesta de modelos para la gestión de colecciones digitales con esquemas de catalogación reconfigurables dinámicamente, de forma que se aborden las limitaciones identificadas en los trabajos previos desarrollados por el Grupo de Investigación: por una parte las relativas al alcance de las propuestas ya realizadas (que se restringen a esquemas de metadatos), por otra parte las referentes al apoyo ofrecido a los expertos para reconfigurar los esquemas, y por otra las relativas a la eficiencia en los sistemas de gestión frente a las reconfiguraciones.

De esta forma, de los objetivos planteados en el apartado 3.1 , se derivan una serie de actividades específicas orientadas a la consecución de los mismos, que permiten estructurar el trabajo llevado a cabo en esta tesis. A continuación, se describen dichas actividades.

3.2.1 *Actividades relativas a los esquemas de catalogación reconfigurables*

Para lograr el objetivo 1 se abordará el problema de la reconfiguración de los esquemas de catalogación utilizados más habitualmente en la organización de colecciones digitales, tanto esquemas de metadatos como vocabularios controlados, utilizando, para ello, operaciones análogas a las usadas en sistemas como ODA. De esta forma, el trabajo a realizar en relación con dicho objetivo se desgranará en las dos actividades específicas que se describen a continuación.

3.2.1.1 Reconfiguración de esquemas de metadatos

Esta actividad se centra en el trabajo que se está desarrollando en el contexto del proyecto RedR+Human, del Plan Nacional de I+D+i. En este proyecto se está desarrollando un modelo de gestión de repositorios específicos de objetos educativos con estructuras

reconfigurables. Uno de los objetivos centrales del proyecto es diseñar un modelo de esquema de metadatos reconfigurable que permita también abordar de forma eficiente las reconfiguraciones de cara a las distintas funcionalidades proporcionadas por la plataforma de gestión (navegación, búsqueda, etc.).

De esta forma, esta actividad se centra en torno al diseño de la versión inicial del modelo de esquemas de metadatos reconfigurables de RedR+Human, así como en la caracterización de las operaciones de reconfiguración soportadas por dicho modelo. En relación con los requisitos básicos que debe cumplir el modelo se plantea, en primer lugar, que éste debe ser lo suficientemente simple como para permitir a los expertos en el dominio editar por sí mismos los esquemas de metadatos. Así mismo, el modelo debe soportar mecanismos básicos de reconfiguración de los esquemas, utilizando operaciones de reconfiguración tipo ODA, tales como renombrado de elementos y reorganización estructural, ya que la utilidad de dichas operaciones está ya ampliamente contrastada en las experiencias de uso previas de ODA en distintos dominios (Arqueología[69], Filología[70] [71] [72], Educación[73], Biología[74], Física[75], Historia[76] etc.). Igualmente, el modelo debe mantenerse lo suficientemente agnóstico como para poder acomodar una variedad suficientemente amplia de modelos de metadatos. Por último, el modelo debe tener una contrapartida operacional adecuada, permitiendo tratar de manera razonablemente eficiente el impacto causado por las reconfiguraciones realizadas en la estructura interna de los repositorios.

3.2.1.2 Reconfiguración de vocabularios controlados

Además del problema de la reconfiguración de esquemas de metadatos, un aspecto básico en la consecución del objetivo 1 es abordar la reconfiguración del otro tipo básico de esquemas de catalogación: vocabularios controlados. Para ello se plantea una actividad

[69] http://oda-fec.org/ucm-chasqui/

[70] http://repositorios.fdi.ucm.es/ciberia_oda/

[71] http://repositorios.fdi.ucm.es/DiccionarioDidacticoLatin/

[72] http://repositorios.fdi.ucm.es/DiccionarioDidacticoAleman/

[73] http://repositorios.fdi.ucm.es/Tropos/

[74] http://repositorios.fdi.ucm.es/especiesParqueGuadarrama/

[75] http://repositorios.fdi.ucm.es/Fisicas/

[76] http://repositorios.fdi.ucm.es/Mythos/

orientada a abordar dicho aspecto. Más concretamente, en esta actividad se propone realizar las siguientes experiencias de reconfiguración de vocabularios controlados:

- Reconfiguración de *folksonomías*. Se propondrá la organización colaborativa de las etiquetas que conforman una folksonomía en categorías que organicen dichas etiquetas, así como la reestructuración colaborativa de dichas etiquetas. Esta experiencia se llevará a cabo en el contexto del proyecto MUGECODER.
- Reconfiguración de *taxonomías*. Más concretamente, se abordará la reconfiguración de las taxonomías para la catalogación de anotaciones de textos literarios digitalizados que surgen en el proyecto *Collaborative Annotation of Digitalized Literary Texts*.
- Reconfiguración de *tesauros facetados*. También en el contexto del proyecto MUGECODER se abordará el problema de la reconfiguración de tesauros facetados utilizados para organizar colecciones de objetos digitales en el campo de las Humanidades. Con ello se obtendrán tesauros facetados dinámicamente reconfigurables.

3.2.2 *Actividades relativas al soporte a los expertos para la reconfiguración*

El desarrollo del objetivo 2 se plantea también en términos de dos actividades, la primera de ellas orientada a la formulación del modelo de inferencia de la estructura de catalogación a partir de la colección digital, y la segunda orientada a la validación inicial de la propuesta. A continuación, se detallan cada una de estas actividades.

3.2.2.1 Desarrollo del modelo de inferencia de estructuras de catalogación

El aspecto crítico en relación con el objetivo 2 es concebir un modelo que permita inferir las estructuras conceptuales asociadas al uso de los esquemas durante las actividades de catalogación de las colecciones. Efectivamente, utilizando dicho modelo, será posible ofrecer al experto una estructura de alto nivel comparable con la del esquema original. A partir de dicha estructura, el experto podrá descubrir potenciales limitaciones y, por tanto, ingeniar las reconfiguraciones necesarias.

Para llevar a cabo esta actividad se deberá aplicar técnicas que permitan inferir estructuras conceptuales y relaciones jerárquicas entre dichas estructuras, con el fin de adecuarse a las necesidades del objetivo, permitiendo la comparación entre los esquemas originales y los inferidos. Como ya se ha indicado anteriormente, a fin de mantener la dificultad de esta tarea dentro de límites razonables, se abordará únicamente el caso específico de las

taxonomías para la anotación de textos literarios digitalizados, en el contexto del proyecto financiado por Google. No obstante, se espera que el método aplicado y los resultados obtenidos puedan extrapolarse a otros tipos de esquemas de catalogación.

3.2.2.2 Validación inicial de la propuesta de soporte a la reconfiguración

El modelo de inferencia de estructuras de catalogación permitirá llevar a cabo la propuesta de soporte a la reconfiguración planteada en el objetivo 2. En esta actividad se propone, por tanto, validar la adecuación del enfoque mediante su aplicación en un caso de estudio real, en el ya citado escenario de la anotación de textos literarios digitalizados.

El caso de estudio elegido involucrará tanto expertos en el dominio, encargados de diseñar los esquemas, como catalogadores, encargados de catalogar adecuadamente los objetos de la colección. De esta forma, los expertos podrán aplicar la herramienta de inferencia de estructuras propuesta en la actividad anterior a las colecciones producidas por los catalogadores para inferir estructuras conceptuales con una organización jerárquica. Dichas jerarquías se podrán comparar, entonces, con el esquema propuesto para permitir a los expertos llevar a cabo las reconfiguraciones que consideren oportunas.

3.2.3 *Actividades relativas a la adaptación de las plataformas a las reconfiguraciones*

Para lograr el objetivo 3 es necesario formular, primeramente, un modelo conceptual de organización interna de la colección que facilite de manera eficiente la navegación guiada por dicha colección (remarcar, de nuevo, que el alcance del objetivo 3 se ha restringido a la funcionalidad de navegación), y que, a la vez, sea invariante a las posibles reconfiguraciones de los esquemas de catalogación. Seguidamente es necesario buscar mecanismos prácticos de indexación que permitan realizar dicho modelo. Dichas consideraciones se plasman en las dos actividades que se indican a continuación.

3.2.3.1 Formulación del modelo de navegación

Esta actividad está orientada a formular un modelo conceptual de organización de la colección que soporte eficientemente la navegación guiada por los esquemas de catalogación. Así mismo, dicho modelo no debe verse afectado por las distintas reconfiguraciones en el esquema.

Para soportar este modelo se propone aplicar un enfoque *dirigido por lenguajes*. De esta forma, las posibles interacciones por el sistema durante la navegación guiada se entenderán

como un lenguaje formal, y la organización de la colección surgirá como un reconocedor apropiado para dicho lenguaje. Los estados de dicho reconocedor se identificarán con posibles estados en la navegación, mientras que los movimientos se identificarán con selección de filtros por parte del usuario para restringir el conjunto de objetos seleccionados.

3.2.3.2 Desarrollo de métodos prácticos de indexación

Esta actividad se orientará, por último, a buscar mecanismos prácticos que permitan representar el modelo conceptual de organización de la colección. Efectivamente, dado que el conjunto posible de estados de navegación puede ser muy grande, será necesario aplicar estrategias que permitan reducir los requisitos de espacio exigido.

De esta forma, se propondrán métodos alternativos para indexar las colecciones, así como mecanismos que permitan recrear dinámicamente el modelo conceptual de organización durante la navegación. Para tal fin, se buscará un equilibrio entre la eficiencia en la navegación y el espacio requerido por los índices, y se realizará una comparativa entre las distintas soluciones propuestas, a fin de permitir elucidar la más conveniente de cara a dicho equilibrio.

3.3 A modo de conclusión

En esta tesis se abordan distintos aspectos relativos a la gestión de colecciones digitales con esquemas de catalogación dinámicamente reconfigurables. A este respecto, y en base a las investigaciones llevadas a cabo por el Grupo de Investigación en el que se integra este trabajo de tesis, se han propuesto objetivos orientados a abordar tres de los aspectos básicos que surgen durante dicha gestión:

1. La reconfiguración de esquemas de catalogación que se adecuen a los dos estilos de catalogación de colecciones que, tal y como se ha indicado en la Sección 3.2.1 , resultan más habituales: la catalogación basada en metadatos, y la basada en vocabularios controlados. De esta forma, se propone abordar la reconfiguración de distintos modelos en ambas categorías, utilizando un repertorio básico de operaciones de reconfiguración basadas en las utilizadas en los trabajos previos en el Grupo sobre esquemas de metadatos reconfigurables.
2. La formulación de una estrategia orientada a facilitar la reconfiguración de esquemas de catalogación por parte de los expertos en el dominio. Dicha estrategia se apoyará principalmente en un modelo de análisis de las colecciones catalogadas mediante el esquema a reconfigurar, modelo que permitirá inferir una estructura organizativa

comparable con el esquema. Los expertos podrán, entonces, comparar ambas estructuras (la originalmente representada en el esquema, y la inferida mediante el modelo de análisis) y proponer las reconfiguraciones necesarias para acercar el esquema a la realidad de su uso. Aunque el trabajo en este sentido se restringirá a las taxonomías, se espera que el método pueda ser extrapolado en el futuro a otros tipos de esquemas de catalogación.

3. La formulación de mecanismos de indexación de las colecciones que permitan producir índices invariantes ante las reconfiguraciones en los esquemas. En concreto, estos mecanismos permitirán una navegación guiada eficiente por las colecciones, incluso cuando los esquemas de catalogación se modifiquen, todo ello de forma dinámica, sin necesidad de realizar fuera de línea costosos procesos de re-indexado o mantener temporalmente la incoherencia entre las estructuras internas de organización y los esquemas de catalogación.

De esta forma, en el siguiente capítulo se presentan y discuten los artículos que se adjuntan a la presente tesis, con el fin de integrar de manera apropiada sus contenidos y de relacionar los mismos con los objetivos planteados en este capítulo. Así mismo, en el Capítulo 5 se analiza el grado de cumplimiento de los objetivos planteados respecto a los resultados presentados en los artículos, y se describen también las líneas de trabajo futuro que emergen del trabajo realizado.

Capítulo 4 - Discusión de las Contribuciones de los Artículos

Este capítulo contextualiza los resultados de investigación obtenidos en esta tesis en base a las publicaciones editadas que la integran. El apartado 4.1 contextualiza los resultados relativos a los esquemas de catalogación reconfigurables (objetivo 1 de la tesis). El apartado 4.2 contextualiza los resultados relativos al enfoque seguido para apoyar a los expertos en el dominio en la reconfiguración de los esquemas (objetivo 2 de la tesis). El apartado 4.3 resume los resultados relativos a los mecanismos que permiten a los sistemas de gestión de colecciones adaptarse a las reconfiguraciones (objetivo 3 de la tesis). Por el último, el apartado 4.4 cierra el capítulo.

4.1 Esquemas de catalogación reconfigurables

Como primer objetivo de esta tesis se plantea abordar el problema de la reconfiguración de los principales tipos de esquemas de catalogación, tanto esquemas de metadatos, como vocabularios controlados. Este apartado describe los trabajos realizados en esta tesis en relación con este objetivo.

De acuerdo con el planteamiento de trabajo realizado en el capítulo anterior, este objetivo se desarrolla en dos actividades diferentes. Los trabajos realizados en relación con la primera de estas actividades, que está dirigida a la reconfiguración de esquemas de metadatos, se describen en la sección 4.1.1 . Por su parte, en la sección 4.1.2 se describen los trabajos relativos a la segunda actividad, orientada a la reconfiguración de distintos tipos de vocabularios controlados.

4.1.1 *Reconfiguración de esquemas de metadatos*

El trabajo llevado a cabo en relación con la reconfiguración de esquemas de metadatos toma como puntos de partida:

- *El modelo de definición de esquemas de metadatos de* ODA. Como ya se ha comentado en los capítulos previos, básicamente dicho modelo concibe los esquemas como jerarquías de atributos de distintos tipos (atributos de tipo texto, atributos con valores controlados, etc.). Además, dicho modelo permite reconfigurar estructuralmente las jerarquías, modificando la filiación de los atributos (es decir, cambiando sus padres). Los esquemas definidos en ODA no tienen una semántica específica, sino que se limitan a diferenciar de manera adecuada los distintos atributos que participan en el esquema, mediante la elección de nombres apropiados y el establecimiento de relaciones

jerárquicas entre los mismos. Dichas relaciones son meramente estructurales (es decir, no tienen carga semántica, a diferencia de otros enfoques, en los que existen relaciones de especialización, relaciones de agregación, etc.).

- *El modelo subyacente a los lenguajes de marcado generalizados*. Tal y como se ha comentado en el Capítulo 2, dichos lenguajes (lenguajes basados en XML) norman como definir y representar las estructuras de documentos. De forma similar al caso anterior, este tipo de lenguajes conciben la estructura de los documentos como jerarquías de *elementos* y tienen un carácter meramente descriptivo (se limitan a diferenciar estructuralmente los documentos, sin reflejar, en ningún momento, cuestiones semánticas). De hecho, la semántica aparece posteriormente mediante artefactos que aprovechan la diferenciación estructural realizada en los documentos (nombres de etiquetas en los elementos, relaciones *padre-hijo* entre las etiquetas, etc.) para añadir procesamiento.

De esta forma, para analizar la reconfiguración de esquemas de metadatos se ha comenzado abstrayendo un modelo agnóstico de definición de este tipo de esquemas de metadatos centrado en la definición de estructuras jerárquicas (árboles). Utilizando la nomenclatura de los lenguajes de marcado generalizado, los nodos en estas estructuras se denominan *elementos*. Así mismo, haciendo uso de la distinción realizada en ODA, el modelo distingue dos tipos básicos de elementos:

- Elementos *descriptivos*. Estos elementos tienen asociados valores en las catalogaciones de los objetos.
- Elementos *estructurales*. Estos elementos no tienen asociados valores, sino que tienen un carácter booleano (es decir, aparecen o *no* aparecen en la catalogación).

Así pues, la catalogación de un objeto digital con un esquema de metadatos definido por el modelo consiste en la confección de un *documento* estructurado que:

- Tendrá la estructura especificada en el esquema de metadatos. Por tanto, el esquema definido puede entenderse como la definición de la estructura jerárquica seguida por todos los documentos que catalogan los objetos.
- Asignará valores apropiados a cada uno de los elementos descriptivos. Dichos valores diferenciarán los documentos asignados a objetos diferentes.

En línea con el enfoque seguido en ODA, es importante indicar que, en la confección de un documento de catalogación, no es necesario utilizar toda la estructura representada en el

esquema, sino seleccionar únicamente una sub-jerarquía de la misma (dicha sub-jerarquía deben contener, en cualquier caso, los elementos raíz).

Dado que todos los documentos que catalogan los objetos siguen la estructura común del esquema, no es necesario representar explícitamente dicha estructura en dichos documentos, sino que basta con representar la asignación de valores a aquellos elementos descriptivos que se consideren relevantes, y seleccionar los elementos estructurales oportunos. En este sentido, los ancestros descriptivos de dichos elementos tomarán, como valor por defecto, el valor *indefinido*, y los ancestros estructurales simplemente se anexarán al documento.

Asímismo, y de forma similar a ODA, el modelo permite reconfigurar los esquemas mediante la reorganización jerárquica de los elementos. Con el fin de evitar tener que replicar dichas reconfiguraciones en cada uno de los documentos, los documentos en sí se conciben como *tablas* que asignan valores a los elementos descriptivos relevantes. De esta forma, cuando se accede a los mismos, el esquema se utiliza para recuperar sus estructuras.

El modelo presentado, se describe en (Joaquín Gayoso-Cabada et al., 2016b) y ha sido implementado completamente en la herramienta *Clavy* ya aludida en el capítulo anterior.

4.1.2 *Reconfiguración de vocabularios controlados*

En relación con la reconfiguración de vocabularios controlados, se ha abordado el problema para tres tipos diferentes de tales vocabularios, de complejidad creciente: listas de términos, taxonomías y tesauros facetados (ver sección 2.3.4). A continuación, se detalla cada uno de estos aspectos.

4.1.2.1 Listas de términos

El estilo de catalogación basada en vocabularios controlados más sencillo que se ha abordado es el basado en listas de términos (ver sección 2.3.4). El añadido de mecanismos de reconfiguración conduce a un enfoque análogo al de las *folksonomías* introducido en el capítulo anterior. La principal diferencia del enfoque propuesto con el clásico de las *folksonomias* es la disponibilidad, en el enfoque propuesto, de un vocabulario explícito que debe ser mantenido colaborativamente. De hecho, es posible estructurar la comunidad encargada de crear colecciones en dos grupos diferenciados:

- Los expertos en el dominio que se encargan de producir y mantener colaborativamente los vocabularios de listas de términos. Dichos vocabularios pueden evolucionar conforme se descubren nuevas necesidades de etiquetado.
- Los catalogadores que utilizan los vocabularios para describir los objetos digitales.

Con el fin de facilitar la organización y mantenimiento de estos vocabularios, es posible organizarlos en categorías jerárquicas (dichas categorías pueden introducir, por ejemplo, temas o propósitos de etiquetados) susceptibles de ser reconfiguradas, mediante la reordenación de sus nodos.

Este enfoque se ha implementado en la herramienta *Clavy* ya aludida anteriormente. El trabajo se describe en (Joaquín Gayoso-Cabada et al., 2016c).

4.1.2.2 Taxonomías

La necesidad de utilizar *taxonomías* (ver sección 2.3.4.2) surgió en el marco del proyecto financiado por Google (*Collaborative Annotation of Digitalized Literary Texts*). El objetivo de este proyecto era el desarrollo de una herramienta educativa para el anotado colaborativo de textos literarios digitalizados. La herramienta desarrollada fue la herramienta @note ya descrita en el capítulo anterior. Como se ha indicado en el capítulo anterior, dicha herramienta permite definir actividades de anotado que constan de un anotado colaborativo del texto por parte de los alumnos, y en la clasificación de las anotaciones utilizando una taxonomía de términos (i.e., palabras seleccionadas para denotar de forma única un concepto) desarrollada colaborativamente por los diseñadores de las actividades (los docentes de Filología, en este caso).

Así mismo, @note permite la reestructuración de las taxonomías creadas colaborativamente por los expertos. Las operaciones de reestructuración contemplan:

- La reorganización de las relaciones *generalización-particularización* de los términos en la taxonomía (es decir, la reorganización de los nodos en la jerarquía de conceptos).
- La fusión de términos para resolver problemas de sinonimia. Esto permite identificar como iguales conceptos que inicialmente se consideraban como diferentes durante la creación colaborativa de las taxonomías.

El enfoque resultante se describe en (Joaquin Gayoso-Cabada et al., 2012) y (Joaquín Gayoso-Cabada et al., 2013). @note ha sido y está siendo utilizada de manera extensiva por las investigadoras del Grupo de Investigación *Literaturas Españolas y Europeas: del Texto al*

Hipertexto (LEETHI) de la Universidad Complutense Madrid en diversos escenarios educativos.

4.1.2.3 Tesauros facetados

Siguiendo un enfoque similar al adoptado en la definición de taxonomías, también se ha experimentado con el uso del modelo para definir tesauros facetados (ver 2.3.4.3). Este trabajo se ha desarrollado en el contexto del proyecto HUM14_251 de la Fundación BBVA.

Para este fin, se ha comenzado definiendo un modelo simple de tesauro facetado, formado por *facetas* que agrupan *términos*, y que, a su vez, pueden refinarse mediante sub-facetas. Los recursos en sí se etiquetan mediante términos elegidos de las facetas apropiadas. De esta forma, la reconfiguración implica reorganizar la jerarquía de facetas mediante el cambio de filiación de las mismas (es decir, modificando las facetas padre).

Este enfoque se ha implementado también en *Clavy*, y ha sido aplicado parcialmente por las investigadoras del grupo LEETHI en el desarrollo de distintas bibliotecas digitales especializadas: *Mnemosine*[77], una biblioteca digital sobre textos raros y olvidados de la Edad de Plata en España, *Ciberia*[78], una biblioteca sobre literatura digital en español, y *Tropos*[79] una biblioteca digital sobre escritura creativa. El trabajo centrado en los tesauros se describe en (Joaquín Gayoso-Cabada et al., 2016a).

4.1.3 *Conclusiones*

En relación con la reconfiguración de esquemas de metadatos, el trabajo realizado en torno al objetivo 1 de esta tesis se ha enfocado definiendo, primeramente, un modelo agnóstico que concibe dichos esquemas como estructuras jerárquicas. Dichas estructuras pueden integrar tanto elementos estructurales, como elementos descriptivos. Los elementos estructurales sirven como componentes organizativos en los esquemas. Por su parte, los elementos descriptivos son esenciales para configurar los esquemas de metadatos, ya que la catalogación con dichos esquemas supone proporcionar valores apropiados para los distintos elementos a nivel de cada objeto digital catalogado. Las estructuras jerárquicas resultantes pueden reconfigurarse mediante la reconfiguración de la filiación de los elementos en los esquemas. Además, la

[77] http://repositorios.fdi.ucm.es/mnemosine/

[78] http://repositorios.fdi.ucm.es/ciberia/

[79] http://repositorios.fdi.ucm.es/Tropos/

representación tabular de los documentos de metadatos permanece invariante a dichas reconfiguraciones.

En relación con la reconfiguración de vocabularios controlados, se ha abordado la reconfiguración en vocabularios de etiquetas, taxonomías, y tesauros facetados. De nuevo, en dichos escenarios es posible aplicar operaciones de reconfiguración estructural análogas a las anteriores, basadas en el cambio de filiación de los nodos en las jerarquías. Así mismo, en el caso de las taxonomías se han aplicado reconfiguraciones basadas en la identificación de conceptos y orientadas a la resolución de problemas de sinonimia.

4.2 Soporte a los expertos para la reconfiguración

El segundo objetivo planteado en esta tesis aborda el aspecto *humano* del proceso de reconfiguración de esquemas de catalogación. Su propósito es encontrar mecanismos que puedan servir de apoyo a los expertos para refinar iterativamente sus esquemas, aplicando aquellas reconfiguraciones que estimen más relevantes. Este apartado se centra en los trabajos llevados a cabo en relación con este segundo objetivo.

Estos trabajos se han desarrollado en dos actividades diferentes. La primera de estas actividades estaba dirigida a la concepción de un modelo de inferencia de las estructuras de catalogación a partir de las colecciones anotadas, que se describe en la sección 4.2.1 . La segunda actividad, se describe en la sección 4.2.2 , y se centra en los trabajos relativos a la validación inicial del modelo.

4.2.1 *Desarrollo del modelo de inferencia de estructuras de catalogación*

Para llevar a cabo la inferencia de las estructuras de catalogación a partir de las colecciones se ha aplicado la técnica de *análisis de conceptos formales* (Sarmah, Hazarika, & Sinha, 2015). El punto de partida de dicha técnica es un *contexto formal*: un conjunto de *objetos*, donde cada uno de ellos viene descrito por un conjunto de *atributos*. La técnica, entonces, determina los posibles *conceptos formales* que se derivan de dicho contexto, así como su estructura. Formalmente, cada concepto formal se puede representar como un par (O,A), donde O es un conjunto de objetos y A es un conjunto de atributos tal que:

- A contiene exactamente todos aquellos atributos que son comunes a todos los objetos en O.
- O contiene exactamente todos aquellos objetos que contienen, en sus descripciones, todos los atributos en A.

Los conceptos formales pueden ordenarse de acuerdo a una relación de orden $\sqsubseteq$ definida como $(O,A) \sqsubseteq (O',A) \Leftrightarrow A \subseteq A'$ (de manera equivalente, $O' \subseteq O$). De hecho, dicha relación de orden define una estructura de *retículo* sobre el conjunto posible de conceptos formales asociado a un contexto formal (que es la estructura que, junto al conjunto de conceptos formales, suelen producir los algoritmos asociados con la técnica de análisis de conceptos formales) (Davey & Priestley, 2002).

De esta forma, el método básico propuesto consiste en:

- Abstraer de manera adecuada la colección catalogada mediante un contexto formal.
- Aplicar el análisis de conceptos formales para construir un retículo que pueda compararse, de manera significativa, con el esquema de catalogación propuesto.

Tal y como se ha anticipado en el planteamiento del trabajo, este método se ha aplicado únicamente al caso particular de las taxonomías. No obstante, se piensa que probablemente dicho método podría generalizarse a otro tipo de esquemas, sin más que idear mecanismos adecuados para llevar a cabo la abstracción de las colecciones como contextos formales, de forma que los retículos resultantes fueran comparables con los esquemas utilizados.

En el de caso particular de las taxonomías, en el proceso de obtención del contexto formal asociado a la colección, se consideran como *objetos* los propios objetos digitales, y como *atributos* el cierre respecto a la relación "*es un*" (generalización-especialización) en la taxonomía de los términos *conceptos taxonómicos* que etiquetan a dichos objetos (es decir, se incluyen como atributos tanto los conceptos taxonómicos explícitamente indicados en el etiquetado, como todos los superconceptos de dichos conceptos). Es por esta razón, que el retículo resultante no será otra cosa que una taxonomía alternativa, inferida del uso real de la colección, y comparable con la taxonomía original propuesta. Los expertos podrán inspeccionar dicho retículo, detectando posibles modificaciones y mejoras en la taxonomía original propuesta.

Este enfoque, que se ha implementado a nivel de prototipo en la herramienta @note, se detalla en (Cigarrán-Recuero et al., 2014).

4.2.2 *Validación inicial de la propuesta de soporte a la reconfiguración*

La validación del modelo de inferencia de estructuras de catalogación se ha llevado a cabo en el contexto de @note mediante una actividad de anotación centrada en el relato "La Biblioteca de Babel" (Borges, 1944). Para ello, los expertos en el dominio que diseñaron la

actividad (profesores de la Facultad de Filología de la UCM) propusieron una taxonomía basada en la metodología *close reading* (DuBois, 2003; Fisher & Frey, 2012; Van Looy & Baetens, 2003) para la realización de análisis crítico de textos. La actividad fue llevada a cabo por alumnos (que actuaron como catalogadores), y, seguidamente analizada utilizando el modelo de inferencia. Como resultado:

- Los expertos en el dominio detectaron interpretaciones incorrectas en el uso de la taxonomía por parte de los catalogadores, y pudieron tomar acciones correctivas al respecto (instruir a los alumnos acerca de dichas interpretaciones erróneas).
- Así mismo, los expertos también detectaron oportunidades de mejora de la taxonomía, lo que condujo al refinamiento de la misma.

  La experiencia se describe con detalle en (Cigarrán-Recuero et al., 2014).

4.2.3 *Conclusiones*

Los trabajos llevados a cabo en relación con el objetivo 2 ponen de manifiesto la factibilidad de utilizar modelos de inferencia de estructuras de catalogación a partir de las colecciones catalogadas, así como de usar dichas estructuras para mejorar tanto los esquemas como su uso. En concreto, se ha propuesto un método basado en análisis de conceptos formales para llevar a cabo la inferencia de las estructuras, y se ha validado dicho método con catalogaciones taxonómicas en la herramienta de anotación de textos literarios digitalizados @note. El aspecto más crítico de dicho método es encontrar una abstracción adecuada de las colecciones mediante contextos formales, de tal forma que los retículos inferidos aporten información útil sobre la estructura y el uso de los esquemas (idealmente, dichos retículos deben representar, al menos estructuralmente, esquemas de catalogación alternativos, que puedan ser comparados con los originales).

El método se ha aplicado a una experiencia real de anotación con @note, obteniéndose resultados positivos, tanto en lo referente a la obtención de información útil para la reconfiguración, como también información útil para guiar a los catalogadores en un mejor uso de los esquemas.

4.3 Adaptación de las plataformas a las reconfiguraciones

El tercer objetivo de la tesis se centra en la búsqueda de mecanismos que permitan que el sistema de gestión de colecciones se adapte dinámicamente a las reconfiguraciones de los esquemas, sin que ello suponga una interrupción de la operación normal del sistema, ni

tampoco una degradación significativa de su rendimiento. En este apartado se describen, por tanto, los trabajos realizados en relación con dicho objetivo.

Siguiendo el planteamiento de trabajo propuesto en el capítulo anterior, el trabajo se ha restringido a una de las funcionalidades más básicas y, a la vez, más fundamentales en el sistema de gestión de colecciones: la navegación guiada por los esquemas de catalogación. El trabajo en sí se ha abordado a través de dos actividades diferentes, la formulación de un modelo de navegación invariante a las reconfiguraciones (sección 4.3.1) y la formulación de métodos de indexación que permitan implementar en la práctica dicho modelo (sección 4.3.2).

4.3.1 *Formulación del modelo de navegación*

Tal y como se ha anticipado en el capítulo anterior, para formular el modelo de navegación se ha adoptado un enfoque *dirigido por lenguajes*. Más concretamente, el modelo de procesamiento de lenguaje considerado es un autómata finito determinista, el *autómata de navegación*, que reconoce todas las posibles secuencias de *interacciones* del usuario durante la navegación, y cuyos estados son los conjuntos de objetos digitales acotados por dichas secuencias de interacciones. Las interacciones en sí vienen dadas por las instancias de los elementos de información en el esquema que pueden ser utilizados para restringir el conjunto de objetos seleccionado. De esta forma:

- El estado inicial del autómata está formado por el conjunto de todos los objetos de la colección.
- Dados dos estados Q y Q', y una interacción *i*, se define la transición $Q \xrightarrow{i} Q'$ si Q' contiene aquellos objetos de Q filtrados por *i*.

Es importante notar que, en la determinación de las interacciones, únicamente se tienen en cuenta las posibles instancias de elementos de información, pero no la estructura particular del esquema. Por tanto, cuando dicha estructura cambia, el autómata permanece invariante, ya que dicho autómata es, en realidad, una caracterización de la navegación *para todas las posibles reconfiguraciones del esquema*. Como resultado, en el peor de los casos, el autómata puede crecer exponencialmente con el tamaño de la colección.

En el contexto de *Clavy*, el modelo se ha concretado para distintos tipos de esquemas de catalogación:

- *Listas de términos*. En este caso, las interacciones vienen dadas por los términos del vocabulario: el usuario puede restringir el conjunto de objetos seleccionado en el estado

actual hasta el momento mediante la selección de uno de los términos asociadas con los objetos de dicho estado. Como resultado, el modelo proporciona una solución muy elegante y eficiente para la navegación multi-nivel en escenarios tales como los sistemas basados en *folksonomías*. Efectivamente, tal y como resulta evidente, el autómata no cambia si se cambia la organización en categorías de los términos. Esta concreción del modelo se describe en (Joaquín Gayoso-Cabada et al., 2016c).

- *Tesauros facetados*. Las interacciones vienen dadas por los términos del tesauro. Dado que en dicha elección de las interacciones no se tiene en cuenta la estructura jerárquica de facetas, el autómata es invariante a las posibles reorganizaciones de dicha jerarquía. Por tanto, el autómata sirve como modelo de navegación para todas las posibles reorganizaciones del tesauro. La aplicación del modelo a tesauros facetados reconfigurables se describe en (Joaquín Gayoso-Cabada et al., 2016a).
- *Esquemas de metadatos*. El modelo se ha aplicado también a los esquemas de metadatos que se ajustan al modelo desarrollado durante la consecución del objetivo 1 (véase la 2.3.4). Como en el caso de los tesauros facetados, el autómata es invariante a los posibles cambios de filiación de los elementos en el esquema, lo que permite utilizarlo para guiar la navegación independientemente de la organización jerárquica de los elementos de metadatos. La aplicación del modelo a los esquemas de metadatos se describe en (Joaquín Gayoso-Cabada et al., 2016b).

Por último, es interesante reseñar que, en todos los casos indicados, los autómatas de navegación que resultan están estrechamente relacionados con los retículos de conceptos que se infieren directamente a partir de las colecciones catalogadas (cuando se ignoran las estructuras de los esquemas). En particular, los estados del autómata se corresponden con conceptos formales, y las transiciones refinan la relación de orden con información explícita de las interacciones lícitas en cada estado.

4.3.2 *Desarrollo de métodos prácticos de indexación*

El modelo basado en autómatas de navegación proporciona una solución elegante y eficiente al problema de la reconfiguración de esquemas de catalogación. No obstante, tal y como se ha comentado anteriormente, en el peor de los casos la representación puede requerir un tamaño exponencial. Aunque tales casos no suelen aparecer en la práctica, el factor exponencial no puede ignorarse, máxime si se desea proporcionar sistemas de gestión de

colecciones generales, que permitan tratar, no colecciones particulares, sino clases completas de colecciones. Así mismo, y aunque el autómata no crezca exponencialmente para una determinada colección, sí puede demandar una cantidad de espacio desmesuradamente grande. Lo ideal sería, por tanto, contar con representaciones más compactas, idealmente del orden del tamaño de la colección. Es por ello que se ha trabajado en encontrar dichas representaciones prácticas para soportar el modelo de navegación. A continuación, se detallan las representaciones exploradas.

4.3.2.1 Índices invertidos

Los índices invertidos se utilizan extensivamente en contextos de recuperación de la información (Chowdhury, 2010; Kriegel, 1984). Básicamente, un índice invertido indica, para cada elemento de información, los objetos seleccionados por dicho elemento de información. De esta forma, el índice permite recrear dinámicamente los estados del autómata de navegación sin más que intersecar, para cada posible secuencia de interacciones, las entradas en el índice para cada interacción en la secuencia.

La solución basada en índices invertidos es una solución simple y habitual para abordar el problema de la reconfiguración en gestores con esquemas dinámicamente reconfigurables (de hecho, tal y como se ha indicado en los capítulos anteriores, es la solución adoptada en ODA). No obstante, su principal inconveniente es el coste de computar las intersecciones de conjuntos de objetos requeridas por el cómputo dinámico de los estados. Conforme aumenta el tamaño de las colecciones, este coste puede impactar negativamente en la experiencia de usuario.

Tal y como se describe en (Joaquín Gayoso-Cabada et al., 2016b, 2016a, 2016c), esta técnica se ha implementado en *Clavy*, tanto en el contexto de las *folksonomías*, como en el de los tesauros facetados y el de los esquemas de metadatos. La implementación se basa en el marco de recuperación de información *Lucene* (McCandless, Hatcher, & Gospodnetić, 2010), que soporta de manera muy eficiente la indexación de contenidos digitales mediante índices invertidos.

4.3.2.2 Dendrogramas de navegación

Con el fin de mejorar el rendimiento obtenido mediante índices invertidos, se ha diseñado una estrategia alternativa inspirada en el uso de *dendrogramas* en escenarios de agrupamiento jerárquico (Perugini, 2010). El resultado se denomina *dendrograma de*

navegación, y es una estructura en árbol en la que los nodos representan conjuntos de estados, y los arcos están etiquetados con interacciones. Así mismo:

- El nodo raíz está asociado con la totalidad de la colección.
- De cada nodo surge un arco para cada posible interacción asociada con los objetos de dicho nodo capaz de restringir dicho conjunto de objetos (es decir, se excluyen las interacciones compartidas por todos los objetos).
- Cada objeto en un nodo, bien se filtra únicamente por una interacción (en caso de que haya varias interacciones posibles, se elige una aleatoriamente), o bien se aloja en dicho nodo (en caso de que no existan interacciones capaces de filtrarlo).

Como resultado, los objetos representados por un nodo vienen dados por los alojados en dicho nodo, y los alojados en todos sus descendientes. Así mismo, el número de nodos en la estructura crece linealmente con respecto al número de objetos en la colección. El enfoque se describe en (Joaquín Gayoso-Cabada et al., 2016a) y en (Joaquín Gayoso-Cabada et al., 2016b). El aspecto negativo es que, durante la navegación, será necesario mantener, no un único nodo en el dendrograma, sino todo un conjunto posible de nodos en el mismo. Este hecho se explicita en (Joaquín Gayoso-Cabada et al., 2016c), donde el dendrograma se identifica con una representación de una versión no determinista del autómata de navegación.

4.3.2.3 Comparativa

En el contexto de *Clavy* se ha comparado la implementación basada en índices invertidos (utilizando Lucene) con la implementación basada en dendrogramas. Para ello, se ha utilizado como caso de estudio el sistema *Chasqui* ya referido en el 2.6.2 , planteándose un experimento en el que se intercala el añadido de objetos provenientes de *Chasqui* con la simulación de distintas sesiones de navegación. Tal y como se describe en (Joaquín Gayoso-Cabada et al., 2016a, 2016b, 2016c), el experimento se ha replicado tanto en el caso de *folksonomías*, como de *tesauros facetados* y *esquemas de metadatos* (en todos los casos se importó *Chasqui* en *Clavy* utilizando una representación adecuada del esquema de catalogación), induciéndose, en todos los casos, el mismo autómata de navegación. Los resultados obtenidos evidencian la superioridad, en el caso de estudio elegido, del enfoque basado en dendrogramas frente al basado en índices invertidos.

4.3.3 *Conclusiones*

El objetivo 3 se ha abordado formulando, primeramente, un modelo de navegación basado en autómatas finitos. Se ha comprobado cómo es posible obtener, para distintos tipos de esquema, autómatas que permanecen invariantes frente a la reconfiguración de dichos esquemas. Así mismo, dichos autómatas soportan la navegación de manera muy eficiente, ya que codifican explícitamente los posibles estados de navegación, así como las posibles interacciones. La principal desventaja, sin embargo, radica en la complejidad del método, que, en el peor de los casos, puede ser exponencial con respecto al tamaño de la colección.

Para abordar el problema de complejidad subyacente al método se han analizado alternativas más prácticas de indexado, tanto índices invertidos, como dendrogramas de navegación. Se ha comprobado también empíricamente la superioridad de los dendrogramas de navegación con respecto a los índices invertidos que habían sido utilizados en investigaciones anteriores en el Grupo de Investigación, como las relativas al sistema ODA.

4.4 A modo de conclusión

En este capítulo se ha descrito cómo se han abordado los distintos objetivos planteados en esta tesis en relación con la gestión de colecciones digitales con esquemas de catalogación dinámicamente reconfigurables. Más concretamente:

1. El primero de estos objetivos, la reconfiguración de esquemas de catalogación, se ha abordado mediante la realización de distintas experiencias relativas a la formulación de modelos reconfigurables, tanto de esquemas de metadatos, como de vocabularios controlados. En este sentido, en el contexto de la herramienta @note, se han definido mecanismos para la reconfiguración de taxonomías, y en el contexto de la herramienta *Clavy* se ha abordado la reconfiguración de organizaciones jerárquicas de vocabularios de etiquetas, de tesauros facetados, y de un modelo de esquema de metadatos especialmente diseñado para soportar la reconfiguración eficiente de las estructuras internas de un sistema de gestión de colecciones.
2. El segundo de los objetivos, la provisión de soporte a los expertos en el dominio para la reconfiguración, se ha abordado mediante la aplicación de la técnica de análisis de conceptos formales. Aunque el trabajo se ha enfocado al caso particular de las taxonomías para la organización de anotaciones de textos literarios digitalizados en el

contexto de @note, se piensa que el método propuesto podría ser extrapolable en a otro tipo de esquemas de catalogación y a otro tipo de colecciones.

3. Finalmente, el último objetivo, la provisión de mecanismos de indexado que permanezcan invariantes ante las reconfiguraciones de los esquemas, se ha abordado para el caso de la navegación, modelando del proceso de navegación guiada por colecciones mediante autómatas finitos deterministas. Así mismo, y con el fin de evitar el potencial crecimiento exponencial de los autómatas, se han propuesto enfoques prácticos de indexado, basados en índices invertidos y en dendrogramas, así como en la reconstrucción dinámica de las partes relevantes del autómata de navegación durante la navegación por las colecciones.

En el próximo capítulo se discutirán con más detalle el alcance y las limitaciones de los resultados obtenidos, así como las posibles vías de investigación que surgen de los mismos.

Capítulo 5 - Conclusiones y Trabajo Futuro

En los capítulos anteriores se ha discutido la problemática relativa a la gestión de colecciones de objetos digitales, así como la asociada a la reconfiguración de los esquemas de catalogación de dichas colecciones. El problema de la reconfiguración de esquemas de catalogación se ha abordado, así mismo, desde el punto de vista lógico (es decir, desde la perspectiva relativa a la manipulación y edición de los esquemas), desde el punto de vista humano (es decir, desde la perspectiva relativa al guiado de los expertos para la reconfiguración de los esquemas), y desde el punto de vista tecnológico (es decir, mediante la propuesta de mecanismos que permitan paliar el impacto en el rendimiento causado por la reconfiguración de los esquemas). Tal y como se ha detallado en el capítulo anterior, los resultados obtenidos a la hora de cubrir los objetivos propuestos en este trabajo de tesis se detallan en diferentes publicaciones que integran esta memoria. De este modo, este último capítulo concluye esta memoria de tesis resumiendo las aportaciones principales y presentando algunas líneas de trabajo futuro.

5.1 Principales aportaciones

En esta sección se presentan las principales aportaciones realizadas en este trabajo de tesis. Más concretamente, y en base a la discusión mantenida en los capítulos precedentes, se destacan las siguientes aportaciones:

- Estrategias para la reconfiguración de esquemas de catalogación, tanto esquemas de metadatos, como vocabularios controlados, orientadas a la edición de las estructuras de dichos esquemas de catalogación.
- Una propuesta para guiar a los expertos en el dominio durante la reconfiguración de esquemas de catalogación basados en taxonomías.
- Un modelo genérico para la navegación en las colecciones de objetos digitales guiada por los esquemas de catalogación, y distintas propuestas prácticas de indexado de colecciones fundamentadas en dicho modelo.

A continuación, se describen en detalle cada una de estas aportaciones.

5.1.1 *Reconfiguración de esquemas de catalogación*

En este trabajo de tesis se ha partido del enfoque a la reconfiguración de esquemas de metadatos en el sistema ODA, sistema para la gestión de repositorios educativos especializados desarrollado en trabajos previos, y se ha extrapolado este enfoque a la reconfiguración de

distintos tipos de esquemas de catalogación. Más concretamente, dicho enfoque se centra en la representación *estructural* o *sintáctica* de los esquemas, y propone abordar la reconfiguración como un proceso de edición de dicha representación. Las representaciones en sí tienen naturaleza arborescente, y las operaciones propuestas permiten, típicamente:

- Crear y renombrar nodos en las jerarquías.
- Cambiar las filiaciones de los nodos en dichas jerarquías, moviendo nodos de una posición a otra en los árboles que representan estructuralmente a los esquemas.
- Fusionar nodos en dichas jerarquías, lo que permite, entre otros aspectos, resolver los problemas de sinonimia que surgen frecuentemente en escenarios en los que los esquemas se crean de forma colaborativa, por grupos de expertos con distintas opiniones y distintas percepciones de los dominios.

De esta forma, en esta tesis se ha aplicado este modelo básico de reconfiguración, abstraído a partir del trabajo previo realizado en ODA, así como a partir de propuestas como las descritas en (J.-L. Sierra & Fernández-Valmayor, 2008), a distintos tipos de esquemas:

- Esquemas de metadatos que siguen un modelo basado en los lenguajes de marcado generalizado (tipo XML), y que se estructuran como jerarquías de *elementos*, tanto *estructurales*, como *descriptivos*. Los elementos estructurales tienen un propósito exclusivamente organizativo, mientras que los elementos descriptivos permiten introducir valores (los *metadatos* propiamente dichos) en la catalogación descriptiva de los objetos. Dicha catalogación se lleva a cabo asociando *documentos* a los objetos, documentos que se representan como tablas que asignan valores a los elementos descriptivos. Dicha representación tabular permanece, por tanto, invariante a reconfiguraciones tales como el renombrado de elementos, o, más relevantemente, el cambio de filiación de dichos elementos en los esquemas.
- Vocabularios basados en listas de términos que, al estilo de las *folksonomías*, se crean colaborativamente por una comunidad de expertos, quienes, además, pueden organizarlos arborescentemente, utilizando categorías. Dichos expertos pueden, así mismo, reconfigurar dichas organizaciones reordenando adecuadamente las categorías, así como reubicando los términos dentro de las distintas categorías.
- Vocabularios basados en taxonomías. Sobre estos vocabularios se ha soportado el rango completo de reconfiguraciones citado anteriormente, incluyendo la fusión de términos en la taxonomía para permitir resolver problemas de sinonimia.

- Vocabularios basados en tesauros facetados. Estos vocabularios se estructuran en facetas que agrupan términos y que a su vez pueden refinarse mediante subfacetas con sus propios conjuntos de términos, y así sucesivamente… La reconfiguración, en este contexto, se lleva a cabo editando la jerarquía de facetas.

Estas experiencias han permitido contrastar cómo es posible abordar el proceso de reconfiguración de esquemas a un nivel sintáctico, que enfatiza la estructura de dichos esquemas frente a otros aspectos, como el significado o el uso. Dichos aspectos semánticos y pragmáticos serán tenidos en cuenta por el experto que realiza la reconfiguración, pero no necesariamente reflejados en el sistema de soporte a la reconfiguración. Efectivamente, dicho sistema se concibe como un mero editor de jerarquías (árboles) que describen la estructura de los esquemas. El enfoque resultante es efectivo desde un punto de vista práctico, tal y como permite constatar el hecho de que se haya podido aplicar satisfactoriamente tanto a esquemas de metadatos (en repositorios tipo ODA), como a esquemas basados en vocabularios controlados (taxonomías en el contexto de @note, listas de términos y tesauros facetados en el contexto de *Clavy*).

5.1.2 *Soporte a los expertos durante el proceso de reconfiguración*

En esta tesis se ha mostrado la factibilidad de ofrecer ayuda automatizada para la reconfiguración de los esquemas de catalogación a los expertos involucrados en la creación y mantenimiento de una colección de objetos digitales en un determinado dominio. El enfoque propuesto consiste en utilizar técnicas de análisis de datos para, a través del análisis de las propias catalogaciones de los objetos, inducir el esquema que *realmente* se está utilizando en la catalogación. Esto permite a los expertos conocer cómo se está usando realmente el esquema inicialmente propuesto, y, por tanto, cómo debe reconfigurarse dicho esquema para adaptarlo a las necesidades específicas del proceso de catalogación. El enfoque cobra, así mismo, especial sentido cuando el proceso de catalogación en sí se lleva a cabo por una comunidad de *catalogadores* segregada de la comunidad de expertos en el dominio. En este escenario:

- Los expertos proponen los esquemas de catalogación a utilizar.
- Los catalogadores utilizan dichos esquemas para catalogar objetos en la colección.
- Los mecanismos de análisis de datos permiten inducir estructuras conceptuales comparables a los esquemas de partida directamente de las colecciones.

- Los expertos pueden explotar estas estructuras para: (i) orientar a los catalogadores en el proceso de catalogación, en caso de que detecten algún uso incorrecto o poco ortodoxo de los esquemas propuestos, o (ii), reconfigurar estos esquemas para ajustarlos mejor a la realidad del proceso de catalogación.

En particular, en esta tesis se ha aplicado este enfoque al dominio de la catalogación temática de anotaciones sobre textos literarios digitalizados utilizando taxonomías. El uso de taxonomías ha permitido utilizar la técnica del análisis de conceptos formales como técnica de análisis de datos. Efectivamente, dicha técnica ha permitido encontrar, dada una colección de anotaciones debidamente clasificada, un retículo de conceptos que refleja la estructura conceptual de dicha colección. Dicho retículo es, a su vez, comparable con la taxonomía original, lo que muestra la factibilidad práctica del método propuesto.

Es interesante notar, así mismo, que el enfoque basado en análisis de conceptos formales empleado en esta tesis podría, en principio, utilizarse en otros escenarios y con otro tipo de esquemas de catalogación. El punto crítico para llevar a cabo esta aplicación es, tal y como ya se ha comentado en el capítulo anterior, encontrar un mecanismo de abstracción apropiado de las colecciones como contextos formales a partir de los cuales puedan inducirse retículos comparables con los esquemas.

5.1.3 *Navegación eficiente en colecciones reconfigurables*

La última aportación de esta tesis tiene relación con la forma de acomodar eficientemente la reconfiguración de los esquemas de catalogación en el funcionamiento general de los sistemas de gestión de colecciones. Efectivamente, tal y como se ha argumentado ya repetidamente a lo largo de esta memoria, la reconfiguración del esquema implica, en muchos casos, la reconstrucción parcial o total de los índices internos que organizan los objetos de la colección, y que son utilizados de forma intensiva para llevar a cabo funcionalidades críticas, tales como la recuperación de objetos, o la navegación guiada por las colecciones. En particular, en esta tesis se ha abordado el problema de la degradación del rendimiento en el sistema de navegación guiada provocado por la reconfiguración de los esquemas. Para ello:

- Se ha comenzado proponiendo un modelo de navegación, basado en autómatas finitos deterministas, para colecciones con esquemas de catalogación reconfigurables. La caracterización de la navegación mediante autómatas concibe las posibles interacciones que hacen progresar la navegación como un lenguaje formal reconocible mediante tales

autómatas. Dichas interacciones dependen de las instancias de los elementos de información en los esquemas, pero no de la forma en la que dichos elementos se organizan en dichos esquemas. Como resultado, los autómatas resultantes permanecen invariantes ante las posibles reconfiguraciones.

- Se ha visto, así mismo, que los autómatas de navegación resultantes de aplicar el modelo propuesto pueden, en ciertos casos, tener tamaños inaceptables (en el peor de los casos el número de estados de estos autómatas puede crecer exponencialmente con el tamaño de las colecciones). Por tanto, se han propuesto alternativas más prácticas al indexado de las colecciones, que eviten la potencial explosión exponencial de los autómatas de navegación y que, por otra parte, posibiliten recrear dinámicamente las partes relevantes de dichos autómatas de forma razonablemente eficiente.

En lo que se refiere a las propuestas de indexado, se ha discutido cómo la representación clásica basada en índices invertidos es, en realidad, una alternativa a la representación del autómata de navegación. No obstante, y aunque el mantenimiento de estos índices es relativamente sencillo conforme la colección evoluciona (es decir, conforme se añaden, eliminan o modifican objetos), la principal desventaja que presentan es la necesidad de computar explícitamente costosas intersecciones de conjuntos de objetos cada vez que se transita a un nuevo estado de navegación. Para paliar esta desventaja se ha propuesto una representación jerárquica alternativa, inspirada por los dendrogramas utilizados en agrupamiento jerárquico. La representación resultante puede concebirse, de hecho, como una representación explícita, y razonablemente elaborada, de un autómata no determinista equivalente al autómata de navegación. Los experimentos llevados a cabo con la colección *Chasqui* importada a *Clavy* sugieren la superioridad de esta representación basada en dendrogramas frente a la basada en índices invertidos.

5.2 Trabajo futuro

Esta sección finaliza este trabajo de tesis exponiendo las líneas de trabajo futuro más prometedoras, derivadas de las propuestas presentadas en dicha tesis. Dichas líneas de trabajo futuro se enumeran a continuación:

- Formulación de un meta-modelo genérico para esquemas de catalogación reconfigurables

- Formulación de un enfoque ETL (*extract – transform – load*) para colecciones digitales basado en especificaciones declarativas.
- Búsqueda de mecanismos genéricos de guiado a los expertos en el dominio durante el proceso de reconfiguración, que sean independientes del tipo de esquema de catalogación.
- Búsqueda de mecanismos de indexado de colecciones alternativos a los propuestos en esta tesis.
- Estudio del impacto de las reconfiguraciones en otras funcionalidades de los sistemas de gestión de colecciones, así como búsqueda de mecanismos que permitan paliar dicho impacto.

  Los siguientes puntos motivan cada una de estas líneas de trabajo futuro e investigación.

5.2.1 *Meta-modelo para esquemas de catalogación dinámicamente reconfigurables*

En esta tesis se ha analizado cómo es posible abordar la reconfiguración de distintos tipos de esquemas de catalogación mediante: (i) la representación arborescente de dichos esquemas, y (ii), la reconfiguración de dichas estructuras jerárquicas, entendida como un proceso de edición de las mismas. Este hecho parece sugerir que es posible formular un *meta-modelo* genérico para la caracterización estructural de esquemas de catalogación, sobre el que poder, así mismo, formular el proceso de reconfiguración de forma independiente a cada tipo de esquema particular. A este respecto, se piensa que un buen punto de partida para formular dicho meta-modelo es basando el mismo en *gramáticas incontextuales*, tales como estas se entienden en el diseño de lenguajes informáticos (Aho, Lam, Sethi, & Ullman, 2006). Efectivamente, una gramática incontextual no es más que una descripción finita de un conjunto infinito de árboles. De esta forma, no resulta descabellado caracterizar, al menos estructuralmente, un esquema de catalogación en términos de una notación basada en gramáticas incontextuales. El enfoque resultante será similar a los ya utilizados anteriormente en el Grupo de Investigación ILSA para modelar, mediante gramáticas, la estructura de documentos XML (Sarasa-Cabezuelo & Sierra, 2013b; Sarasa-Cabezuelo, Temprado-Battad, Rodríguez-Cerezo, & Sierra, 2012), documentos JSON (Sarasa-Cabezuelo & Sierra, 2013a), o redes de objetos en el contexto del desarrollo de software dirigido por modelos (Sarasa-Cabezuelo & Sierra, 2015).

5.2.2 *Enfoque* etl *basado en especificaciones declarativas*

Tal y como se ha comentado en esta memoria, el sistema *Clavy* (utilizado como uno de los bancos básicos de experimentación en esta tesis) incluye una arquitectura basada en conectores (*plug-ins*) para la importación – exportación de colecciones de objetos digitales desde / hacia distintas fuentes externas de información. El resultado es, en realidad, un sistema ETL (*extract – transform – load*) (Kimball & Caserta, 2004) para gestionar colecciones de objetos digitales. En dicho sistema es posible, así mismo, disponer de componentes *extract* (los encargados de importar colecciones de fuentes externas) y *load* (los encargados de volcar las colecciones en fuentes externas) genéricos, orientados a distintos formatos de información (p.e., componentes *extract* y *load* para cargar – volcar colecciones desde/a bases de datos relacionales, documentos XML, hojas de cálculo, etc.). De esta forma, el grueso del sistema recaerá en los componentes *transform*, encargados de transformar colecciones en colecciones. Dichos componentes serán conscientes de la *semántica* particular de los esquemas de las colecciones involucradas, lo que les permitirán llevar a cabo transformaciones específicas, dependientes de las colecciones concretas.[80] De esta forma, y conectando con la propuesta de representar esquemas de catalogación como gramáticas sugerida en la sección anterior, parece razonable describir dichos componentes *transform* como extensiones semánticas (con acciones de procesamiento) de las gramáticas incontextuales que, según dicha propuesta, caracterizarán los esquemas. Esto conduce a un enfoque declarativo a la especificación de componentes *transform*, mediante extensiones semánticas de gramáticas incontextuales. Tales componentes podrán generarse, entonces, automáticamente a partir de dichas especificaciones, de forma análoga a como en los trabajos previos del grupo ILSA anteriormente citados (Sarasa-Cabezuelo & Sierra, 2013b, 2013a, 2015; Sarasa-Cabezuelo et al., 2012) se genera automáticamente componentes de procesamiento de documentos XML o JSON, o transformaciones entre modelos, a partir de especificaciones basadas en esquemas de traducción o en gramáticas de atributos (Aho et al., 2006).

[80] En cierta forma, el proceso de reconfiguración explorado en esta tesis puede verse como un caso particular de proceso *transform*, operado por un humano, en lugar de por un programa.

5.2.3 *Enfoque genérico para el guiado durante el proceso de reconfiguración*

Tal y como se ha sugerido ya en esta memoria, parece razonable estudiar formas de generalizar el enfoque basado en análisis de conceptos formales aplicado a las actividades de anotación con @note a otros escenarios de catalogación, apoyados en otro tipo de esquema. A este respecto, se ha sugerido ya abordar el enfoque resultante mediante un proceso de abstracción de cada tipo de colección como un contexto formal apropiado, a partir del cual sea posible inferir retículos significativos. Así mismo, otro posible camino que resulta prometedor es explotar, para tal fin, el propio autómata de navegación, en virtud de la relación existente entre los estados de dicho autómata y los conceptos formales que integran el hipotético retículo inducido por la colección (véase la discusión mantenida al respecto en el capítulo anterior). Por último, un aspecto igualmente relevante que debe explorarse es el hecho de que, conforme el tamaño de la colección crece, la visualización de los retículos o de las estructuras equivalentes a los mismos carece de sentido. Será necesario, entonces, contar con soporte automatizado a la comparación entre tales estructuras y los esquemas de catalogación, así como mecanismos que permitan sugerir automáticamente reconfiguraciones a los expertos a partir de dicha comparación.

5.2.4 *Enfoques de indexado alternativos*

Otra línea de trabajo prometedora es seguir refinando los enfoques de indexado de las colecciones. A este respecto, si bien en esta tesis se ha hecho palpable que el tamaño del autómata de navegación puede crecer, en el peor de los casos, exponencialmente con el tamaño de la colección, dicho resultado es, en realidad, un resultado *en el peor de los casos*. De esta forma, para colecciones concretas puede ser perfectamente posible tener una representación explícita de dicho autómata. Merece la pena, por tanto, evaluar empíricamente el tamaño de los autómatas para distintas colecciones. A este respecto, se pueden utilizar las distintas colecciones que están desarrollando con *Clavy* los grupos de Humanidades con los que el grupo ILSA colabora actualmente, tanto en la UCM (grupos LEETHI y LOEP), como en Panamá (Fundación El Caño para la preservación del patrimonio arqueológico panameño). También merece la pena realizar un estudio empírico de cómo influye la catalogación en el tamaño del autómata (p.e., cómo influye la densidad de elementos de información utilizados para describir cada objeto en el número de estados del autómata resultante). Por último, se estima también interesante utilizar mecanismos que permitan construir el autómata de manera *perezosa*,

durante la navegación, así como mecanismos que permitan organizar una memoria *caché* basándose en dicha idea, que guarde, en cada momento, las partes del autómata más utilizadas.

5.2.5 *Otras funcionalidades en sistemas de gestión de colecciones reconfigurables*

Por último, al igual que en esta tesis se ha estudiado la manera de mejorar la eficiencia de la navegación guiada en colecciones con esquemas reconfigurables, surge de manera natural hacer este estudio extensivo a otro tipo de funcionalidades. En particular, se considera interesante abordar el problema de la recuperación de objetos a partir de consultas. Tal y como ya se ha comentado en esta memoria, dicha funcionalidad puede, en cierto sentido, reducirse a la funcionalidad de navegación. Efectivamente, dada una consulta booleana, es posible transformar la misma a forma normal conjuntiva. Si el resultado no contiene negaciones, cada término podrá resolverse directamente interpretándolo como un camino de navegación, y utilizando el autómata de navegación para proporcionar directamente el resultado (como un conjunto de estados). Las negaciones podrán tratarse, bien de manera no determinista, bien extendiendo el autómata para considerar información negativa, además de la información positiva considerada en el modelo actual.

Capítulo 6 - Artículos Presentados

6.1 A flexible model for the collaborative annotation of digitized literary works

Cita Completa:

Gayoso-Cabada, J., Ruiz, C., Pablo-Nuñez, L., Sarasa-Cabezuelo, A., Goicoechea-de-Jorge, M., Sanz-Cabrerizo, A., & Sierra-Rodriguez, J.-L. (2012). A flexible model for the collaborative annotation of digitized literary works. En Proceedings of the 2012 Digital Humanities Conference, DH 2012 (pp. 190–193).

Resumen original de la contribución:

@Note 1.0 allows us to retrieve digitized works from Google Books collection and add annotations to enrich the texts with research and learning purposes: critical editions, reading activities, e-learning tasks, etc. One of the main features of @Note annotation model, which distinguishes it from similar approaches (Azouaou & Desmoulins 2006; Bechhofer at al. 2002; Koivunen 2005; Rios da Rocha et al. 2009; Schroeter et al. 2006; Tazi et al. 2003), is to promote the collaborative creation of annotation schemas by communities of researchers, teachers and students, and the use of these schemas in the definition of annotation activities on literary works. It results in a very flexible and adaptive model, able to be used by many different communities of experts in literature defending different critical literary theories and for different annotation tasks. In this paper we present this annotation model.

Referencias Bibliográficas:

(Azouaou & Desmoulins, 2006; Bechhofer, Carr, Goble, Kampa, & Miles-Board, 2002; Brachman & Levesque, 2004; Da Rocha, Willrich, Fileto, & Tazi, 2009; Fraternali, Rossi, & Sánchez-Figueroa, 2010; Guermeur & Unruh, 2010; Koivunen, 2005; Polsani, 2006; Richardson & Ruby, 2008; Rumbaugh, Jacobson, & Booch, 2005; Schroeter, Hunter, Guerin, Khan, & Henderson, 2006; Tazi, Al-Tawki, & Drira, 2003)

Figuras originales:

Debido a que las figuras de la versión editada no presentan calidad suficiente, se incluyen a continuación versiones de las figuras más relevantes con mayor calidad, a fin de facilitar el seguimiento del trabajo.

Figure 1

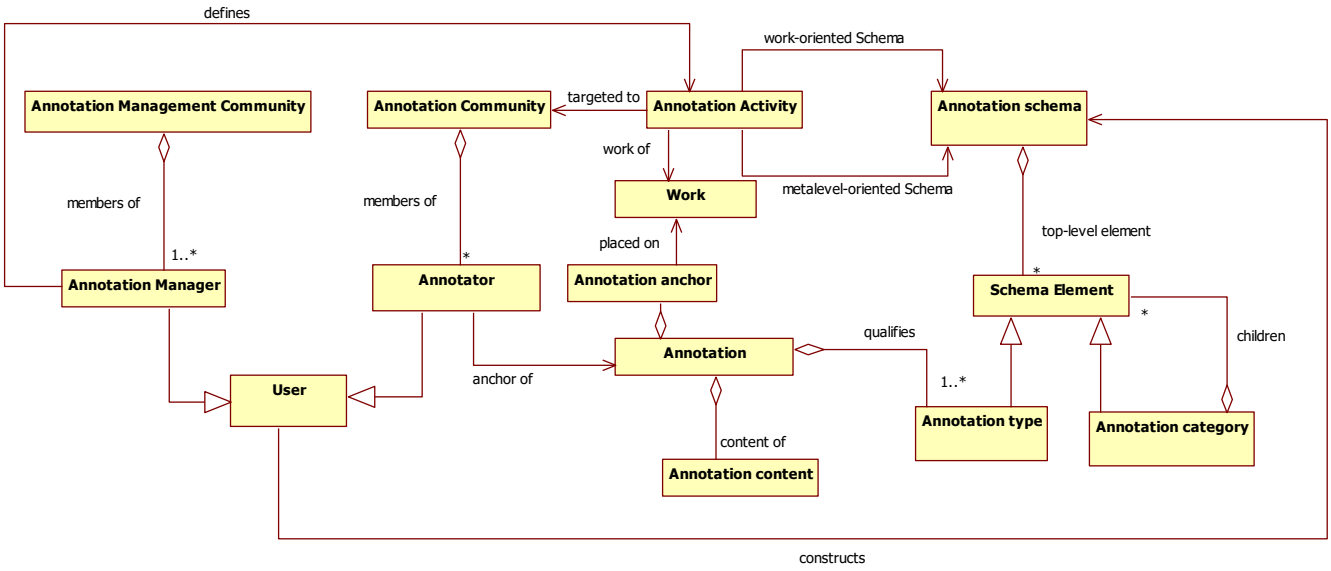
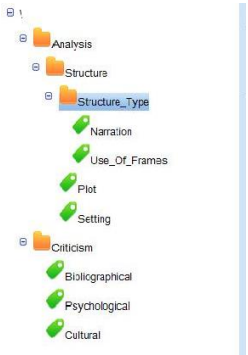


Figure 2

In the context of the annotation management community, an annotation schema can be *public* or *private*. A private schema is only accessible for the annotation manager who created it. On the contrary, a public schema is accessible for all the annotation managers. Annotation managers have unlimited privileges on all the schemas to which they can access (i.e., they can create new annotation types and categories, they can blend two different types/categories in a single one, they have renaming and erasing privileges, etc), with the exception of modifying the *public*/*private* character (it only can be done by the schema's creator). In addition, when annotation managers create annotation activities, they only can choose those schemas to which they have access grants. Concerning annotators, they can add new types and categories to the book-oriented annotation schema, but they can't perform any other modification.

Figure 3

(a)



(b)

```
Structure ⊆ Analysis
Structure_Type  ⊆ Structure
Narration  ⊆ Structure_Type
Use_Of_Frames ⊆ Structure_Type
Plot ⊆ Structure
Setting ⊆ Structure
Bibliographical ⊆ Criticism
Psychological ⊆ Criticism
Cultural ⊆ Criticism
...
```

A flexible model for the collaborative annotation of digitized literary works

Gayoso-Cabada, Joaquin
gayoxo@gmail.com
Universidad Complutense de Madrid, Spain

Ruiz, Cesar
cruiz85@gmail.com
Universidad Complutense de Madrid, Spain

Pablo-Nuñez, Luis
lpnunez@filol.ucm.es
Universidad Complutense de Madrid, Spain

Sarasa-Cabezuelo, Antonio
asarasa@fdi.ucm.es
Universidad Complutense de Madrid, Spain

Goicoechea-de-Jorge, Maria
mgoico@filol.ucm.es
Universidad Complutense de Madrid, Spain

Sanz-Cabrerizo, Amelia
amsanz@filol.ucm.es
Universidad Complutense de Madrid, Spain

Sierra-Rodriguez, Jose-Luis
jlsierra@fdi.ucm.es
Universidad Complutense de Madrid, Spain

1. Introduction

The Complutense University has been one of the first European universities that has collaborated with Google's project[1] by putting on the Web 100,000 volumes from its ancient fund. However scholars notice that these digitized texts are often of no much use to professors-researchers-students in literature unless additional tools are provided, to enhance the educational and research value of this material. In particular, the ability of making annotations on these texts has been largely recognized as a basic mean of adding value to this kind of digitized resources (Rios da Rocha et al. 2009). In this paper we present the annotation model used in @Note 1.0, a system developed at UCM funded by the Google's 2010 Digital Humanities Award program.

@Note 1.0 allows us to retrieve digitized works from Google Books collection and add annotations to enrich the texts with research and learning purposes: critical editions, reading activities, e-learning tasks, etc. One of the main features of @Note annotation model, which distinguishes it from similar approaches (Azouaou & Desmoulins 2006; Bechhofer at al. 2002; Koivunen 2005; Rios da Rocha et al. 2009; Schroeter et al. 2006; Tazi et al. 2003), is to promote the collaborative creation of annotation schemas by communities of researchers, teachers and students, and the use of these schemas in the definition of annotation activities on literary works. It results in a very flexible and adaptive model, able to be used by many different communities of experts in literature defending different critical literary theories and for different annotation tasks. In this paper we present this annotation model.

2. The @Note Annotation Model

2.1. Structure of the model

The structure of the @Note annotation model is summarized in the UML class diagram (Booch et al. 2005) of Fig. 1. In this model:

- *Annotation management communities* are groups of *annotation managers*, experts in literature (teachers, researchers, etc) who act as administrators to create activities, to select works and to organize activity groups.
- *Annotation communities*, in their turn, are groups of *annotators*, students / pupils interested in literature who perform proposed annotation activities.
- Each a*nnotation activity* comprises (i) a *digitized work*, (ii) a *metalevel-oriented annotation schema*, (iii) a *work-oriented annotation schema*.
- In this context, the *works* are the literary texts that can be annotated during the annotation activities. *Annotations*, in their turn, are characterized by: (i) an *annotation anchor* (the region of the work to which the annotation refers), (ii) an *annotation content* (a free rich-text piece that actually configure the annotation), (iii) a set of *annotation types* (semantic qualifiers for annotations) chosen from the annotation schemas attached to the annotation activity (at least one from the metalevel-oriented annotation schema).

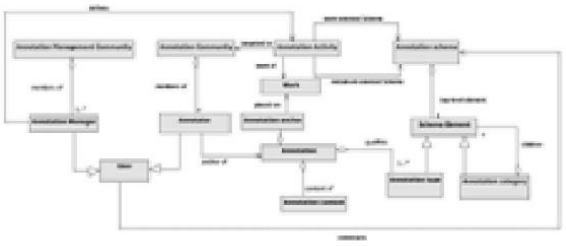


Figure 1: @Note information model

- The *annotation schemas* are explicit formalization of the types of annotations that can be carried out on works. In @Note, annotation schemas are hierarchies formed by annotation types and *annotation categories* (sets of annotation types and/or others, more specifically, annotation categories). In their turn, they can be *metalevel-oriented annotation schemas* (schemas which usually comprise concepts concerning particular literary theories around which the annotation activities are articulated), or *work-oriented annotation schemas* (schemas that capture aspects relative to the relationships between annotations and their anchors). While schemas of the first type are created by annotation managers, schemas of the second type are usually created by annotators.

In the context of the annotation management community, an annotation schema can be *public* or *private*. A private schema is only accessible for the annotation manager who created it. On the contrary, a public schema is accessible for all the annotation managers. Annotation managers have unlimited privileges on all the schemas to which they can access (i.e., they can create new annotation types and categories, they can blend two different types/categories in a single one, they have renaming and erasing privileges, etc), with the exception of modifying the *public/private* character (it only can be done by the schema's creator). In addition, when annotation managers create annotation activities, they only can choose those schemas to which they have access grants. Concerning annotators, they can add new types and categories to the book-oriented annotation schema, but they can't perform any other modification.

Figure 2: Example of rules governing the annotation process (informally described using natural language)

2.2. The annotation process

The @Note annotation process governs how to create the different types of information elements envisioned in the annotation model. For this purpose, @Note introduces a set of rules governing aspects like information visibility, creation and modification privileges of annotations and annotation schemas, etc. Although, by lack of space, these rules will not be detailed here, in Fig. 2 we include an example concerning an informal description of some of the rules governing the management of annotation schemas.

2.3. Annotation browsing and recovering

Annotation schemas in @Note are seen as T-*boxes* of description logic theories (Brachman & Levesque 2004). For instance, Fig 3a shows, edited in @Note, a fragment of the annotation schema used at UCM in an English Literature introductory course, while Fig 3b despites the description logic's counterpart. This simple interpretation is still powerful-enough to enable powerful *annotation browsing* and *annotation recovering* behavior. Indeed:

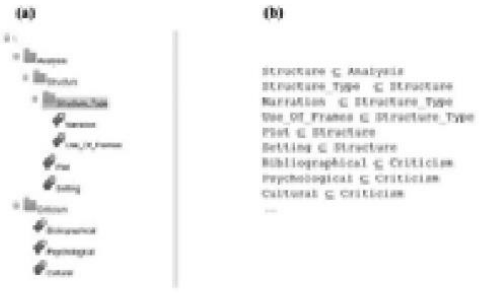


Figure 3: (a) A fragment of annotation schema (b) description logic counterpart

- Annotations can be browsed using annotation schemas, in a similar way to a folder explorer in a filesystem. In each step, there is a category or a type selected, and the user will *see* all the annotations entailed by such a selected element.
- Queries consist of arbitrary Boolean formulae involving annotation types and categories, being the outcomes the annotations entailed by such formulae.

In both cases, since entailment will be performed according to the description logic principles, the process will take into account the *is-a* relationship made explicit by the annotation schema.

2.4. Some technical details

The system has been entirely developed using Google technologies for the development of Rich-Internet Applications (RIAs) (Fraternali et al. 2010): GWT in the client side and the Google App Engine's facilities in the server side (Unruh 2010). Fig. 4 shows some snapshots of the system. The current version runs on the fully free-access books integrated in Google Books, and, in particular, on the UCM-Google collection. The works retrieval is achieved by the use of the Google Books API through REST (Richardson & Ruby 2007), and then presented to clients in an asynchronous way to keep them responsive to their events.

Figure 4: Some snapshots of @Note

3. Conclusions and Future Works

@Note promotes a fully collaborative annotation process, in which not only literary works are collaboratively annotated, but also annotation schemas are collaboratively created. The @Note system has been evaluated at UCM by several researchers and students in literature. They highlighted the flexibility of the annotation model, and, in particular, the ability to create and share annotation schemas tailored according to different critical perspectives and annotation activities. Additionally, they appreciate a sufficient expressive power from a browsing and recovering point of view. They also remarked the educational potential of the tool, although some advanced features could add some conceptual difficulties for students.

Currently we are working to adapt the annotation tool in order to facilitate its connection to a repository of learning objects, so as to allow the storage of literary texts' annotations as learning objects (Polsani 2003), and to make possible the recovery of those annotations and move about them according to the associated metadata. Thus, we are developing a communal working space for the creation of written compositions in different traditions and languages. We are also experimenting with the students' capability for developing their own catalogues, annotating the literary texts according to them and reusing their annotations in the production of critical essays. Additionally, we are working on connecting our system with other digital libraries (in particular, with Hathi Trust[2]). Finally, we are planning to address interoperability issues, in order to enable the interchange of annotations according to some of the emerging standards proposed by the digital humanities community (e.g., OAC[3]).

Acknowledgements

This work has been funded by Google with a grant of the Google's 2010 Digital Humanities award program entitled *Collaborative annotation of digitalized literary texts*. Additionally, this work has been performed in the context of the project grants of the Spanish Ministry for Research and Innovation (FF12008-06924-C02-01 and TIN2010-21288-C02-01), UCM (PIMCDs 2010/177 and 2011/313) and Santander-UCM (GR 42/10 - 962022).

References

Azouaou, F., and C. Desmoulins (2006). MemoNote, a context-aware annotation tool for teachers. *Proceedings of the 7th International Conference on Information Technology Based Higher Education and Training* ITHET'06, Sidney, Australia, July 2006.

Bechhofer, S., L. Carr, C. Goble, S. Kampa, and T. Miles-Board (2002). The Semantics of Semantic Annotation. *Proceedings of the First International Conference on Ontologies, DataBases, and Applications of Semantics for Large Scale Information Systems* ODBASE'02, Irvine, CA, USA, October 2002.

Booch, G., J. Rumbaugh, and I. Jacobson (2005). *The Unified Modeling Language User Guide (2nd Edition)*. Boston: Adisson-Wesley.

Brachman, R., and H. Levesque (2004). *Knowledge Representation and Reasoning*. Amsterdam: Morgan-Kauffmann.

Fraternali, P., R. Gustavo, and F. Sánchez-Figueroa (2010). Rich Internet Applications. IEEE *Internet Computing* 14(3): 9-12.

Koivunen, M.-R. (2005). Annotea and Semantic Web Supported Collaboration. *Proceedings of the UserSWeb Workshop – 2nd European Semantic Web Conference*, Heraklion, Grece, June 2005.

Polsani, P. (2003). Use and abuse of reusable learning objects. *Journal of Digital Information* 3(4).

Richardson, L., and S. Ruby (2007). *Restful web services*. Beijing: O'Reilly.

Rios da Rocha, T., R. Willrich, R. Fileto, and S. Tazi (2009). Supporting Collaborative Learning Activities with a Digital Library and Annotations. *Proceedings of the 9th IFIP World Conference on Computers in Education* WCCE'09, Bento Gonçalves, Brazil, July 2009.

Schroeter, R., J. Hunter, J. Guerin, I. Khan, I., and M. A. Henderson (2006). Synchronous Multimedia Annotation System for

Secure Collaboratories. *Proceedings of the Second IEEE International Conference on e-Science and Grid Computing* e-Science '06, Amsterdam, Netherlands, December 2006.

Tazi, S., Y. Al-Tawki, and K. Drira K. (2003). Editing pedagogical intentions for document reuse. *Proceedings of the 4th IEEE Technology Based Higher Education and Training* ITHET'03, Marrakesh, Morocco, July 7-9, 2003.

Unruh, A. (2010). *Google App Engine Java and GWT Application Development*. Birmingham: Packt Publishing.

Notes

1. http://www.ucm.es/BUCM/atencion/25403.php
2. http://www.hathitrust.org/
3. http://www.openannotation.org/

HyperMachiavel: a translation comparison tool

Gedzelman, Séverine
severine.gedzelman@ens-lyon.fr
ENS de Lyon, France

Zancarini, Jean-Claude
jean-claude.zancarini@ens-lyon.fr
ENS de Lyon, France

1. Introduction

The HyperMachiavel project started with the idea of a tool that would aid research communities comparing several editions of one text and in particular comparing translations.

The Italian studies department (Triangle laboratory) at ENS de Lyon has been working for many years on fundamental texts, from Machiavelli, Guicciardini and other contemporary followers, that put forward new political concepts throughout Europe in the 16th century. The question addressed in the project was mainly about the transfer of these concepts from one language to another, and especially their reception in France. The first aligned corpora tested in our tool gathers different editions of Machiavelli's *Il Principe*, the *princeps edito* from Blado in 1532 and the first four French translations of the 16th century.

Inspired by machine translation and lexicographic domains, the system presented in this paper proposes an annotation environment dedicated to the edition of lexical correspondences and offers different views to assist humanities researchers in their interpretations of the quality and the specificities of translator's work.

2. Viewing and Searching in Aligned Corpora

2.1. Synoptic View

To be able to identify lexical correspondences, machine translation tools usually propose a frame of two panels, one for the source text and the other for the target text. The visualized interface is meant for annotators to easily revise the results obtained from automatic word alignment. In general it only considers a pair of texts at a time.

In the world of digital editions, text comparison has always been of great interest and the request to view diplomatic vs normalized transcriptions, or simply

6.2 @Note: an electronic tool for academic readings

Cita Completa:

Gayoso-Cabada, J., Sanz-Cabrerizo, A., & Sierra, J.-L. (2013). @Note: An Electronic Tool for Academic Readings. En Proceedings of the 1st International Workshop on Collaborative Annotations in Shared Environment: Metadata, Vocabularies and Techniques in the Digital Humanities, DH-CASE 2013 (paper 17, 4 pags). Florence, Italy: ACM.

Resumen original de la contribución:

@note is a rich-internet application for the collaborative annotation of digitized literary texts. It enables the collaborative definition of annotation activities by a community of annotation managers, as well as the collaborative accomplishment of these activities by communities of annotators. For this purpose, @note lets annotator managers define the different components of annotation activities, among which it is possible to distinguish annotation schemata: conceptual structures abstracting the semantics of annotations that are collaboratively defined and that can be used to catalogue the annotations. Annotations themselves are conceived as discussion forums. Later they can be filtered and browsed according to the annotation schemata, and they can be organized in exportation templates and exported in multiple formats (p.e., RTF or HTML). This demo will be focused on the operation of @note, both from the perspective of annotators, who collaboratively annotate digitized texts, and annotator managers, who collaboratively define annotation activities.

Referencias Bibliográficas:

(Azouaou & Desmoulins, 2006; Bechhofer et al., 2002; Da Rocha et al., 2009; Fraternali et al., 2010; Joaquin Gayoso-Cabada et al., 2012; Guermeur & Unruh, 2010; Koivunen, 2005; Polsani, 2006; Richardson & Ruby, 2008; Ruiz et al., 2012; Schroeter et al., 2006; Tazi et al., 2003; Yang, 2010)

@note: an electronic tool for academic readings

Joaquín Gayoso-Cabada
Fac. Informática
Univ. Complutense de Madrid
C/ Prof. José García Santesmases s/n
28040 Madrid (Spain)
+34913947548
jgayoso@ucm.es

Amelia Sanz-Cabrerizo
Fac. Filología
Univ. Complutense de Madrid
Ciudad Universitaria s/n
28040 Madrid (Spain)
+34913945401
amsanz@ucm.es

José-Luis Sierra
Fac. Informática
Universidad Complutense de Madrid
C/ Prof. José García Santesmases s/n
28040 Madrid (Spain)
+34913947548
jlsierra@ucm.es

ABSTRACT
@note is a rich-internet application for the collaborative annotation of digitized literary texts. It enables the collaborative definition of annotation activities by a community of *annotation managers*, as well as the collaborative accomplishment of these activities by communities of *annotators*. For this purpose, *@note* lets annotator managers define the different components of annotation activities, among which it is possible to distinguish *annotation schemata*: conceptual structures abstracting the semantics of annotations that are collaboratively defined and that can be used to catalogue the annotations. Annotations themselves are conceived as discussion forums. Later they can be filtered and browsed according to the annotation schemata, and they can be organized in exportation templates and exported in multiple formats (e.g., RTF or HTML). This *demo* will be focused on the operation of *@note*, both from the perspective of annotators, who collaboratively annotate digitized texts, and annotator managers, who collaboratively define annotation activities.

Categories and Subject Descriptors
J. [**Computer Applications**]: (J.5) Arts and Humanities – Literature. K. [**Computer Milieux**]: (K.3) Computers and Education – (K.3.1) Computer Uses in Education – Collaborative Learning. H. [**Information Systems**]: (H.3) Information Storage and Retrieval – (H.3.7) Digital Libraries – User issues

General Terms
Human Factors

Keywords
Collaborative annotation, reading activity, e-library, e-learning

1. INTRODUCTION
@note [4] is a rich-internet application (RIA) developed at Complutense University of Madrid and funded by the Google's 2010-2011 Digital Humanities Award programs. This application is oriented to the collaborative annotation of digitized literary texts, and supports two levels of collaboration:

- On one hand, communities of *annotators* use *@note* for working together in *annotation activities* oriented to annotate digitized literary works, and for organizing the annotations using suitable *annotation schemata*.
- On another hand, the definition of annotation activities is collaboratively carried out in *@note* by communities of *annotation managers*, who collaborate in the definition of the different aspects of such activities (and, in particular, in the definition of the annotation schemata).

This demo will focus on the different aspects of the application. In Section 2 the application is presented from the perspective of its different users (annotators and annotation managers). Section 3 summarizes some details concerning *@note* internals. Section 4 presents some results of using *@note* in educational settings. Finally, Section 5 outlines some conclusions and lines of future work.

2. The *@note* Collaborative Annotation Tool
This section describes the main features of *@note* from a user perspective. Subsection 2.1 is focused on annotators, while Subsection 2.2 is focused on annotation managers.

2.1 Annotators' Perspective
As indicated earlier, annotators use *@note* to collaborate in annotation activities. The main constituents of an annotation activity are:

- The digitized literary work to be annotated. In its current version, *@note* makes possible the selection of digitized texts from Google Books collection, as well as from a local library. However, the extension of the tool for supporting other sources of digitized texts is straightforward. Indeed, with the exception of the books stored in the local library, annotations in *@note* are kept separated from the annotated texts.
- The annotation schemata to be used for organizing the annotations. In *@note*, annotation schemata are hierarchical arrangements of *annotation types* (atomic concepts that can be used to catalog annotations) and *annotation categories* (composite concepts used to contain annotation types and / or simpler annotation categories). Figure 1a shows a fragment of such an annotation schema in *@note*. Indeed, an annotation activity in *@note* can comprise two different types of annotation schemata: a *metalevel-oriented annotation schema*, and a *work-oriented annotation schema*. Metalevel-oriented schemata are provided and maintained by annotation managers, and usually represent concepts concerning particular literary theories around which the annotation activities are articulated. Therefore, these schemata cannot be modified by annotators. On the other hand, work-oriented schemata capture aspects relative to the relationships between annotations and their anchors. On the contrary to metalevel-oriented schemata, annotators are allowed to add new annotation types and categories to work-oriented schemata.
- An optional *exportation template*. Templates constitute simple document models, conceived as hierarchical

DH-Case'13, September 10, 2013, Florence, Italy.

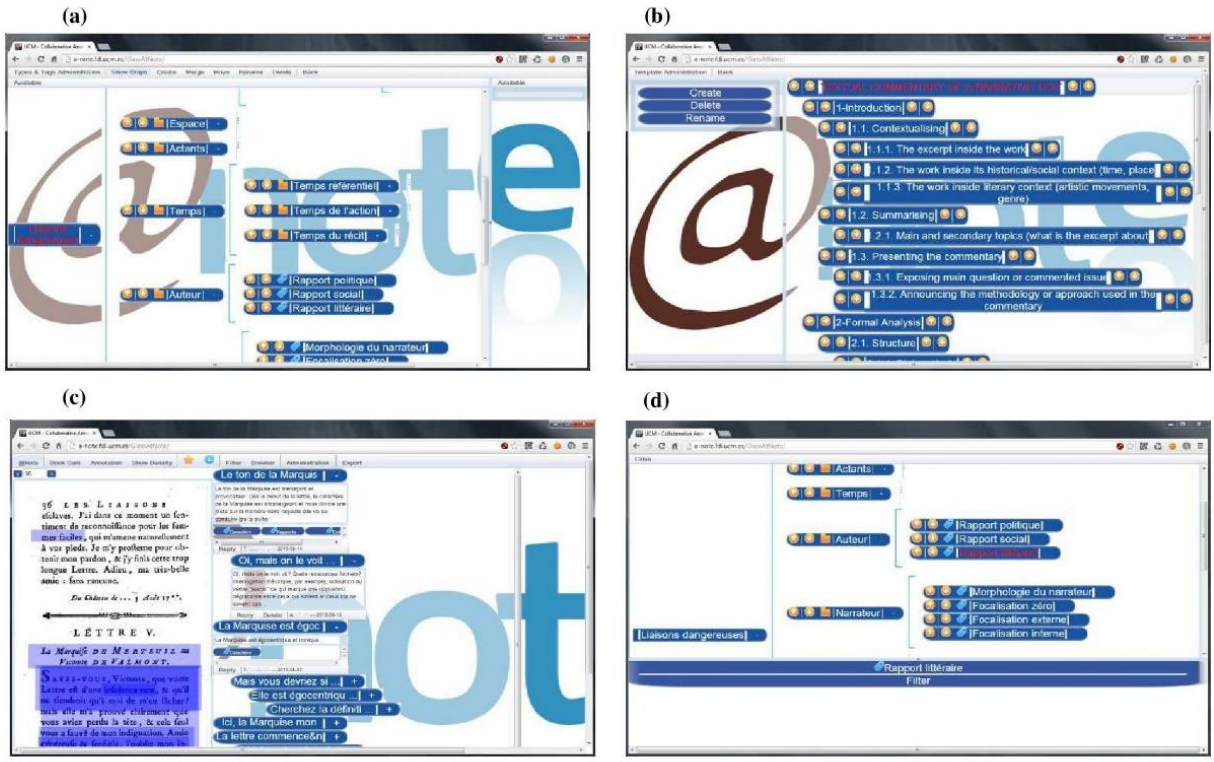


Figure 1 (a) Fragment of an annotation schema; (b) an exportation template; (c) an annotation; (d) annotation schemata-based browsing.

arrangements of *sections* (Figure 1b). Such templates can be used by annotators to annotators to prepare documents with the contents of selected annotations

Thus, when carrying out annotation activities, annotators collaboratively create annotations on the texts associated to the activities, as well as contribute to the annotations created by others. In *@note*, annotations are characterized by an *anchor* (the region in the text to which the annotation refers), the actual *content* of the annotation, and a set of *annotation types* that catalog the annotation. *@note* makes it necessary to associate with each annotation a type chosen from the metalevel-oriented annotation schema of the activity. Besides this constraint, annotators are free to choose other annotation types from this schema, as well as from the work-oriented one.

@note includes a complete permission system, which lets annotation creators to limit the operations that other annotations can perform on their annotations (they can define annotations *private*, *public* and *publicly editable*).

Another interesting feature of *@note* is to enable content of annotations structured as a discussion forum (Figure 1c). In this way, annotators can contribute to an annotation by adding new contributions to the associated forum, which, in turn, are arranged in a tree-shaped structure. Each contribution, in its turn, is a piece of free rich-text that, in addition to text, can include references to images, video, external URLs, etc.

In order to facilitate collaboration, *@note* makes it possible to browse the annotated text of an annotation activity by using the annotation schemata in a similar way to a folder explorer in a file system (Figure 1d). In addition, it makes it possible to perform more sophisticated searches based on more complex Boolean conditions involving both annotation types and annotation categories. Thus, using the collaborative catalog of annotations created by their peers, annotators can find interesting annotations and then add their contributions.

Finally, as indicated before, annotators can create skeletons of documents by selecting annotations. For this purpose, they fill the different sections of the exportation template with annotations localized using the described browsing and searching facilities. Once the desired annotations have been pushed into the template, annotators can get a document containing the data and contents of the annotations. Currently, *@note* supports the generation of HTML and editable RTF documents. This feature is especially relevant in educational settings, where students can use the generated documents as preliminary drafts for preparing their works summarizing their readings of the documents. As with the inclusion of support for new libraries, doting to *@note* of support for new document formats is a straightforward programming task.

2.2 Annotation Managers' Perspective

The main task of annotation managers is to collaborate in the creation of annotation activities. For this purpose, they can collaboratively define all the elements making up these activities. In particular, and in addition to choose the works to be annotated (Figure 2a), annotation managers can collaboratively define the annotation schemata to be incorporated in such activities. Indeed, this feature (the possibility of collaborative creation of annotation

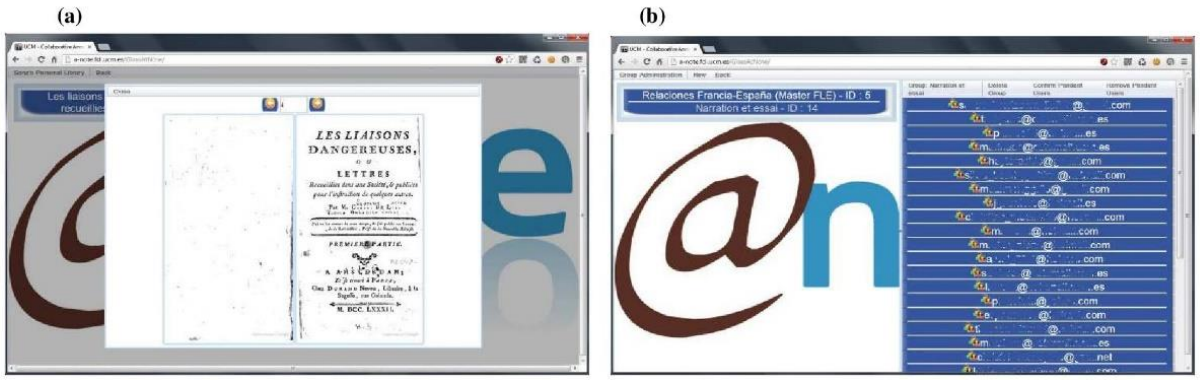


Figure 2 (a) Visualization of a book to be included in an activity; (b) Formation of a group of annotators

schemata) distinguishes @note from similar approaches [1,2,5,8,10,11]. As a consequence, it facilitates the use of *@note* by many different communities of experts in literature defending different critical literary theories and for different annotation tasks.

In addition to select the digitized texts and to define the annotation schemata, annotation managers can define the other aspects of annotation activities: constituting groups of annotators and assign these groups to the activities (Figure 2b), setting up the labels and texts shown in the annotation user interface in order to tailor it to the particular cultural idiosyncrasy of annotators (e.g., by adapting the interface to Spanish, French, etc.; the default language is English), or choosing the exportation templates to be used in the activities.

Finally, as in the case of annotations, annotator managers have available a complete repertory of permissions, making it possible, for instance, to maintain an annotation schemata, an exportation template, a digitized text uploaded in the local library, or even a complete annotation activity private, public, or publicly editable.

3. Some Technical Details

As indicated earlier, *@note* has been designed as a RIA [3] in order to provide users with a satisfactory user experience. For this purpose, the client side of the application was developed using the Google Web Toolkit (GWT) [12], a Google's generative technology that makes it possible to generate dynamic web interfaces (i.e., HTML + JavaScript) from Java code, which substantially facilitates the development, portability of maintenance of the resulting applications. On the other hand, the server side has been programmed as a set of GWT-compliant services. Persistence, in its turn, has been provided using the standard Java Persistence API (JPA) [13], which facilitates the deployment of the application in different environments (indeed, the first version of *@note* was hosted in the Google App Engine -GAE- [12], and then it was ported to an independent server with a minimum of effort). Finally, *@note* exposes many of its functionality using RESTFul web services [7][9] .

4. Teachers' experience

The primary use of *@note* has been in education. The digital annotation process, based on digitized documents, makes possible new learning environments in university providing an interaction between the teacher and the students that substantially differs from the current media, and which allow teachers to exploit social learning skills through collaborative annotation activities. Three teachers at Complutense University are working currently with *@note*. Their experiences aim to evaluate the versatility and the efficiency of our tool. Next some of the questions asked to 87 students in French, English and Spanish Literary Studies after trying *@note* and their answers are presented:

- *Were annotations of literary texts important for you before this experience*? As evidenced in Figure 3a, the outcome of this answer was mostly positive (almost 70% of students considered annotation as an important or a very important activity for their learning process).
- *Is @note a useful tool for you?* Again the outcome was very positive, since almost 80% of students interviewed considered *@note* a useful or very useful tool (Figure 3b).
- *Do you think @note can help you to improve your literary reading?* 80% of students gave a positive response to this item, thus considering *@note* as a valuable instrument to help them to improve their ability to read literature (Figure 3c)
- *Have the tags been useful for you?* Although concerning this item the response was not so spectacularly positive, around 70% of the students considered the cataloguing feature of *@note* useful or very useful, which, taking into account the difficulty and more effort-demanding character of the feature, makes up a very positive and promising outcome (Figure 3d).
- *Is it interesting to collaborate with your class-mates for this kind of activities?* Again almost 80% of students gave a positive response to this item, which also can be considered as a very positive and stimulating outcome (Figure 3e).

5. Conclusions and future Works

@note makes possible the collaboration of not only annotators engaged in annotation activities, but also of annotation managers involved in the conception, definition and customization of such activities. In particular, the support of *@note* for the collaborative definition of annotation schemata constitutes one of its main features. *@note* has been used at Complutense University with educational purposes. Preliminary evaluation results make the educational potential of the tool apparent, regardless of the intrinsic difficulty of more advanced features (e.g., cataloguing and creation of work-oriented annotation schemata).

Currently we are working on connecting *@note* with a repository of learning objects, in order to make it possible to store

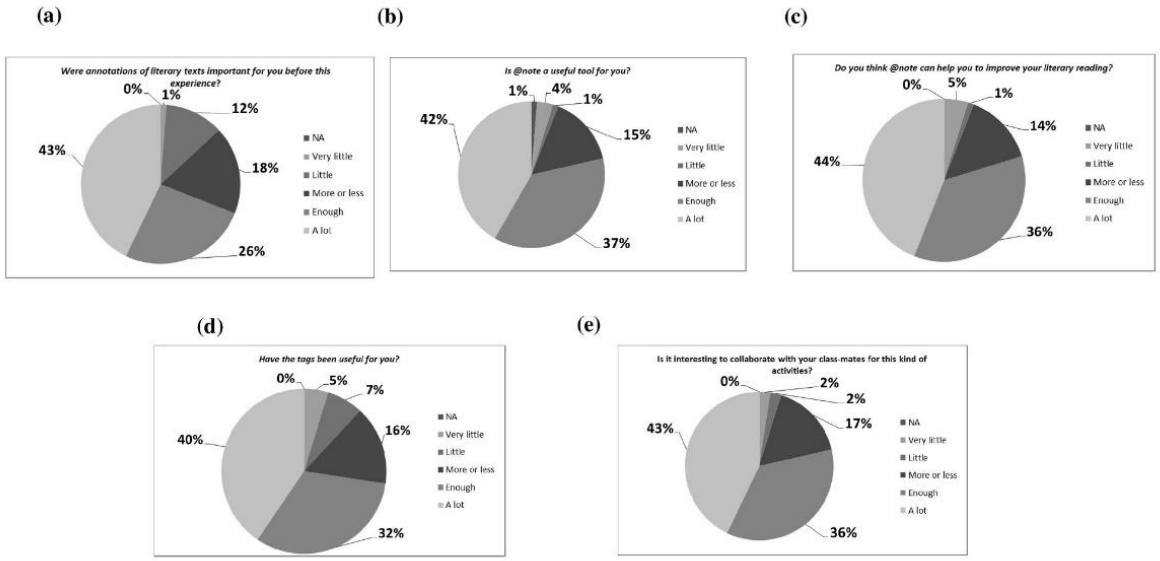


Figure 3 Answers to some of the questions of the survey done to students using *@note* at Complutense University.

annotations as learning objects that can be subsequently deployed on standard e-learning platforms [6]. We are also planning to connect *@note* to other digital libraries (in particular, Hathi Trust[1]). Finally, we are planning to incorporate to *@note* support for standard annotation representation formats (e.g., OAC[2]).

6. Acknowledgements

This work has been funded by Google with a grant of the Google's 2010-2011 Digital Humanities award program entitled Collaborative annotation of digitalized literary texts. Additionally, this work has been performed in the context of the project grants of the Spanish Council for Research and Innovation (FFI2012-34666 and TIN2010-21288-C02-01). We also would like to thank César Ruiz for his work in the development of *@note*, as well as to the members of ILSA and LEETHI research groups for their effort in conceiving and evaluating the application.

7. References

1. Azouaou, F., Desmoulins, C. MemoNote, a context-aware annotation tool for teachers. 7th Int. Conf. on Information Technology Based Higher Education and Training (ITHET´06). 2006.
2. Bechhofer, S., Carr, L., Goble C., Kampa, S. and Miles-Board T. The Semantics of Semantic Annotation. In ODBASE: First Int. Conf. on Ontologies, DataBases, and Applications of Semantics for Large Scale Information Systems. 2002.
3. Fraternali, P., Gustavo, R., Sánchez-Figueroa, F. Rich Internet Applications. IEEE Internet Computing 14(3), 9-12. 2010
4. Gayoso, J., Ruiz, C., Pablo, L., Sarasa, A., Goicoechea, M., Sanz, A., Sierra J.L. A Flexible Model for the Collaborative Annotation of Digitized Literary Works. Proc. of the 2012 Digital Humanities Conference. 2012
5. Koivunen, M-R. Annotea and Semantic Web Supported Collaboration. UserSWeb Workshop. 2nd European Semantic Web Conference. 2005.
6. Polsani, P. Use and abuse of reusable learning objects. Journal of Digital Information, 3(4). 2003
7. Richardson, L., Ruby, S. Restful web services. O'Reilly. 2007
8. Rios da Rocha, T., Willrich, R., Fileto, R., Tazi, S. Supporting Collaborative Learning Activities with a Digital Library and Annotations. 9th IFIP World Conference on Computers in Education (WCCE 2009). 2009
9. Ruiz, C., Gayoso, J., Sarasa, A., Pablo, L., Sanz, A., Sierra, J.L. Web-services API in @Note. Proc. of the INTEREDITION Symposium on Scholarly Digital Editions, Tools and Infrastructure. 2012
10. Schroeter, R. Hunter, J. Guerin, J. Khan, I. Henderson, M. A Synchronous Multimedia Annotation System for Secure Collaboratories. Second IEEE International Conference on e-Science and Grid Computing (e-Science '06). 2006.
11. Tazi, S. Al-Tawki Y. and Drira K. Editing pedagogical intentions for document reuse, 4th IEEE Technology Based Higher Education and Training. 2003
12. Unruh, A. Google App Engine Java and GWT Application Development. Packt Publishing. 2010
13. Yang, D. Java Persistence with JPA. Outskirts Press. 2010

[1] http://www.hathitrust.org/

[2] http://www.openannotation.org/

6.3 Assessing semantic annotation activities with formal concept analysis

Cita Completa:

Cigarrán-Recuero, J., Gayoso-Cabada, J., Rodríguez-Artacho, M., Romero-López, M.-D., Sarasa-Cabezuelo, A., & Sierra, J.-L. (2014). Assessing semantic annotation activities with formal concept analysis. Expert Systems with Applications, 41(11), 5495–5508.

Resumen original de la contribución:

This paper describes an approach to assessing semantic annotation activities based on formal concept analysis (FCA). In this approach, annotators use taxonomical ontologies created by domain experts to annotate digital resources. Then, using FCA, domain experts are provided with concept lattices that graphically display how their ontologies were used during the semantic annotation process. In consequence, they can advise annotators on how to better use the ontologies, as well as how to refine these ontologies to better suit the needs of the semantic annotators. To illustrate the approach, we describe its implementation in @note, a Rich Internet Application (RIA) for the collaborative annotation of digitized literary texts, we exemplify its use with a case study, and we provide some evaluation results using the method.

Referencias Bibliográficas:

(Aroyo & Dicheva, 2004; Azouaou & Desmoulins, 2006; Baader, 2003; Bao, Zhou, & He, 2005; Baumeister, Reutelshoefer, Haupt, & Nadrowski, 2008; Baumeister, Reutelshoefer, & Puppe, 2011; Baumgartner, Flesca, & Gottlob, 2001; Baumgartner, Frölich, & Gottlob, 2007; Bendaoud, Toussaint, & Napoli, 2008; D. M. Berry, 2012; Borges, 1944; Brewster & O'Hara, 2004; Brin, 1998; Calhoun, 2013; Carpineto & Romano, 2004; Chi, Hsu, & Yang, 2005; Cigarrán, Gonzalo, Peñas, & Verdejo, 2004; Cigarrán, Peñas, Gonzalo, & Verdejo, 2005; Cimiano, Handschuh, & Staab, 2004; Cimiano, Handschuh, et al., 2004; Cimiano, Hotho, & Staab, 2005; Cimiano, Hotho, Stumme, & Tane, 2004; Cole & Eklund, 1996; Da Rocha et al., 2009; Dasiopoulou, Giannakidou, Litos, Malasioti, & Kompatsiaris, 2011; de Souza, Davis, & de Medeiros Evangelista, 2006; Devedzic, Jovanovic, & Gasevic, 2007; Di Donato et al., 2013; Dill, Eiron, Gibson, Gruhl, Guha, Jhingran, Kanungo, McCurley, et al., 2003; Dill, Eiron, Gibson, Gruhl, Guha, Jhingran, Kanungo, Rajagopalan, et al., 2003; Dingli, Ciravegna, & Wilks, 2003; DuBois, 2003; Etzioni et al., 2005; Fan & Xiao, 2007; Formica, 2006; Gamallo, Lopes, & Agustini, 2007; Ganter & Wille, 1999; Gayo, De Pablos, & Lovelle, 2010; Joaquin Gayoso-Cabada et al., 2012; Joaquín Gayoso-Cabada et al., 2013; Giannopoulos, Bikakis, Dalamagas, & Sellis, 2010; Greaves, 2004; Handschuh & Staab, 2002; Handschuh, Staab, & Ciravegna, 2002; Hogue & Karger, 2005; Hunter & Gerber, 2010; Huynh, Karger, & Quan, 2002; Jiang & Chute, 2009; Jiang, Ogasawara, Endoh, & Sakurai, 2003, p.; Jiang, Pathak, & Chute, 2009; Katifori, Halatsis, Lepouras, Vassilakis, & Giannopoulou, 2007; Keyser, 2007; Kim, Hwang, & Kim, 2007; Kiu & Lee, 2008; Kiyavitskaya, Zeni, Cordy, Mich, & Mylopoulos, 2009; Kiyavitskaya, Zeni, Mich, Cordy, & Mylopoulos, 2007; Kogut & Holmes III, 2001; Koivunen, 2005; Krötzsch, Hitzler, & Zhang, 2005; Krötzsch, Vrandečić, & Völkel, 2006; Kurilovas, Kubilinskiene, & Dagiene, 2014;

Kushmerick, 2000; Malik, Prakash, & Rizvi, 2010; Maynard, 2003; Mu, 2010; Noy et al., 2001; Oliveira & Rocha, 2013; Poelmans, Ignatov, Kuznetsov, & Dedene, 2013; Popov et al., 2003; Popov, Kiryakov, Ognyanoff, Manov, & Kirilov, 2004; Reeve & Han, 2005; Richards, 2004, 2006; S. Rudolph, 2004; Sebastian Rudolph, 2008; Sebastian Rudolph, Völker, & Hitzler, 2007; Schroeter et al., 2006; Sertkaya, 2009; Šimko, Tvarožek, & Bieliková, 2013; Soon & Kuhn, 2004; Stumme & Maedche, 2001, p.; Tazi et al., 2003; Tiropanis, Davis, Millard, & Weal, 2009; Vargas-Vera et al., 2002; Vargas-Vera, Moreale, Stutt, Motta, & Ciravegna, 2007; Völker & Rudolph, 2008; Wille, 1992, 2009; H. Xu & Xiao, 2009; W. Xu, Li, Wu, Li, & Yuan, 2006; Zhao, Halang, & Wang, 2007)

Expert Systems with Applications 41 (2014) 5495–5508

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Assessing semantic annotation activities with formal concept analysis

Juan Cigarrán-Recuero [a], Joaquín Gayoso-Cabada [b], Miguel Rodríguez-Artacho [a], María-Dolores Romero-López [c], Antonio Sarasa-Cabezuelo [b], José-Luis Sierra [b,*]

[a] Escuela Técnica Superior de Ingeniería Informática, Universidad Nacional de Educación a Distancia, 28040 Madrid, Spain
[b] Facultad de Informática, Universidad Complutense de Madrid, 28040 Madrid, Spain
[c] Facultad de Filología, Universidad Complutense de Madrid, 28040 Madrid, Spain

ARTICLE INFO

Keywords:
Semantic annotation
Formal concept analysis
Ontology
Annotation tool

ABSTRACT

This paper describes an approach to assessing semantic annotation activities based on formal concept analysis (FCA). In this approach, annotators use taxonomical ontologies created by domain experts to annotate digital resources. Then, using FCA, domain experts are provided with concept lattices that graphically display how their ontologies were used during the semantic annotation process. In consequence, they can advise annotators on how to better use the ontologies, as well as how to refine these ontologies to better suit the needs of the semantic annotators. To illustrate the approach, we describe its implementation in @note, a Rich Internet Application (RIA) for the collaborative annotation of digitized literary texts, we exemplify its use with a case study, and we provide some evaluation results using the method.

1. Introduction

The enormous efforts to digitize physical resources (documents, books, museum exhibits, etc.), along with recent advances in information and communication technologies, have democratized access to a cultural, scientific and academic heritage previously available to only a few. Likewise, the current trend is to produce new resources in a digital format (e.g., in the context of social networks), which entails an in-depth paradigm shift in almost all the humanistic, social, scientific and technological fields. In particular, the field of the humanities is one which is going through a significant transformation as a result of these digitalization efforts and the paradigm shift associated with the digital age. Indeed, we are witnessing the emergence of a whole host of disciplines, those of Digital Humanities (Berry, 2012), which are closely dependent on the production and proper organization of digital collections.

As a result of the undoubted importance of digital collections in modern society, the search for effective and efficient methods to carry out the production, preservation and enhancement of such digital collections has become a key challenge in modern society (Calhoun, 2013). In particular, the annotation of resources with metadata that enables their proper cataloging, search, retrieval and use in different application scenarios is one of the key elements to ensuring the profitability of these collections of digital objects. While the cataloging and retrieval of resources (whether digital or non-digital) have been the object of study in library sciences for decades (Calhoun, 2013), modern applications require annotating resources in semantically richer and more flexible ways, in many cases allowing multiple alternative annotations in the same collection. In consequence, the tendency is to introduce the use of ontology-based semantic technologies, in addition to conventional metadata schemas (Keyser, 2012).

While in recent years we have witnessed significant advances in the automatic annotation of resources, in particular of those with heavy text content (see Section 6), there are multiple scenarios in which resource annotation cannot be inferred from the contents of these resources (e.g., scenarios involving resources in which the content is not directly related to the meta–information required). In these cases it is necessary to involve human annotators in the semantic annotation of the resources. The resulting activities are referred to as *semantic annotation activities* in this paper. Some examples of semantic annotation activities are the annotation of digital educational resources (e.g., *learning objects*) in the eLearning domain (Aroyo & Dicheva, 2004; Devedzic, Jovanovic, & Gasevic, 2007; Kurilovas, Kubilinskiene, & Dagiene, 2014; Tiropanis, Davis, Millard, & Weal, 2009), the annotation of media content in the multimedia domain (Hunter & Gerber, 2010; Labra, Ordónez, & Cueva-Lovelle, 2010; Mu, 2010; Šimko, Tvarožek, & Bieliková, 2013), or the one chosen as a case study in this paper: the annotation of digitized literary texts (Azouaou & Desmoulins, 2006; Donato et al., 2013; Gayoso, Sanz, & Sierra, 2013; Gayoso et al., 2012; Koivunen, 2005; Rocha, Willrich, Fileto, & Tazi, 2009; Schroeter, Hunter, Guerin, Khan, & Henderson, 2006; Tazi, Al-tawki, & Drira, 2003).

* Corresponding author. Tel.: +34 913947548.
E-mail address: jlsierra@fdi.ucm.es (J.-L. Sierra).

http://dx.doi.org/10.1016/j.eswa.2014.02.036
0957-4174/

The main objective of any semantic annotation activity should be to produce an annotation of the resources in the underlying digital collection that satisfies all the requirements of accuracy, completeness and adequacy posed by the intended uses of the collection. Therefore, being able to assess to what extent these requirements are accomplished is an obligation in order to guarantee the quality of the final annotation outcomes. On one hand, the result of this assessment could help annotators to make a better use of the semantic models (i.e., the *annotation ontologies*) during the annotation of the resources. On the other hand, it could also be useful to the creators of the ontologies (i.e., the experts in the domain), who could identify how their ontologies should be modified, augmented or refined on the basis of the actual use of these assets during the annotation process. However, for huge collections or dense and semantically-rich annotations, the accomplishment of this assessment by individual inspection of every single annotated resource can become a titanic task. Therefore, providing automatic or semi-automatic assistance in the assessment of semantic annotation activities is an overriding concern in guaranteeing the quality of the annotations performed.

This paper addresses the formulation of mechanisms that support the assessment of semantic annotation activities, in order to enable: (i) better guidance of annotators during the annotation process, and (ii) the iterative refinement of the annotation ontologies. For this purpose, it presents a method of assessing the use of ontologies in semantic annotation activities, based on formal concept analysis (FCA). In this approach, annotators are provided with ontologies specifically designed by domain experts, and they use these ontologies to annotate a collection of digital resources. Then, the annotated collections are automatically analyzed using FCA to allow domain experts access to a lattice-based graphical representation that summarizes the overall annotation activity. This representation is linked to the concepts in the ontology so that at a glance, domain experts can assess how the proposed ontology is being used by annotators. Along with other aspects, they can see which concepts are not being used, which concepts are always used together, and which concepts are used more often than others. As a result, they can provide guidance to the annotators, enabling them to better use the ontologies proposed, or they can find aspects of the ontology that can be improved (e.g., several concepts might be combined into a single concept or they could include new concepts made apparent from the concept lattice). Therefore, and under reasonable assumptions, FCA provides domain experts with the machinery necessary to address the assessment of semantic annotation activities, at least to a semi-automatic extent.

The approach proposed in this paper has been successfully used in @note, a Rich Internet Application (RIA) for the collaborative annotation of digitized literary texts for educational purposes. In @note, teams of annotators (students, in this case) must complete the annotation of digitized literary works with free-text notes, and they must catalogue these notes using concepts taken from an ontology provided by the domain experts (teachers, in this case). Once the annotation activity is complete, and according to the aforementioned approach, @note allows teachers to examine how students performed the annotation activity by showing them a concept lattice created by considering notes as objects and ontology concepts as attributes in a formal context.

The remainder of this paper is organized as follows. In Section 2, we describe the annotation assessment approach. In Section 3, we describe its implementation in @note. In Section 4, we present a case study, i.e., an annotation activity of a literary work (*The Library of Babel*, a short story authored by the Argentinian writer Jorge Luis Borges). In Section 5, we present some evaluation results. In Section 6, we describe some related works. Finally, in Section 7, we present the conclusions and directions for future work.

2. The assessment approach

This section describes our approach to the assessment of semantic annotation activities using FCA. In Subsection 2.1, we summarize the elements of FCA required in the approach. In Subsection 2.2, we present an overview of such an approach. In Subsection 2.3, we describe the nature of annotation ontologies. Finally, in Subsection 2.4, we present the use of FCA to facilitate the assessment of annotation activities by domain experts.

2.1. The elements of FCA

The annotation assessment approach proposed in this paper relies heavily on the construction of concept lattices from annotated digital resources. As mentioned earlier, we use the well-known FCA technique. FCA is a mathematical theory of concept formation derived from lattice and ordered set theories that provides a theoretical model for organizing information and revealing relationships (Wille, 1992; Ganter & Wille, 1999; Carpineto, & Romano, 2004; Wille, 2009). The main construct of the theory is the *formal concept,* which is derived from a *formal context*.

A *formal context* can be defined as a set of objects, a set of attributes and a set of *is-a* or *has-a* relationships between objects and attributes. A formal concept is a pair (A,B), where A is a set of objects (also known as the *extent* of the formal concept), and B is a set of attributes (also known as the *intent* of the formal concept). The extent and the intent of a formal concept are connected as follows:

- The extent A consists of all the objects that are related to all the attributes in the intent B.
- The intent B consists of all the attributes shared by the objects in the extent A.

Formal concepts can be ordered by their extents. More formally, $(A,B) \subseteq (C,D) \Leftrightarrow A \subseteq C$; in this case, (C,D) is called a *super-concept* of (A,B) and, conversely, (A,B) a *sub-concept* of (C,D). This order relation is a generalization-specialization, and it can be proven to be a *lattice* (i.e., a concept lattice) based on the basic theorem of FCA (Ganter & Wille, 1999; Wille, 1992).

In a concept lattice, two important types of formal concepts are *object concepts* and *attribute concepts*:

- The *object concept* associated with an object o is the most specific concept that includes o in its extent. The intent of an object concept is defined by all the attributes of o, whereas the extent contains not only object o but also all those objects related to all the attributes of o.
- The *attribute concept* associated with attribute a is the most generic concept that includes a in its intent. Its extent contains all the objects with attribute a, and its intent is defined by all the attributes shared by the objects belonging to the extent set.

Because concept lattices are ordered sets, they can be displayed naturally in terms of *Hasse diagrams* (Ganter & Wille, 1999). In a Hasse diagram: (a) there is exactly one node for each formal concept; (b) if, for concepts $C1$ and $C2$, $C1 \subseteq C2$ holds, then $C2$ is placed above $C1$; and (c) if $C1 \subseteq C2$ but there is no other concept $C3$ such that $C1 \subseteq C3 \subseteq C2$, there is a line joining $C1$ and $C2$.

Fig. 1(a) shows an example of a formal context, and Fig. 1(b) shows its associated concept lattice using a Hasse diagram.[1] This example illustrates that Hasse diagrams are particularly useful for visualizing concept lattices; thus they will be used in our approach as the primary means of presenting lattices to domain experts. The

[1] Concept lattices in section 2 have been generated with the ConExp application (http://conexp.sourceforge.net/).

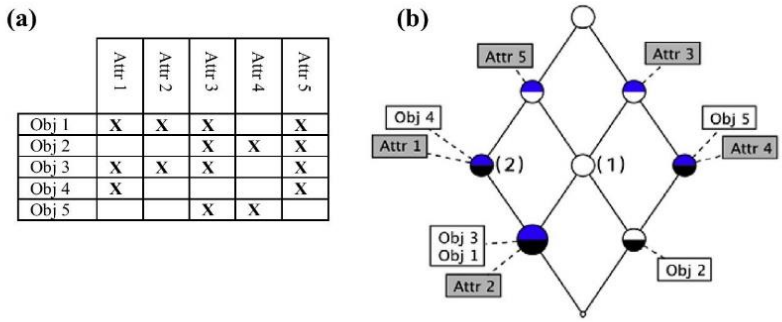


Fig. 1. (a) A sample formal context and (b) the concept lattice associated with the formal context in (a).

example also illustrates the method of marking the diagrams to facilitate the identification of formal concepts. To avoid an overloaded representation, each formal concept (i.e., a node in the diagram) is depicted with a minimal set of objects and a minimal set of attributes. Hence, each formal concept can be easily reconstructed from the diagram as follows:

- The extent is given by the union of all the objects depicted in the nodes on the paths leading from the formal target concept to the bottom concept in the diagram. For example, the extent of the formal concept associated with the node marked (1) in Fig. 1(b) is $\{obj1, obj2, obj3\}$.
- The intent is given by all the attributes depicted by the nodes on the paths from the formal target concept to the top node in the diagram. For example, in Fig. 1(b), the intent of the concept represented in node (1) is $\{attr3, attr5\}$.

In Fig. 1(b), concept (2) is an object concept of $obj4$ defined as $(\{obj1, obj3, obj4\}, \{attr1, attr5\})$ (object concepts in Fig. 1(b) are represented by coloring the lower half of the node), which means this concept is the most specific concept containing $obj4$ in its extent. Concept (2) is also the attribute concept of $attr1$ (attribute concepts are represented in Fig. 1b by coloring the upper half). Concept (1), described as $(\{obj1, obj2, obj3\}, \{attr3, attr5\})$, is neither an object nor an attribute concept.

Quantitative information can also be attached to each node (e.g., the absolute size of the extent of each concept, or its percentage with respect to the overall number of objects).

2.2. Overview of the approach

Fig. 2 outlines our approach to the assessment of annotation activities with FCA. This approach is inspired from our previous experiences using FCA for information retrieval (Cigarrán, Gonzalo, Peñas, & Verdejo 2004) and for browsing search results (Cigarrán, Penas, Gonzalo, & Verdejo 2005), as well as in our previous experiences in the collaborative authoring of annotation ontologies by domain experts during the design of annotation activities (Gayoso et al., 2012, 2013). The steps in the approach are as follows:

- The *domain experts design* the annotation activities. They begin by analyzing the collections of digital *resources* to be annotated and providing suitable formal *annotation ontologies* for them. They also create suitable *annotation guidelines* to be followed while performing the annotations.
- The *annotators* use the formal ontologies designed by the domain experts to perform the *annotation* of digital resources. During this process, they tag resources with one of several concepts selected from the ontology.
- Once the annotation process is complete, the resources annotated can be automatically analyzed by using FCA. In the *analysis*, the digital resources are the objects of a formal context, and the ontology concepts used to tag them are the attributes of the context. Thus, a *concept lattice* associated with the annotated collection can be automatically constructed.
- The concept lattice is then graphically inspected by the domain experts to assess how their ontologies were used for annotation. As result of this *assessment* stage, domain experts can contribute to improvements in the annotation process. On the one hand, they can advise annotators on the better use of the ontologies by refining the annotation guidelines. On the other hand, they can re-structure their ontologies to better accommodate the pragmatic needs of the annotators. Ontology re-structuring can include the elimination of unused concepts, the fusion of concepts that are commonly used together, the addition of new concepts revealed by the concept lattice, etc.

In addition, driven by the concept-lattice supported assessment, the process can be applied several times. The result is an ontology better suited to the real-world needs of digital-resource annotators.

2.3. Annotation ontologies

Our approach promotes the iterative provision of annotation ontologies. For this purpose:

- During the design step, domain experts provide an initial version of the annotation ontology. To this end, experts begin by characterizing the annotation activity itself. This annotation activity is characterized in terms of: (i) the digital resources to be annotated, (ii) the agents that must carry out the annotation process (i.e., the annotators), and (iii) the goals of the annotation (these goals can vary, depending on the application; typically, annotated resources enable semantic searching and retrieval and ontology-driven semantic browsing). Then, following a conventional, conceptualization-oriented, ontology design process, they provide an ontology specifically oriented to the features of the annotation task.
- The initial ontology is refined as a consequence of the *assessment* stage. As we will indicate later in the paper, this refinement is basically structural: adding new concepts associated to the combination of existing ones, removing useless concepts, melding equivalent concepts into a single one, etc. It lets domain experts solve structural design misconceptions and mistakes on the basis of the evidence gathered from the use of the ontology.

While our proposal is iterative in nature, both the initial provision of the ontology and its refinement after each assessment step

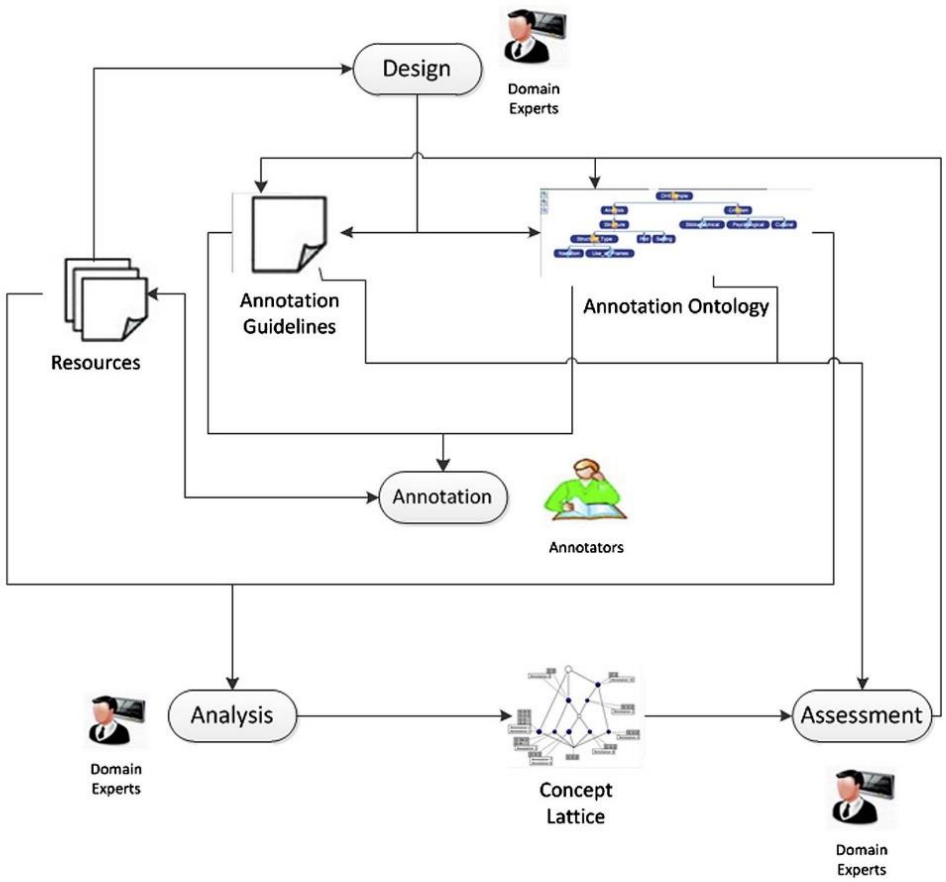


Fig. 2. Workflow of the assessment method.

require the basic conceptualization skills of domain experts. To make this assumption feasible, it is not reasonable to presuppose domain experts with advanced computer science knowledge, knowledge representation education, formal logic knowledge or experience in artificial intelligence, but educated professionals with profound knowledge of the requirements of a particular annotation task. Therefore, we constrain the shape of the ontologies to single concepts arranged in a multiple inheritance hierarchy. Fig. 3(a) shows an (abstract) example of this kind of ontology, while Fig. 3(b) details its description logic formalization (Baader, Calvanese, McGuinness, Nardi, & Paterl-Schneider, 2010). As this example makes apparent, although multiple inheritance is allowed (in the example, there is an ontology concept -i.e., C-34-1- that has more than one parent concept, i.e., C-4 and C-3), other kinds of relationships between concepts beyond the *is-a* relationship are intentionally avoided in order to facilitate the active engagement of domain experts in the design step. While it may hinder capturing more complex conceptualizations, hierarchies of this type constitute, on one hand, the skeleton of more sophisticated ontologies

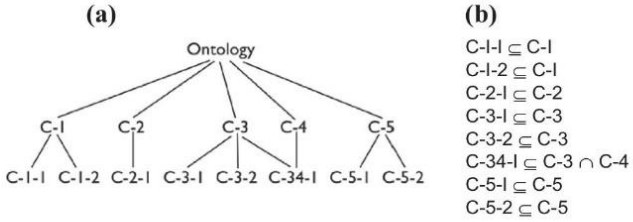


Fig. 3. (a) A simple annotation ontology and (b) description logic formalization of (a).

(Brewster & O'hara, 2004), and, on the other hand, are sufficiently simple to be authored by domain experts with easy-to-use hierarchy editors (Noy et al., 2001), who can determine the concepts required to organize the resources, as well as to arrange these concepts in meaningful taxonomies. (We also found this to be true in our experience with @note, where the gap between domain experts and computer science knowledge was especially noticeable.)

Finally, it is worthwhile to analyze the potential ontological disagreements among different domain experts in the context of our proposal. On one hand, when several domain experts work together on the definition of an annotation activity, our approach enforces the need to reach consensus before passing onto the annotation step. For this purpose, domain experts can take advantage of suitable collaboration mechanisms such as those available, for instance, in @note. On the other hand, although a particular annotation activity requires a single, consensual annotation ontology, it is important to point out that different domain experts can define different annotation ontologies for the annotation of the *same* body of digital resources. The point here is that each (possibly discordant) ontology is oriented to a *different* annotation activity. For instance, we have observed this frequently in our experience with @note, where several experts in literature, working on the same text with different (even opposing) purposes, defined different annotation ontologies, and thus different annotation activities. Once each activity was finished, the use of FCA allowed experts to assess to what extent the different ontologies accomplished aspects like adequacy to the intended annotation tasks, understandability and usability by annotators, etc. The conclusions obtained were very valuable for subsequent discussion among experts and for comparison of different (even divergent) approaches to the organization of digital resources.

2.4. Concept lattices for collections of resources annotated according to taxonomical ontologies

FCA can be straightforwardly applied to our approach by assigning digital resources as the objects and ontology concepts as the attributes. It is important not to confuse ontology concepts, which are one of our primary sources of information, with the formal concepts that are the final entities obtained by combining the digital resources and the ontology concepts via FCA. In this model, *ontology concepts* become *attributes* of the formal contexts.

As we consider our ontologies as taxonomies, we must also include this ontological assumption and define the formal context accordingly. For example, if an annotation is tagged with an ontology concept *OC*, it will also be associated with any other ontology concept *OC'* such that $OC \subseteq OC'$. With these assumptions, concept lattices are constructed as follows:

- The upper part of the lattice is formed by the attribute concepts associated with each ontology concept *OC* in the original ontology. Thus, it is easy to recognize the source ontology in the upper segment of the emerging concept lattice with the exception of the ontology concepts that: (a) were never used in the annotation process (i.e., they will appear in the bottom section of the lattice); (b) always co-occur in the same annotations (i.e., they share the same attribute concepts); and (c) although not related in the original ontology, are used together in some cases by the annotators. This structure facilitates the discovery of new relations not present in the original ontology via the final concept lattice.
- The bottom part of the lattice is associated with the formal concepts that are a combination of the upper attribute concepts related to primary ontology concepts. The latter concepts are very interesting, as they reflect how the annotators have combined the original ontology concepts.

To illustrate this process, Fig. 4(a) shows a formal context depicting a possible way in which the annotators used the ontology displayed in Fig. 3. The annotated resources are the formal objects and the ontology concepts are the formal attributes. At a glance, the formal context shows the ontology concepts shared by more than one annotation. The formal context maps not only the ontology concept used in the annotation but also the more generic ontology concepts that are super-concepts of the ones selected. For example, if the annotator used the ontology concept C-1-1 to annotate, the formal context will also show its most generic ontology concept as a relation (i.e., C-1). Fig. 4(b) shows the concept lattice and the 11 formal concepts obtained from the formal context. This lattice shows the following:

- The upper part maps the generic ontology concepts that have not been combined with any other ontology concept in any of the annotations. This means that, in those cases, the annotator applied the original ontology. This is the case of the attribute concepts C-1 and C-3. It can be read from the lattice that these ontology concepts have been used in isolation (i.e., for resources 9 and 10) or co-occurring with their ontology sub-concepts (i.e., for resources 3 and 4).
- Ontology concepts C-2 and C-5 always co-occur in the same annotations. In this situation, both ontology concepts will be merged into the same attribute concept: ({*C*-2,*C*-5,*C*-2-1,*C*-5-1}, {*Resource* 1, *Resource* 2}), suggesting that they have a close conceptual relation from the annotator's point of view. Thus, on the basics of the quantitative information, or by examining the actual annotations, domain experts could decide, e.g., to change the original ontology to reflect this situation, thus melding C-2 and C-5 into a single concept, or otherwise to instruct annotators to better clarify the distinction between C-2 and C-5.
- The formal concept {(*C*-3, *C*-4, *C*-34-1), (*Resource* 5)} correctly maps the nature of the ontology concept *C*-34-1 with the two aforementioned parents.
- The bottom concept shows the ontology concepts not used by the annotators, the ontology concept *C*-5-2 in this case. Domain experts can use this evidence to modify the ontology or to instruct annotators.
- The remainder of the formal concepts depicts situations where the annotators have combined conceptually different ontology elements. For example, ontology concept *C*-3-1 has been used in combination with ontology concept *C*-1-1 (for resources 6 and 7), as well as ontology concept *C*-3-2 (for resource 8). Eventually, domain experts could consider examining the lattice and the associated annotated resources to potentially refine the ontology by assigning suitable names to these emerging conceptual combinations.

All these considerations are examples of the iterative approach to the formulation of annotation ontologies derived from our approach: using FCA, experts are able to assess how ontologies were actually used by annotators; the resulting analysis helps perform structural refinements on these ontologies.

3. Assessment of annotation activities in @note

This section shows how the approach described in this paper has been implemented in @note, an application for the collaborative annotation of digitized literary works. In Subsection 3.1, we summarize the @note application. In Subsection 3.2, we describe the assessment of the @note annotation activities using FCA.

(a)

	C-1	C-2	C-3	C-4	C-5	C-1-1	C-1-2	C-2-1	C-3-1	C-3-2	C-34-1	C-5-1	C-5-2
Resource 1		X			X			X				X	
Resource 2		X			X			X				X	
Resource 3	X					X							
Resource 4	X						X						
Resource 5			X	X							X		
Resource 6	X		X			X			X				
Resource 7	X		X			X			X				
Resource 8	X		X			X				X			
Resource 9			X										
Resource 10	X												

(b)

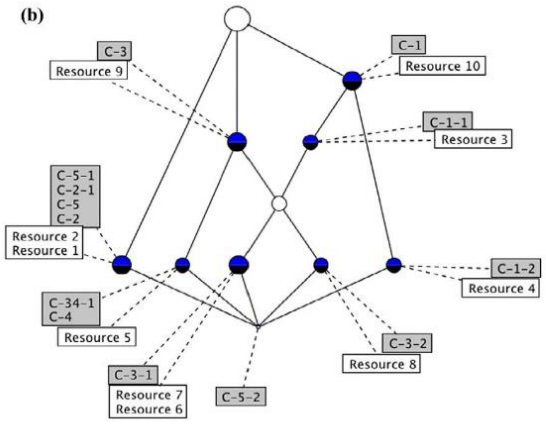


Fig. 4. (a) Formal context derived from the ontology in Fig. 3; (b) the concept lattice associated with the context in (b).

3.1. The @note application

The application @note is an RIA for the collaborative annotation of digitized literary texts (Gayoso et al., 2012, 2013). This collaborative annotation tool adds free-text notes to literary works and classifies these notes with concepts selected from ontologies specifically devised for the annotation of the specific works. As a result, @note enables the collaborative creation of specialized knowledge bases of notes added to a literary work for a given purpose (e.g., to enable critical reading).

The keystone concept in @note is the *annotation activity*. An annotation activity is primarily characterized by the literary work to be annotated, and the annotation ontology to be used during the annotation of the work. These activities are defined by experts in literature (teachers, researchers, etc.), who can collaborate in the selection of the volume to be annotated and, more importantly, in the definition of the annotation ontology. In addition, these activities are targeted to groups of annotators (students, other scholars, etc.) responsible for adding and cataloguing notes.

Following the assumptions used in this study, annotation ontologies in @note are hierarchical arrangements of concepts organized according to an *is-a* relationship. The @note application organizes these concepts into two different groups as follows:

- *Annotation types*, which are the more specific concepts (the leaves in the hierarchy), and are actually those concepts that can be used for classifying notes; and
- *Annotation categories*, which are more general concepts (inner nodes in the hierarchy) that can be specialized in other, simpler categories and/or annotation types. These concepts are used solely for structuring purposes. (@note does not allow them to be used directly for semantically describing notes, although they can be used for searching and browsing.)

Fig. 5(a) shows an excerpt of an annotation ontology in @note, and Fig. 5(b) shows its description logic counterpart.

The ontological bias in @note has been shown to be sufficiently useful to experts and annotators in literature and simple enough to facilitate the collaborative authoring of annotation ontologies by experts. The experts can share their ontologies with other experts and directly edit these ontologies by using a simple tree-based editor integrated in the RIA (Fig. 6(a)). Annotators can also interact easily with the annotation ontologies by using a graph-like view (Fig. 6(b)); see the works of Gayoso et al. (2012, 2013) for more details on the @note functionalities.

3.2. Using FCA to assess annotation activities in @note

To support the assessment approach described in this paper, @note associates a formal context with each annotation activity as follows:

- The set of objects is composed of the set of all the free-text notes added to the literary work (i.e., the digital resources in this case are composed of the digital notes added to the literary works), and
- Following the directives described in Section 2, the attributes associated with each note are composed of all the annotation types that tag the note, as well as all the annotation categories in which these types are included.

Experts can further project this formal context on subgroups of annotators, and even on the notes authored by individual annotators, to better assess the use of the ontology by the subgroups and individual users. Thus, by applying FCA in these formal contexts, @note is able to display how the annotation activity was performed by using Hasse diagrams (see Fig. 7(a)). Each node in the diagram displays the set of new ontology concepts in the annotation ontologies referred to by the node. As shown in Fig. 7(a), such a set can be empty (as it is in the case of newly-formed concepts not present in the original ontology). Each node can be expanded to show complete information about the formal concept (see Fig. 7(b)) as follows:

- Quantitative information on the cardinality of the formal concept's extent (i.e., count and percentage of notes included in the concept),
- The formal concept's intent (i.e., annotation types and induced annotation categories), and
- The concept's extent (i.e., the notes grouped in the concept).

This tool thus directly supports the assessment of annotation activities in @note by applying the considerations outlined in Section 2.

4. Case study: the annotation of "the library of babel"

This section describes how the FCA-based assessment of annotation activities of @note is applied in practice. For this purpose, we focus on the annotation of *The Library of Babel*, a short story written by the Argentinian writer Jorge Luis Borges (Borges, 1944).

4.1. The annotation activity

To provide the annotation ontology for this activity, domain experts (teachers, in this case) applied *close reading* as the basic methodology for the critical literary text analysis (Lentricchia & Dubois, 2003). *Close reading* achieves text interpretation through its primary formal and thematic aspects. *Close reading* promotes several readings of the text: *simple reading* (the reader is able to follow the course of the story), *detailed reading* (the reader achieves an in-depth understanding of the text by consulting, e.g., external

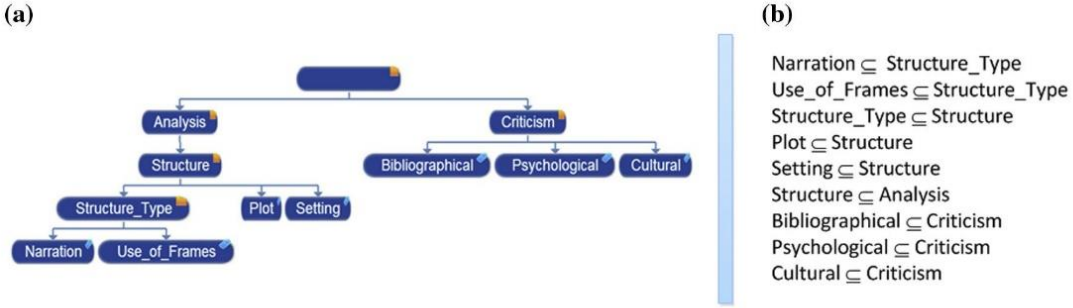


Fig. 5. (a) Sample annotation ontology in @note and (b) description logic counterpart of (a).

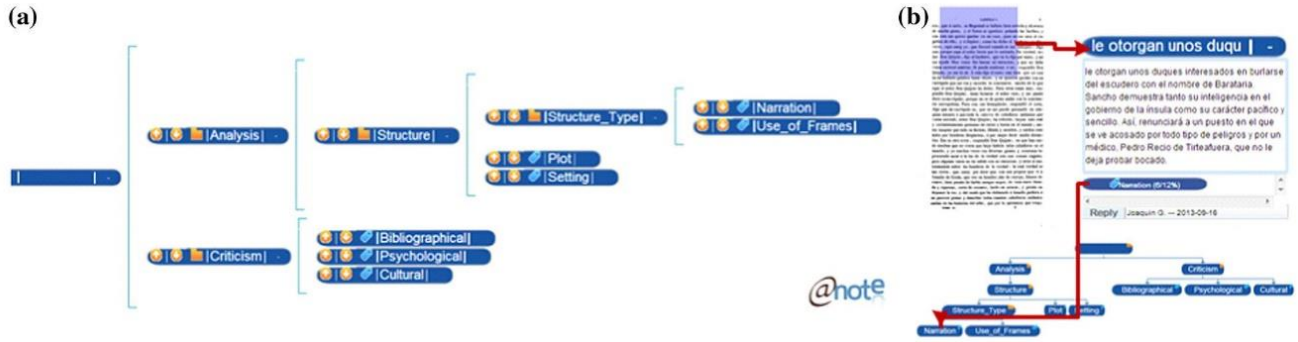


Fig. 6. (a) The ontology editor in @note and (b) the annotation of free-text notes in @note.

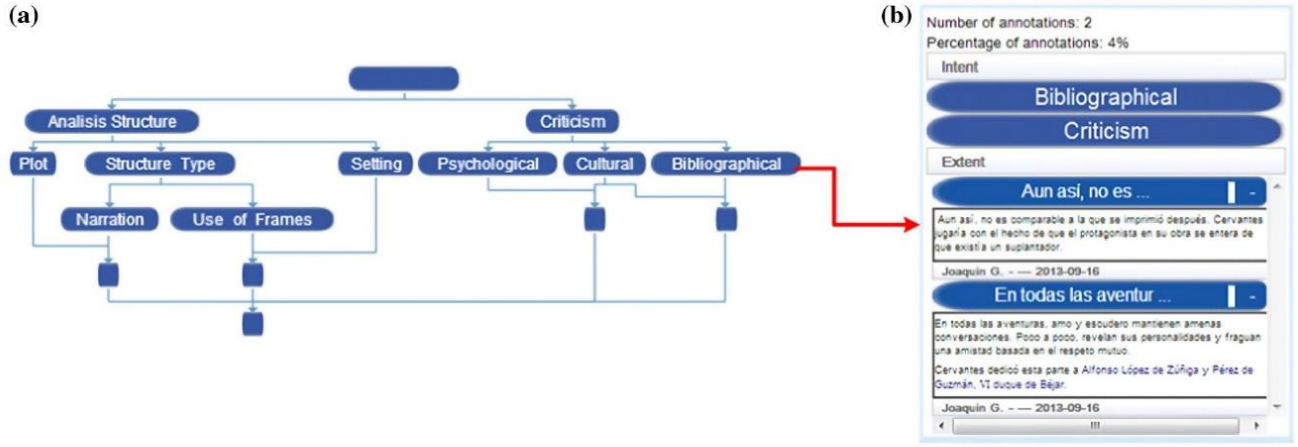


Fig. 7. (a) A Hasse diagram depicted in @note and (b) the expansion of a node from (a).

references, to clarify some text content or to provide additional information), and *interpretative reading* (the reader is able to interpret the contents of the story and to derive ontological and philosophical consequences, e.g., in the case of Borges' story, posing philosophical conjectures about human existence).

Domain experts provided the ontology shown in Fig. 8(a) by applying *close reading*. For this purpose, they first analyzed how *close reading* could be particularized in this text, which led them to formulate a repertory of concepts oriented to the different dimensions of critical text analysis theory. Then they organized these concepts in a meaningfully conceptual hierarchy. This task was facilitated by the collaborative ontology edition mechanisms included in @note (basically, collective edition of the ontology, and use of discussion forums to resolve disagreements on how to materialize *close reading* for the Borges' work).

Concerning the anatomy of the ontology, formal aspects of the text are captured by annotation types in the *References* annotation category. Annotation types include *Authors* referred to in the text, *Citations* made in the text, *Books* referenced, and *Word Meanings* for significant words. Thematic aspects can be catalogued with annotation types under *Characters*, *Time*, *Space* and *Morals*, as well as with the *Authorities* annotation type. Thus, the goal of the domain experts developing this ontology is to allow the annotators (students, in this case) to see how, according to Borges' literary imagination, the *Library* (*Space*), where all the *Books* in the world (and therefore all the *Authors* and all the cultures, represented by *Citations* and *Word Meanings*, anywhere in *Time*) form the *Universe*. Borges' *Universe* includes both a microcosm, the human being, and a macrocosm, divinity. Thus, the order resulting from the union of *Time* and *Space* induces *Morals* (represented by the concepts of *God* and *Devil*). This type of order is perceived only from the *Narrator's* point of view, since the *Other* characters in the story depend upon the *Narrator's* own point of view.

During this annotation activity, annotators (students in this case) created 75 free-text notes, and they catalogued the notes using the ontology provided by the domain experts (their teachers, in this case). Fig. 8(b) shows an example of a note from the ontology catalogued by the students with the annotation type *Citations*.

4.2. Assessment of the activity

When the annotation activity was complete, annotation experts assessed it by using FCA. Fig. 9 shows the concept lattice obtained by applying FCA to the annotation activity depicted in @note. The figure shows that the situations described in Section 2.3 are also present in this scenario. The following examples illustrate the different scenarios:

- The annotation types *Authors*, *Citations* and *Word Meanings*, which are all in the *References* annotation category, were used in isolation (they were used in 3%, 7% and 25% of the notes, respectively) (Fig. 10). This is evidence that the ontology was used at the *detailed reading* level of the *close reading* method, i.e., the annotators found unknown authors, citations and words in the text, then they consulted external sources on the Internet (e.g., web pages), then they created notes explaining these elements, and they classified these notes in one of these concepts (depending on whether the unknown term was an author, a citation, or another word).
- The combination of concepts is evidence that the ontology was used at the *interpretation level* of the *close reading* method. The most prominent example of this is the combination of *Library*

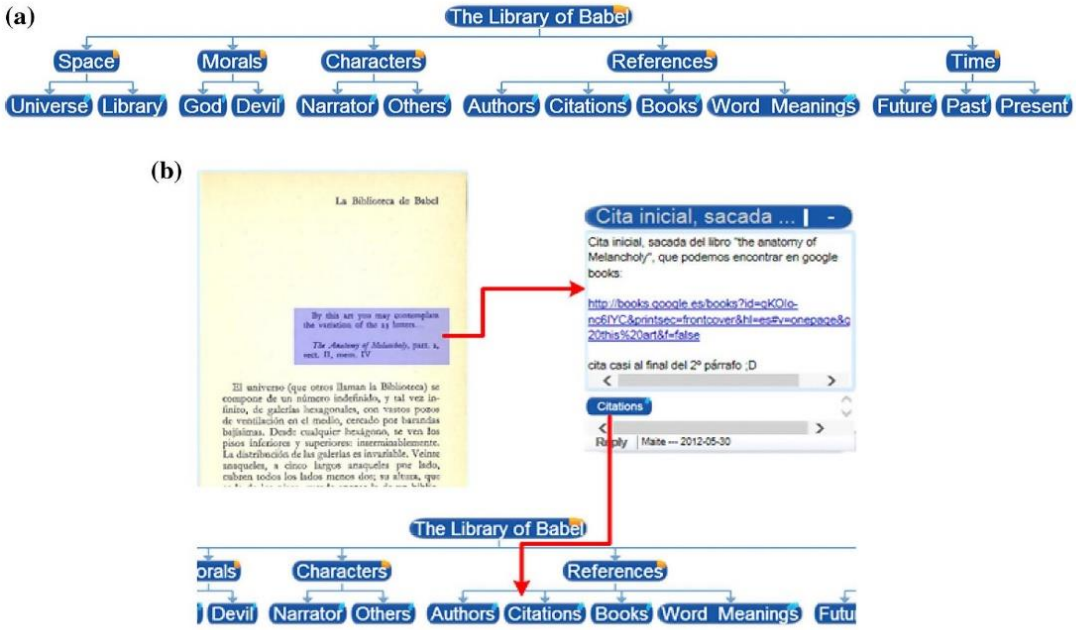


Fig. 8. (a) The annotation ontology for "The Library of Babel" and (b) an example of an annotation.

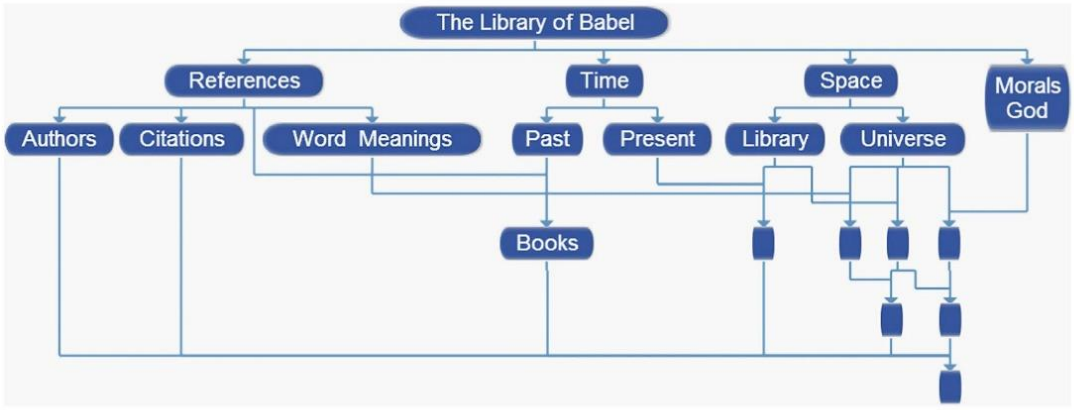


Fig. 9. The concept lattice for the annotation of the *Library of Babel*.

and *Universe* in 5% of the notes (Fig. 11(a)). It is a straightforward ontological interpretation of Borges' text, where the library and the universe share the same time/space/morals paradigm. The combination of *Word Meanings*, *Library* and *Universe* in 4% of the notes reveals a deeper interpretation, in which the term *infinite* is explained, not in terms of its objective (dictionary-based) definition, as would be done at the *detailed reading* level, but in terms of Borges' association of *Library* and *Universe* (Borges' *Library* is infinite, and so is the *Universe*, thus *Library* and *Universe* can be identified) (Fig. 11(b)). The combination of *Word Meanings* and *Universe* in 1% of the notes follows a simpler interpretation of the association, in which *Library* has been omitted because of the prior *Library-Universe* identification (Fig. 11(c)). After examining the notes' contents, the combination of *Books* with *Past* in 1% of the notes revealed (Fig. 11(d)) a philosophical interpretation. Indeed, the pilgrimage of the narrator during his youth, i.e., the *Past*, in search of the *Book*, was interpreted as a metaphor for the meaning of life, wherein life is circular, cyclical and infinite, as are the universe and the library in Borges' literary imagination.

- Other combinations of concepts resulted in incorrect interpretations of the text, however. For example, 4% of the notes were tagged with *Library* and *Present* (Fig. 12(a)). After examining the notes, experts realized that these notes were from the same annotator, who incorrectly related the infiniteness of the Borges' Library with the present, interpretations that could hardly be derived from Borges' text (rather, the library is infinite, and therefore timeless).
- The experts also examined the unused concepts in the activity (those downgraded to the bottom concept), i.e., *Future*, *Narrator*, *Other*, *Devil* and *Authorities*. With respect to *Future*, its lack of use is explained in this particular text because time in the Borges story oscillates between past and present, although the concept could be useful for other works. Regarding the lack of notes involving aspects of the characters, experts realized that to focus this category in terms of particular characters might be unproductive because it is too generic a term (in particular because Borges' text is narrated in the first person, characters other than the narrator are subjected to the narrator's point of view). Similarly, experts recognized the

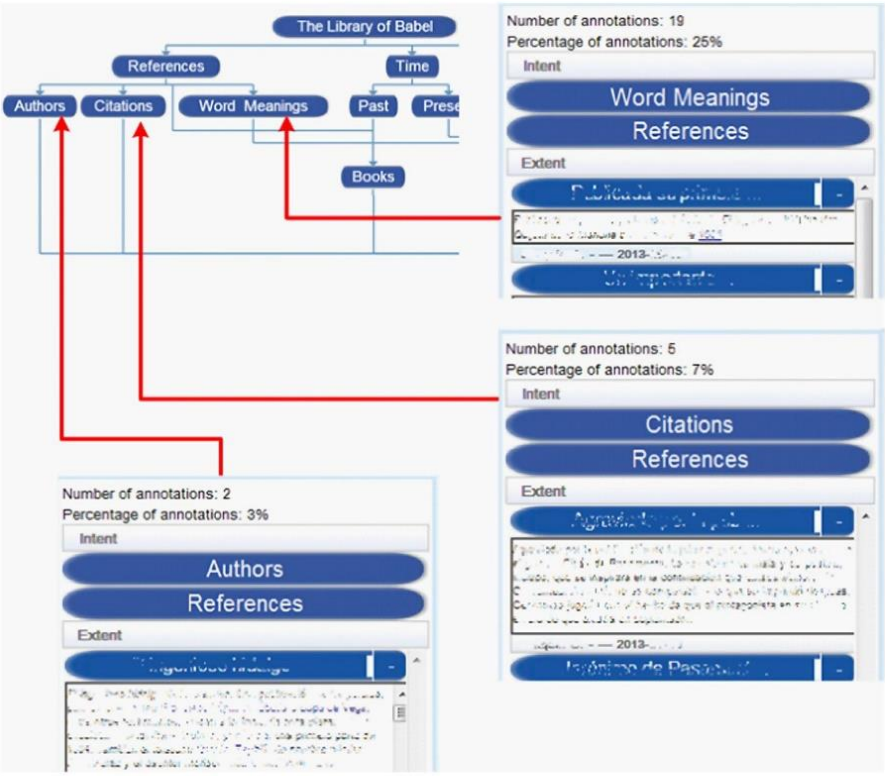


Fig. 10. The formal concepts involving the *Authors*, *Citations* and *Word Meanings* ontology concepts.

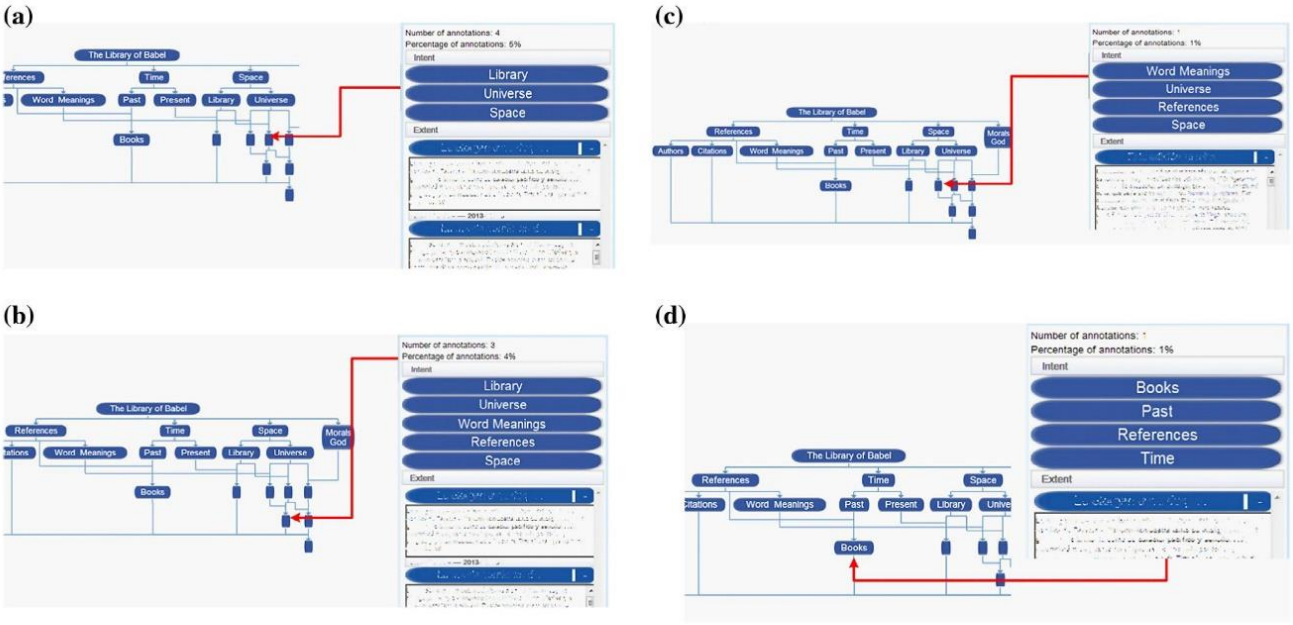


Fig. 11. (a) The combination of *Library* and *Universe*, (b) the combination of *Word Meanings*, *Library* and *Universe*, (c) the combination of *Word Meanings* and *Universe*, and (d) the combination of *Books* and *Past*.

sub-categorization of *Morals* into *God* and *Devil* could lead to a misconception. Indeed, the goal was not to recognize elements concerning deity; although in the text, it is possible to make some interpretations concerning *God* (5% of the notes did, as shown in Fig. 12(b)), it has nothing to do with *Morals*. With respect to *Devil*, the text does not contain any mention of the *Devil*, only *Evil*. The experts also realized that the *Authorities* annotation type is meaningless for this type of work because it can be superseded by using other annotation types in *References*.

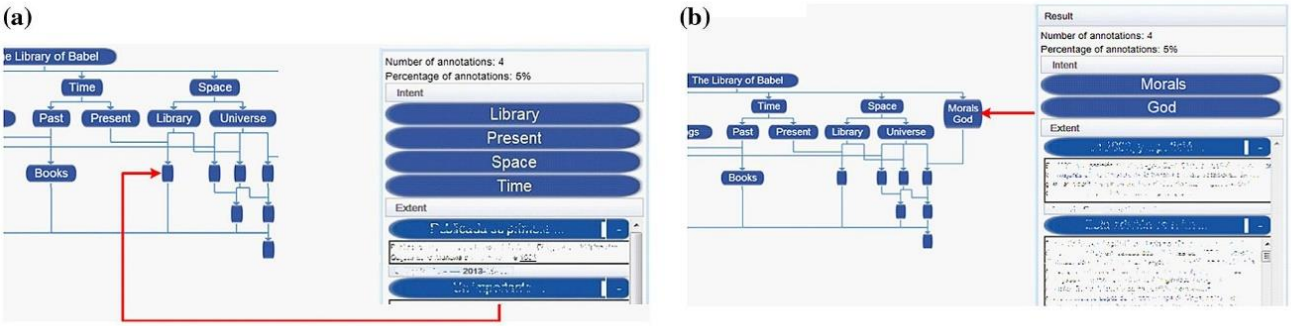


Fig. 12. (a) The combination of *Library* and *Present* and (b) the formal concept for *God*.

4.3. Corrective actions

After completing the assessment, the domain experts performed the following corrective actions based on the results of the FCA applied to the annotation activity:

- The domain experts warned annotators about the potential misconceptions. In particular, authors of the notes in Fig. 11(c) were notified of the possibility of including *Library*, in addition to *Universe* and *Word Meanings*, when discussing the role of *infinite* in the text. They also warned the author of the notes associated with Fig. 12(a) about the incorrect combination of *Library* and *Present*.
- They also refined the annotation ontology. Rather than providing concepts for the different types of characters (*Narrator* and *Other*), they decided to focus the ontology on the analysis of the first-person text, therefore providing concepts for tagging the narrator's different features, i.e., *Physical Description*, *Social Ambit* and *Moral Values*. They decided to eliminate *Authorities*. They chose more abstract names for the *Morals* concepts, i.e., *Good* and *Bad* instead of the more anthropomorphic ones (*God* and *Devil*). The resulting ontology, created in @note, is shown in Fig. 13.

5. Method evaluation

To evaluate the usefulness of the method, we executed an informal survey among domain experts who were using this method in @note. In the survey, we focused on the following three essential aspects of the method:

- *Does the method allow domain experts to assess whether annotators are using the ontology appropriately in the context of the annotation activity*? The response to this question was mostly positive. Experts highlighted the importance of the explicit visualization of the combinations between concepts as one of the key aspects in assessing the use of the ontology during annotation. They also mentioned the possibility of projecting concept lattices onto specific subgroups of annotators and even onto individuals, offered by @note as an essential and very useful feature (in particular in educational settings, where @note is currently used). As a potential feature to be added to the tool, they suggested the inclusion of semi-automatic assessment support (e.g., by adding rules able to detect good and bad uses in the ontology) and the inclusion of dynamic/continuous assessment, so that intermediate assessment results can impact the annotation activity. They also emphasized the specific nature of the @note domain (annotation of literary texts) and the specific purpose of the activities in @note (empowering the reading of literary works through explicit annotation activities), warning us that what is valid for this domain may not necessarily be extrapolated to other annotation domains. Nevertheless, they also agreed that, given the complexity of this domain and of the intended activity (human reading), the possibility of extrapolation is more than a reasonable assumption.
- *Does the method allow domain experts to help annotators make better use of annotation ontologies*? While the experts agreed that the method provides valuable information to help annotators use ontologies more effectively (as the case study in the previous section illustrated), they also agreed that it should be accompanied by better support for offering feedback to the annotators (e.g., informative messages linked to ontology concepts explaining their intended use in the annotation activity). They also noted the need to personalize feedback based on the annotators' expected expertise (this feature is very important in educational settings). The inclusion of mechanisms to allow/forbid certain combinations (e.g., using rules) would also be a welcome improvement.
- *Does the method allow domain experts to enhance their ontologies after the annotation activity*? The response to this question was unanimously affirmative. After applying the method, domain experts refined their annotation ontologies in similar ways to those described in the case study (erasing useless concepts, rethinking concept sub-hierarchies, choosing better names for concepts, etc.). The case study is a clear example of this positive outcome.

6. Related work

Most of the research on the use of artificial intelligence and knowledge-based techniques for enhancing the annotation activities of online resources has been focused on the automation of the semantic annotation process (i.e., the automatic addition of semantic annotations to digital resources), in the context of the

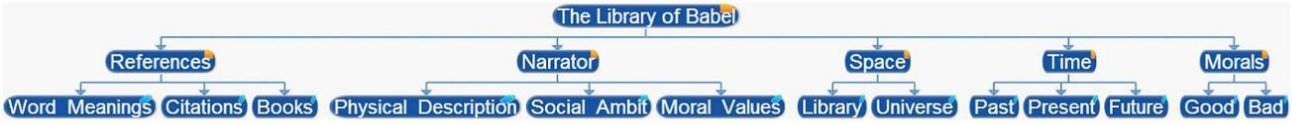


Fig. 13. The evolution of the annotation ontology as a result of the assessment of the annotation activity.

semantic web, and in the annotation of text resources (e.g., text documents and HTML pages) (Oliveira & Rocha, 2013; Reeve & Han, 2005). Typically, these systems use natural language processing techniques to identify the parts of the documents to be annotated and, ideally, to determine the content of the annotations. To acquire and exploit the additional linguistic knowledge required for the annotation process, these systems can adopt different strategies, as described below:

- *Data-driven* strategies. In these strategies, the annotation process is orchestrated by using solely the corpus of pre-annotated texts. An example of a system following this strategy is SemTag, the semantic annotation component of the Seeker platform (Dill, 2003; Dill et al., 2003), which tags parts of the documents with concepts taken from a taxonomy. The only additional knowledge required by SemTag is the annotated corpus to determine the corpus-wide distribution of terms at each node of the taxonomy.
- *Knowledge base* strategies. According to these strategies, the user manually provides knowledge bases that contain the additional knowledge required to perform the annotation. These knowledge bases can conform to formalisms such as regular expression-based patterns or rules. Examples of systems following this strategy are AeroDAML (Kogut & Holmes, 2001), a system used to annotate documents with the DAML agent annotation language (Greaves, 2004), KIM (Malik, Prakash, & Rizvi, 2010; Popov, Kiryakov, Kirilov, & Manov, 2003; Popov, Kiryakov, Ognyanoff, Manov, & Kirilov, 2004), a flexible platform for pattern-based semantic text annotation, MUSE (Maynard, 2003), a system for named entity recognition and co-referencing in text documents, Cerno (Kiyavitskaya, Zeni, Cordy, Mich, & Mylopoulos, 2009; Kiyavitskaya, Zeni, Mich, & Cordy, 2007), a system for domain-specific text annotation, KnowWE (Baumeister, Reutelshoefer, Haupt, & Nadrowski, 2008; Baumeister, Reutelshoefer, & Puppe, 2010), a semantic wiki system, Lixto (Baumgartner, Frölich, & Gottlob, 2007; Baumgartner, Flesca, & Gottlob, 2001), which helps users in the creation of rule-based wrappers and which uses a visual style, and Semantic Wikipedia (Krötzsch, Vrandečić, & Völkel, 2006), a framework for extracting semantic information from plain text.
- *Pattern-discovering* strategies. According to these strategies, the user provides an initial knowledge base made of patterns, which is automatically extended (Brin, 1999). Armadillo (Dingli, Ciravegna, & Wilks, 2003; Kiyavitskaya et al., 2009), a system for the domain-specific annotation of large repositories of texts, or PANKOW (Cimiano, Handschuh, & Staab, 2004), a pattern-based annotation algorithm, are practical examples of these strategies.
- *Wrapper induction* strategies. These systems use machine-learning techniques to induce wrappers to locate the sections in the documents to be annotated (Kushmeric, 2000). Usually (although not necessarily) these wrappers are described as automatically-induced knowledge bases. Examples of systems using wrapper induction are MnM (Vargas-Vera, Moreale, Stutt, Motta, & Ciravegnan, 2007; Vargas-Vera et al., 2002), a semantic annotation environment able to infer tagging rules from an annotated corpus, Ont-O-Mat (Handschuh, Staab, & Ciravegna, 2002), a system implementing the CREAtion of Metadata (CREAM) framework (Handschuh, & Staab, 2002), GoNTogle (Giannopoulos, Bikakis, Dalamagas, & Sellis, 2010), a system that exploits both the text structure and previous annotations made by users, KnowItAll (Etzioni et al. 2005), a system oriented to extracting large collections of facts from the web, and Thresher (Hogue & Karger, 2005; Huynh, Karger, & Quan 2002), a system able to produce wrappers from examples provided by the end user.

Contrary to the approach used in the above systems, the use of artificial intelligence techniques in the method described in this paper is oriented to analyzing the resources annotated, not to automating the annotation process. It is especially well suited to domains where annotations cannot easily be inferred from the contents automatically (e.g., annotation of literary texts according to some literary criteria, such as the case study presented in this paper illustrates). It is also especially well suited to annotation activities comprising non-text data, such as the annotation of images, video and other media (Dasiopoulou, Giannakidou, Litos, Malasioti, & Kompatsiaris, 2011), or the annotation of text resources using free-text notes with a clearly defined scholarly purpose (e.g., critical text analysis), as enabled by tools similar to @note (Azouaou & Desmoulins, 2006; Donato et al., 2013; Koivunen, 2005; Rocha et al., 2009; Schroeter et al., 2006; Tazi et al., 2003). Additionally, because this method is focused on the assessment stage of the annotation process, it could be meaningfully combined with any of the previously mentioned massive annotation methods.

FCA has been used as a primary tool in ontology engineering for different purposes (Cimiano, Hotho, Stumme, & Tane, 2004; Poelmans, Ignatov, Kuznetsov, & Dedene, 2013) as follows:

- *Ontology construction*. This activity addresses the construction of ontologies for a given domain. For this purpose, FCA supports bottom-up approaches to ontology construction, in which the process focuses on building formal contexts with concept instances as objects and features of concept instances as attributes. Xu and Xiao (2009) exemplify this approach in the computer network management domain; Bao, Zhou, and He (2005) exemplify it in the domain of pressure component design; and Chi, Hsu, and Yang (2005) in the domain of digital archives. Richards (2004, 2006) have developed a method to combine rule bases and FCA in the construction of ontologies. In their method, classification rules characterizing objects are used, which makes it possible to identify rules as objects and rule conditions as attributes in the formal concepts. Another typical application of FCA in ontology construction is to build initial ontologies from a set of documents. In this case, documents are preprocessed using standard natural language processing techniques, and then ontologies are built from the result of this preprocessing (Bendaoud, Toussaint, & Napoli, 2008; Cimiano, Hotho, & Staab, 2005; Cole, & Eklund, 1996; Gamallo, Lopes, & Agustini, 2007; Jiang, Ogasawara, Endoh, & Sakuari, 2003; Soon & Kuhn, 2004; Xu, Li, Wu, Li, & Yuan, 2006). Kiu and Lee (2008) use FCA to edit existing ontologies (i.e., to add, delete and modify existing concepts) instead of constructing ontologies from scratch.
- *Ontology enhancement and quality management*. This activity addresses ontology refinement to better suit the target domain. Rudolph (2004) uses FCA to add axioms, in the form of implication rules, incrementally and interactively to a description logic-based representation of an ontology. He focuses on a finite universe of objects and on pairs of these objects. Then, he uses ontology concepts as attributes for these objects and ontology roles as attributes for the pairs. By using the associated concept lattices to approximate hypothetical axioms and by asking domain experts about the validity of these axioms when they are not covered by the current ontology, the method can enrich the ontology with new axioms or otherwise enrich the formal contexts with appropriate counterexamples. Rudolph, Volker, and Hitzler (2007) and Völker and Rudolph (2008) extended the technique and combined it with natural language processing to cope with the refinement of lexical ontologies, and Rudolph (2008) extended it to acquire complete sets of domain-range restrictions. Sertkaya (2009) used a similar

approach to complete ontologies with relevant information about a domain. Kim, Hwang, and Kim (2007) considered sets of ontology constructs as objects and sets of binary relations as attributes. They then mapped ontologies onto formal contexts and applied FCA to detect potential problems in the ontologies. Jiang and Chute (2009) and Jiang, Pathak, and Chute (2009) used FCA to audit the quality of two real-world ontologies.

- *Ontology mapping and merging.* Ontology mapping is the transformation of source ontologies into target ones, i.e., with knowledge representations of overlapping fields likely to represent the same concept with different names, while ontology merging is the amalgamation of several ontologies into a single one. These two interrelated activities have also been addressed by using FCA. Stumme and Maedche (2001) proposed a merging method focused on concept instances. These instances were used to construct formal contexts, which, in turn, were merged. The resulting concept lattice was pruned using information from the original ontologies, and, finally, the merged ontology was generated from the pruned lattice with the help of the domain experts. De souza, Davis, and Evangelista (2006) solved the interoperability issues of overlapping ontologies by extracting similarity measures for the identification of concepts related across ontologies. They used thesauri as a bridge representation, i.e., they associated terms in thesauri with concepts in ontologies, and then mapped the thesauri in concept lattices. Similarity distances were defined in terms of the resulting lattices. Fan and Xiao (2007) approach focused on similarity measures between ontologies in terms of subclass mapping, rather than in terms of entity. To do so, it computed inclusion measures to map the ontologies. Other works have used ontologies to uncover similarities between FCA concepts, as in Formica (2006). Zhao, Halang, and Wang (2007) proposed transforming ontologies into formal contexts and then merging them to obtain a concept lattice, while concurrently developing a similarity measure based on a rough FCA. Finally, Krotzsch, Hitzler, and Zhang (2005) proposed modeling complex relationships by using morphisms to formalize the interplay between two knowledge bases.

Thus, while most of the works dealing with the use of FCA in ontological engineering are focused on ontology management, assuming that FCA can facilitate ontology management operations such as merging, mapping, assessment, and quality assurance, our approach is more focused on the annotation activities themselves. Domain experts use concept lattices induced by annotated resources and the structural organization of ontological concepts to assess particular annotation activities. In consequence, they can either instruct annotators on the better use of ontologies, enhance annotation ontologies (as many of the aforementioned works on the use of FCA in ontological engineering do), or adopt a mixture of both types of corrective actions.

7. Conclusions and future work

The semantic annotation of collections of digital resources enhances the cataloguing and retrieval of the resources and, more importantly, enables a more sophisticated use of these resources in different applications. Semantic annotation can require both standardized annotation schemas and domain-specific ontologies specifically designed by domain experts to suit the features of the collection and the intended use of the resources therein. However, in order to ensure the quality of the annotations it is necessary to assess to what extent annotators made full use of the ontologies, and to what extent the ontologies provided were suitable to the annotation task envisioned. By doing so, domain experts can, on one hand, advise annotators on how to improve their ontology usage. On the other hand, domain experts can detect aspects from the ontology that can be improved in order to better meet annotation requirements. In consequence, they are able to re-structure their ontologies, which leads to an incremental and iterative process of ontology enhancement. This paper has shown how to achieve these features by using FCA.

From a theoretical point of view, the main contribution of this paper is to developing a generic approach to the assessment of semantic annotation activities, based on FCA. This approach is particularly suited to settings where the annotation of resources cannot easily be automated on the basis of the resource structure, and therefore, must be performed by a community of annotators though a collaborative and iterative process. Since in this approach the responsibility of ontology design is assigned to domain experts, we constrain the ontologies to hierarchical arrangements of concepts, i.e., concepts related by an *is-a* relationship. Other kinds of relationships are intentionally excluded in order to facilitate the authoring of ontologies by using suitable hierarchy editors. In this way, by considering annotated resources as objects and ontology concepts as attributes of a formal context, FCA is used to create a concept lattice from annotated collections of digital resources. The upper part of the lattice contains the ontology concepts from the original ontology used to annotate the resources, as well as the *is-a* relationships among these concepts, re-structured according to evidence of use gathered from the formal context. The lower part contains the different combinations of ontology concepts meaningfully and distinctively used during the annotation. Thus, by inspecting the lattice, experts can gain insight on how annotators actually used the ontology, uncovering those uses caused by a misconception of the annotation guidelines, and those due to potential problems in the original ontology. In consequence, the aforementioned assessment goals (i.e., to instruct annotators in the better use of the annotation ontology, and to enhance the ontology according to its practical usage) can be achieved.

From a practical point of view, the main contribution of the paper is showing how the proposed approach can be implemented in practice. For this purpose, we have described how the approach has been implemented in @note, a tool for the collaborative annotation of digitized literary works. In addition, we have illustrated how this implementation works in practice, with an annotation activity focused on *The Library of Babel*, a short story written by the Argentinian writer Jorge Luis Borges. In this setting, and in order to evaluate the approach, we ran an informal survey among experts in literature who used @note. The outcomes were mostly positive: domain experts considered the approach valuable to helping annotators improve their annotation skills with respect to established annotation ontologies, and a valuable tool for enhancing annotation ontologies themselves. They also suggested some improvements to the approach, concerning support for greater automation of the assessment process by using rules operating on the concept lattice.

In this way, as main strengths of the approach we can highlight its feasibility and ease of use. Indeed, as the experience of @note with experts in literature has made apparent, domain experts (literature teachers) are able to carry out basic conceptualizations of annotation ontologies in terms of *is-a* arrangements of concepts, they are able to give meaningful interpretations to the resulting concept lattices once annotation activities have finished, and, more importantly, they are able to instruct annotators (students, in this setting) regarding their misconceptions in using the ontology, and to enhance the ontology itself as the result of usage experiences.

Finally, we are aware of some weaknesses in the approach. Perhaps the most significant one is the rather strong ontological

assumption made, which confines the ontologies allowed to taxonomical arrangements of atomic concepts. However, as argued earlier, this assumption is necessary in order to maintain the feasibility of the approach concerning the role of domain experts as ontology designers. Whether this assumption can be relaxed without compromising usability may be the object of future inquiries. Another weakness of the approach is whether the visual representation of the lattice scales well for larger ontologies. In this respect, works like that of Katifori, Halatsis, Lepouras, Vassilakis, and Giannopoulou (2007) suggest that it is possible to use sophisticated visualization techniques to cope with huge hierarchical structures. However, whether these techniques will be well received by domain experts (and, in particular, by literature experts in the context of @note) deserves more research efforts. Lastly, another weakness in the approach is the limited support for helping domain experts analyze the concept lattice. As indicated above, our experiences with domain experts suggested some interesting directions (e.g., to use rules for automating some aspects of the assessment). However, these aspects deserve further investigation.

Currently we are working on the human–computer interaction aspects of the approach in the context of @note on the basis of domain expert feedback. On the basis of this feedback, we are also adding support for automating some assessment aspects by enabling the definition and use of assessment rules able to detect common situations that demand domain expert attention. For future work, we will address other aspects raised by domain experts, i.e., dynamic/continuous assessment, support for attaching feedback to the annotation ontology, etc. In addition, we are also planning to apply the approach to other settings (repositories of learning objects in the educational domain and a collection of digitized and digital objects in the Digital Humanities scenario). Finally, we plan to work on the weaknesses of the approach mentioned above, improving visualization support for concept lattices, and more in-depth research oriented to relaxing the basic ontological assumption adopted.

Acknowledgements

This work was funded by Google (Digital Humanities Award Programs 2010, 2011), as well as by the project grants FFI2012-34666, TIN2010-21288-C02-01, TIN2009-14317-C03-03 and S2009/TIC-1650. We would also like to thank César Ruiz for his work in the development of @note, as well as the members of ILSA and LEETHI research groups for their effort in conceiving and evaluating the application.

References

Aroyo, L., & Dicheva, D. (2004). The new challenges for E-learning: the educational semantic web. *Educational Technology & Society, 7*(4), 59–69.

Azouaou, F., & Desmoulins, C. (2006). MemoNote, a context-aware annotation tool for teachers. In *Proceedings of the 7th international conference on information technology based higher education and training (ITHET '06)* (pp. 621–628). Sydney, Australia: IEEE Computer Society.

Baader, F., Calvanese, D., McGuinness, D. L., Nardi, D., & Paterl-Schneider, P. F. (Eds.). *The description logic handbook: theory, implementation and applications* (2nd ed.. Cambridge University Press.

Bao, S., Zhou, Y., & He, S. (2005). Research on pressure component design ontology building based on knowledge sharing and reusing. In *Proceedings of the 9th international conference on computer supported cooperative work in design (CSCWD'05)* (pp. 1183–1187). Coventry, UK: IEEE Computer Society.

Baumeister, J., Reutelshoefer, R., Haupt, F., & Nadrowski, K. (2008). Capture and refactoring in knowledge wikis coping with the knowledge soup. In *Proceedings of 2nd workshop on scientific communities of practice (SCOOP'08)*. Bremen, Germany.

Baumeister, J., Reutelshoefer, J., & Puppe, F. (2010). KnowWE: a semantic wiki for knowledge engineering. *Applied Intelligence, 35*(3), 323–344.

Baumgartner, R., Flesca, R., & Gottlob, G. (2001). Visual web information extraction with Lixto. In *Proceedings of the 27th international conference on very large data bases (VLDB '01)* (pp. 119–128). San Francisco, USA: Morgan Kaufmann Publishers Inc.

Baumgartner, R., Frölich, O., & Gottlob, G. (2007). The lixto systems applications in business intelligence and semantic web. In E. Franconi et al. (Eds.), *The semantic web: research and applications. LNCS* (4519, pp. 16–26).

Bendaoud, R., Toussaint, Y., & Napoli, A. (2008). PACTOLE: a methodology and a system for semi-automatically enriching an ontology from a collection of texts. In P. Eklund et al. (Eds.). *ICCS LNAI* (5113, pp. 203–216). Springer.

Berry, D. M. (Ed.). (2012). *Understanding digital humanities*. Palgrave Macmillan.

Borges, J. L. (1944). *La biblioteca de Babel, in Ficciones* (Last Ed. 2008). Madrid: Alianza Editorial.

Brewster, C., & O'hara, K. (2004). Knowledge representation with ontologies: the present and future. *IEEE Intelligent Systems, 9*(1), 72–81.

Brin, S. (1999). Extracting patterns and relations from the world wide web. In *Selected papers from the international workshop on the World Wide Web and databases (WebDB '98)* (pp. 172–183). London, UK: Springer.

Calhoun, K. (2013). *Digital libraries*. Facet Publishing.

Carpineto, C., & Romano, G. (2004). *Concept data analysis: theory and applications*. John Wiley & Sons.

Chi, Y. L., Hsu, T. Y., & Yang, W. P. (2005). Building ontological knowledge bases for sharing knowledge in digital archive. In *Proceeding of the 4th IEEE international conference on machine learning and cybernetics* (4, pp. 2261–2266). Guangzhou, China: IEEE Computer Society.

Cigarrán, J., Gonzalo, J., Peñas, A., & Verdejo, F. (2004). Browsing search results via formal concept analysis: automatic selection of attributes. In P. Eklund (Ed.), *ICFCA 2004, LNAI* (pp. 74–87). Sydney, Australia: Springer.

Cigarrán, J., Peñas, A., Gonzalo, J., & Verdejo, F. (2005). Automatic selection of noun phrases as document descriptors in an fca-based information retrieval system. In B. Ganter & R. Godin (Eds.). *ICFCA 2005, LNAI* (pp. 49–63). Lens, France: Springer.

Cimiano, P., Handschuh, S., & Staab, S. (2004). Towards the self-annotating web. In *Proceedings of the 13th conference on World Wide Web (WWW '04)* (pp. 462–471). New York, USA: ACM.

Cimiano, P., Hotho, A., & Staab, S. (2005). Learning concept hierarchies from text corpora using formal concept analysis. *Journal of Artificial Intelligence, 24*, 305–339.

Cimiano, P., Hotho, A., Stumme, G., & Tane, J. (2004). Conceptual knowledge processing with formal concept analysis and ontologies. In P. Eklund (Ed.). *ICFCA, LNAI* (2961, pp. 189–207). Springer.

Cole, R., & Eklund, P. (1996) Application of formal concept analysis to information retrieval using a hierarchically structured thesaurus. In International conference on conceptual graphs, ICCS '96 (pp. 1–12). Sydney: University of New South Wales.

Dasiopoulou, S., Giannakidou, E., Litos, G., Malasioti, P., & Kompatsiaris, Y. (2011). A survey of semantic image and video annotation tools. In G. Paliouras et al. (Eds.). *Knowledge-driven multimedia information extraction and ontology evolution* (6050, pp. 196–239). Springer.

De souza, K. X. S., Davis, J., & Evangelista, S. R. M. (2006). Aligning ontologies evaluating concept similarities and visualizing results. *Journal on Data Semantics V, LNCS, 3870*, 211–236. Springer.

Devedzic, V., Jovanovic, J., & Gasevic, D. (2007). The pragmatics of current E-Learning standards. *IEEE Internet Computing, 11*(3), 19–27.

Dill, S. (2003). A case for automated large-scale semantic annotation. *Web Semantics: Science, Services and Agents, 1*(1), 115–132.

Dill, S., Eiron, N., Gibson, D., Gruhl, D., Guha, R., Jhingran, A., et al. (2003). SemTag and seeker: bootstrapping the semantic web via automated semantic annotation. In *Proceedings of the 12th international conference on world wide web (WWW '03)* (pp. 178–186). New York, USA: ACM.

Dingli, A., Ciravegna, F., & Wilks, Y. (2003). Automatic semantic annotation using unsupervised information extraction and integration. In *Proceedings of workshop on knowledge markup and semantic annotation (K-CAP'03)*. Florida, USA: ACM.

Donato, F., Morbidoni, C., Fonda, S., Piccioli, A., Grassi, A., & Nucci, M., (2013). Semantic annotation with Pundit: a case study and a practical demonstration. In *Proceedings of the 1st international workshop on collaborative annotations in shared environment: metadata, vocabularies and techniques in the digital humanities* Article no. 16. ACM Digital Library.

Etzioni, O., Cafarella, M., Downey, D., Popescu, A., Shaked, T., Soderland, S., et al. (2005). Unsupervised named-entity extraction from the web: an experimental study. *Artificial Intelligence, 165*(1), 91–134.

Fan, L., & Xiao, T. (2007). An automatic method for ontology mapping. In B. Apolloni et al. (Eds.). *KES/WIRN, Part III, LNAI* (4694, pp. 661–669). Springer.

Formica, A. (2006). Ontology-based concept similarity in formal concept analysis. *Information Sciences, 176*, 2624–2641.

Gamallo, P., Lopes, G. P., & Agustini, A. (2007). Inducing classes of terms from text. In V. Matousek et al. (Eds.). *TSD, LNAI* (4629, pp. 31–38). Springer.

Ganter, B., & Wille, R. (1999). *Formal concept analysis – mathematical foundations*. Springer Verlag.

Gayoso, J., Ruiz, C., Pablo, L., Sarasa, A., Goicoechea, M., Sanz, A., & Sierra, J. L., (2012). A flexible model for the collaborative annotation of digitized literary works. In *Proceedings of the 2012 digital humanities conference (DH101)*. Hamburgo, Germany.

Gayoso, J., Sanz, A., & Sierra, J. L., (2013). @note: An electronic tool for academic readings. In *Proceedings of the workshop on collaborative annotations in shared environments: metadata, vocabularies and techniques in the digital humanities (DH-CASE'13)* at ACM DocEng 2013. Florence, Italy.

Giannopoulos, G., Bikakis, N., Dalamagas, T., & Sellis, T. (2010). *GoNTogle: a tool for semantic annotation and search. Proceedings of the 7th international conference on The semantic web: research and applications – volume part II (ESWC'10)*. Berlin,Germany: Springer, pp. 376–380.

Greaves, M. (2004). DAML – DARPA Agent Markup Language http://www.daml.org/.
Handschuh, S., & Staab, S. (2002). Authoring and annotation of web pages in CREAM. In *Proceedings of the 11th international conference on World Wide Web (WWW '02)* (pp. 462–473). New York, USA: ACM.
Handschuh, S., Staab, S., & Ciravegna, F. (2002). S-CREAM - Semi-automatic CREAtion of Metadata. In *Proceedings of the 13th international conference on knowledge engineering and knowledge management (EKAW'02)* (pp. 358–372). Siguenza, Spain: Springer.
Hogue, A., & Karger, D. (2005). Thresher: automating the unwrapping of semantic content from the World Wide Web. In *Proceedings of the 14th international conference on World Wide Web (WWW '05)* (pp. 86–95). New York, USA: ACM.
Hunter, J., & Gerber, A. (2010). Harvesting community annotations on 3D models of museum artefacts to enhance knowledge, discovery and re-use. *Journal of Cultural Heritage, 11*(1), 81–90.
Huynh, D., Karger, D., & Quan, D. (2002). Haystack: A platform for creating, organizing and visualizing information using RDF. In *Proceedings of the 8th international conference on intelligent user interfaces (IUI '03)* (pp. 323). New York, USA: ACM.
Jiang, G., & Chute, C. G. (2009). Auditing the semantic completeness of SNOMED CT using formal concept analysis. *Journal of the American Medical Informatics Association, 16*(1), 89–102.
Jiang, G., Ogasawara, K., Endoh, A., & Sakuari, T. (2003). Context-based ontology building support in clinical domains using formal concept analysis. *International Journal of Medical Informatics, 71*, 71–81.
Jiang, G., Pathak, J., & Chute, C. G. (2009). Formalizing ICD coding rules using formal concept analysis. *Journal of Biomedical Informatics, 42*(3), 504–517.
Katifori, A., Halatsis, C., Lepouras, G., Vassilakis, C., & Giannopoulou, E. (2007). Ontology visualization methods – a survey. *ACM Computing Surveys, 39*(4). Article 10.
Keyser, P. (2012). *Indexing: from thesauri to the semantic web*. Chandos Publishing.
Kim, D. S., Hwang, S. H., & Kim, H. G. (2007). Concept analysis of OWL ontology based on the context family model. In *Proceedings of the 2007 international conference on convergence information technology (ICCIT '07)* (pp. 896–901). Washington, DC, USA: IEEE Computer Society.
Kiu, C. C., & Lee, C. S. (2008). Ontological knowledge management through hybrid unsupervised clustering techniques. In Y. Zhang et al. (Eds.). *APWeb, LNCS* (4976, pp. 499–510). Springer.
Kiyavitskaya, N., Zeni, N., Cordy, j. R., Mich, L., & Mylopoulos, J. (2009). Cerno: light-weight tool support for semantic annotation of textual documents. *Data & Knowledge Engineering, 68*(12), 1470–1492.
Kiyavitskaya, N., Zeni, N., Mich, L., & Cordy, J. (2007). Annotating accommodation advertisements using Cerno. In M. Sigala et al. (Eds.). *Information and communication technologies in tourism* (2007, pp. 389–400). Springer.
Kogut, P., & Holmes, W. (2001). AeroDAML: applying information extraction to generate DAML annotations from web pages. In *Proceedings of the workshop on knowledge markup and semantic annotation at 1st international conference on knowledge capture (K-CAP 2001)* (200). Victoria, Canada: Springer.
Koivunen, M. R. (2005). Annotea and semantic web supported collaboration. In *Proceedings of the workshop on user aspects of the semantic web at European semantic web conference (ESWC'05)* (pp. 5–16). Heraklion, Greece: Springer.
Krötzsch, M., Vrandečić, D., & Völkel, M. (2006). Semantic mediawiki. In I. Cruz et al. (Eds.), *The Semantic Web-ISWC* (pp. 935–942). Springer.
Krötzsch, M., Hitzler, P., & Zhang, G. Q. (2005). Morphisms in context. In F. Dau et al. (Eds.). *ICCS, LNAI* (3596, pp. 223–237). Springer.
Kurilovas, E., Kubilinskiene, S., & Dagiene, V. (2014). Web 3.0 – based personalization of learning objects in virtual learning environments. *Computers in Human Behaviour, 30*, 654–662.
Kushmeric, N. (2000). Wrapper induction: efficiency and expressiveness. *Artificial Intelligent, 118*(1–2), 15–68.
Labra, J. E., Ordónez, P., & Cueva-Lovelle, J. M. (2010). WESONet: applying semantic web technologies and collaborative tagging to multimedia web information systems. *Computers in Human Behaviour, 26*(2), 205–209.
Lentricchia, F., & Dubois, A. (2003). *Close reading: the reader*. Durham, N. C.: Duke University Press.
Malik, S. K., Prakash, N., & Rizvi, S. (2010). Semantic annotation framework for intelligent information retrieval using KIM architecture. *International Journal of Web & Semantic Technology, 1*(4), 12–26.
Maynard, D. (2003). Multi-source and multilingual information extraction. *Expert Update, 6*(3), 11–16.
Mu, X. (2010). Towards effective video annotation: an approach to automatically link notes with video content. *Computers & Education, 55*(4), 1752–1763.
Noy, N. F., Sintek, M., Decker, S., Crubézy, M., Fergerson, R. W., & Musen, M. (2001). Creating semantic web contents with Protégé-2000. *IEEE Intelligent Systems, 16*(2), 60–71.
Oliveira, P., & Rocha, J. (2013). Semantic annotation tools survey. In *Proceedings of the IEEE symposium on computational intelligence and data mining (CIDM'13)* (pp. 301–307). Singapore: IEEE Computer Society.
Poelmans, J., Ignatov, D. I., Kuznetsov, S. O., & Dedene, G. (2013). Formal concept analysis in knowledge processing: a survey. *Expert Systems with Applications, 40*, 6538–6560.
Popov, B., Kiryakov, A., Kirilov, A., & Manov, D. (2003). KIM – semantic annotation platform. In D. Fensel et al. (Eds.). *ISWC, LNCS* (Vol. 2870, pp. 834–849). Springer.
Popov, B., Kiryakov, A., Ognyanoff, D., Manov, D., & Kirilov, A. (2004). KIM–a semantic platform for information extraction and retrieval. *Natural Language Engineering, 10*(3–4), 375–392.
Reeve, L., & Han, H. (2005). Survey of semantic annotation platforms. In *Proceedings of the 2005 ACM symposium on applied computing (SAC '05)* (pp. 1634–1638). New York, USA: ACM.
Richards, D. (2004). Addressing the ontology acquisition bottleneck through reverse ontological engineering. *Knowledge and Information Systems, 6*, 402–427.
Richards, D. (2006). Ad-hoc and personal ontologies: a prototyping approach to ontology engineering. In A. Hoffmann et al. (Eds.). *PKAW, LNAI* (4303, pp. 13–24). Springer.
Rocha, T., Willrich, R., Fileto, R., & Tazi, S. (2009). Supporting collaborative learning activities with a digital library and annotations. In A. Tatnall & A. Jones (Eds.). *Education and technology for a better world* (Vol. 302, pp. 349–358). Berlin Heidelberg: Springer.
Rudolph, S. (2004). Exploring relational structures via FLE. In K. E. Wolff et al. (Eds.). *ICCS, LNAI* (3127, pp. 196–216). Springer.
Rudolph, S. (2008). Acquiring generalized domain-range restrictions. In R. Medina et al. (Eds.). *ICFCA, LNAI* (4933, pp. 32–45). Springer.
Rudolph, S., Volker, J., & Hitzler, P. (2007). Supporting lexical ontology learning by relational exploration. In U. Priss et al. (Eds.). *ICCS, LNAI* (4604, pp. 488–491). Springer.
Schroeter, R., Hunter, J., Guerin, J., Khan, I., & Henderson, M. (2006). A synchronous multimedia annotation system for secure collaboratories. In *Proceedings of the Second IEEE international conference on e-science and grid computing (E-SCIENCE '06)* (pp. 41). Washington, DC, USA: IEEE Computer Society.
Sertkaya, B. (2009). OntoComP: a protege plugin for completing OWL ontologies. In L. Arroyo et al. (Eds.). *ESWC, LNCS* (5554, pp. 898–902). Springer.
Šimko, J., Tvarožek, M., & Bieliková, M. (2013). Human computation: image metadata acquisition based on a single-player annotation game. *International Journal of Human-Computer Studies, 71*(10), 933–945.
Soon, K., & Kuhn, W. (2004). Formalizing user actions for ontologies. In M. J. Egenhofer et al. (Eds.). *GIScience, LNCS* (Vol. 3234, pp. 299–312). Springer.
Stumme, G., & Maedche, A. (2001). FCA-MERGE: Bottom-up merging of ontologies. In *Proceedings of the 17th international joint conference on artificial intelligence (IJCAI 2001)* (pp. 225–234). San Francisco, USA: Morgan Kaufmann Publishers Inc.
Tazi, S., Al-tawki, Y., & Drira, K. (2003). Editing pedagogical intentions for document reuse. In *Proceedings of the 4th international conference on information technology based higher education and training (ITHET '03)* (pp. 274–278). Marrakech: Morocco: IEEE Computer Society.
Tiropanis, T., Davis, H., Millard, D., & Weal, M. (2009). Semantic technologies for learning and teaching in the web 2.0 Era. *IEEE Intelligent Systems, 24*(6), 49–53.
Vargas-Vera, M., Motta, E., Domingue, J., Lanzoni, M., Stutt, A., & Ciravegna, F. (2002). MnM: Ontology driven semi-automatic and automatic support for semantic markup. In *Proceedings of the 13th international conference on knowledge engineering and knowledge management. ontologies and the semantic web (EKAW '02)* (pp. 213–221). London, UK: Springer.
Vargas-Vera, M., Moreale, E., Stutt, A., Motta, E., & Ciravegnan, F. (2007). MnM: semi-automatic ontology population from text. In R. Sharman et al. (Eds.), *Ontologies: a handbook of principles, concepts and applications in information systems* (pp. 373–402). New York, USA: Springer.
Volker, J., & Rudolph, S. (2008). Lexico-logical acquisition of OWL DL axioms: an integrated approach to ontology refinement. In R. Medina et al. (Eds.). *ICFCA, LNAI* (4933, pp. 62–77). Springer.
Wille, R. (1992). Concept lattices and conceptual knowledge systems. *Computers and Mathematics with Applications, 23*, 493–515.
Wille, R. (2009). Restructuring lattice theory: an approach based on hierarchies of concepts. In *Proceedings of the 7th international conference on formal concept analysis (ICFCA '09)* (pp. 314–339). Berlin, DE: Springer.
Xu, H., & Xiao, D. (2009). Building information specification ontology for computer network management based on formal concept analysis. In *Proceedings of the international conference on information and automation (ICIA'09)* (pp. 312–317). Macau, China: IEEE Computer Society.
Xu, W., Li, W., Wu, M., Li, W., & Yuan, C. (2006). Deriving event relevance from the ontology constructed with formal concept analysis. In A. Gelbukh (Ed.). *CICLing, LNCS* (3878, pp. 480–489). Springer.
Zhao, Y., Halang, W. A., & Wang, X. (2007). Rough ontology mapping in e-business integration. *Studies in Computational Intelligence, 37*, 75.

6.4 Browsing digital collections with reconfigurable faceted thesauri

Cita Completa:

Gayoso-Cabada, J., Rodríguez-Cerezo, D., & Sierra, J.-L. (2016). Browsing Digital Collections with Reconfigurable Faceted Thesauri. En Proceedings of the 25th International Conference on Information Systems Development, ISD 2016 (pp. 378-389)

Resumen original de la contribución:

Faceted thesauri group classification terms into hierarchically arranged facets. They enable faceted browsing, a well-known browsing technique that makes it possible to navigate digital collections by recursively choosing terms in the facet hierarchy. In this paper we develop an approach to achieve faceted browsing in live collections, in which not only the contents but also the thesauri can be constantly reorganized. We start by introducing a digital collection model letting users reconfigure facet hierarchies. Then we introduce navigation automata as an efficient way of supporting faceted browsing in these collections. Since, in the worst-case, the number of states in these automata can grow exponentially, we propose two alternative indexing strategies able to bridge this complexity: inverted indexes and navigation dendrograms. Finally, by comparing these strategies in the context of Clavy, a system for managing collections with reconfigurable structures in digital humanities and educational settings, we provide evidence that navigation dendrogram organization outperforms the inverted index-based one.

Referencias Bibliográficas:

(Ben-Yitzhak et al., 2008; Berchtold, Böhm, Keim, Kriegel, & Xu, 2000; Chodorow, 2013; Cigarrán-Recuero et al., 2014; Culpepper & Moffat, 2010; Godin, Saunders, & Gecsei, 1986; Grainger, Potter, & Seeley, 2014; Greene, 2015; Greene, Dunaiski, Fischer, Ilvovsky, & Kuznetsov, 2015; Greene & Fischer, 2015; Hildebrand, van Ossenbruggen, & Hardman, 2006; Huang et al., 2014; Jain, Murty, & Flynn, 1999; Kriegel, 1984; Kuznetsov, 2001; C. Li, Yan, Roy, Lisham, & Das, 2010; R. Li, Bao, Yu, Fei, & Su, 2007; McCandless et al., 2010; Nasir Uddin & Janecek, 2007; Perugini, 2010; Radelaar, Boor, Vandic, Van Dam, & Fasincar, 2014; Sarmah et al., 2015; schraefel, Wilson, Russell, & Smith, 2006; J. L. Sierra & Fernández-Valmayor, 2008; José Luis Sierra et al., 2006; Daniel A. Smith et al., 2005; Daniel Alexander Smith et al., 2007; Tunkelang, 2009; Wray & Eklund, 2010; Yee et al., 2003; Zhang, Li, Gurrin, & Smeaton, 2016; Zheng, Zhang, & Feng, 2013; Zobel & Moffat, 2006)

Fé de erratas:

- En la página 385 del artículo, donde dice:

  *The **vertical** axis corresponds to the number of operations carried out so far.*
  *The **horizontal** axis corresponds to cumulative time (in seconds).*

  debe decir:

  *The **horizontal** axis corresponds to the number of operations carried out so far.*
  *The **vertical** axis corresponds to cumulative time (in seconds).*

Browsing Digital Collections with Reconfigurable Faceted Thesauri

Joaquín Gayoso-Cabada *jgayoso@ucm.es*
Fac. Informática. Complutense University of Madrid
Madrid, Spain

Daniel Rodríguez-Cerezo *drcerezo@ucm.es*
Fac. Informática. Complutense University of Madrid
Madrid, Spain

José-Luis Sierra *jlsierra@ucm.es*
Fac. Informática. Complutense University of Madrid
Madrid, Spain

Abstract

Faceted thesauri group classification terms into hierarchically arranged facets. They enable *faceted browsing*, a well-known browsing technique that makes it possible to navigate digital collections by recursively choosing terms in the facet hierarchy. In this paper we develop an approach to achieve faceted browsing in live collections, in which not only the contents but also the thesauri can be constantly reorganized. We start by introducing a digital collection model letting users reconfigure facet hierarchies. Then we introduce *navigation automata* as an efficient way of supporting faceted browsing in these collections. Since, in the worst-case, the number of states in these automata can grow exponentially, we propose two alternative indexing strategies able to bridge this complexity: *inverted indexes* and *navigation dendrograms*. Finally, by comparing these strategies in the context of *Clavy*, a system for managing collections with reconfigurable structures in digital humanities and educational settings, we provide evidence that navigation dendrogram organization outperforms the inverted index-based one.
Keywords: Faceted Browsing, Faceted Thesauri, Indexing, Reconfigurable Collections

1. Introduction

Faceted navigation is a common interaction technique in business, the cultural industry and many other domains [2],[18],[26],[29],[30],[32]. For this purpose, resources are classified in terms of suitable *faceted thesauri*. A faceted thesaurus groups classification terms into facets, which in turn can have associated sub-facets, yielding a hierarchical arrangement. This hierarchical organization can, in turn, guide navigation through the underlying collection of digital resources (regardless of whether these are records in a database, objects in a virtual museum, entries in a virtual shop catalog, or any other type of digital object). In mature digital collections, in which there are few or no changes in the underlying resources, and in which classification schemata are pre-established and stay immutable, faceted navigation can be accomplished in very efficient ways [7]. However, for live collections, such as those arising in social or other highly dynamic and changing environments, not only are changes in the underlying resources frequent, but these changes can also affect the classification schemata themselves. When faceted thesauri are used in these dynamic settings, reconfiguring the thesaurus can mean a profound rearrangement of the collection's internal structures, which can be costly in time (often, it must be carried out offline). In consequence, user experience can be seriously hindered. Indeed, when a user changes the thesaurus, what he/she probably expects is an almost instant response in navigation; in these cases high response times and/or a temporarily outdated underlying information system are inadmissible. We have realized this fact during the compilation of research and education-oriented collections of digital objects in digital

humanities scenarios [4],[23],[24]. In these scenarios faceted thesauri-like classification schemata were subjected to continuous change, refinement and evolution throughout the collections' life cycles. Many times those reconfigurations in the schemata were performed with experimental and/or exploratory purposes in mind, and domain experts (researchers and/or instructors in charge of compiling and maintaining the collections) were not willing to wait for long periods until the changes were reflected in their collections. On the contrary, they wanted to see the changes in the browsing system immediately after changing the classification schemata, in order to determine whether these changes in the schemata really met their expectations. Thus, in this paper we partially respond to these needs by firstly providing a model of digital collection with a *reconfigurable* faceted thesaurus, in which the facet hierarchy can be freely rearranged, thus accomplishing the exploratory needs of the potential users. Secondly, we also introduce indexing strategies that provide reasonable time-space tradeoffs concerning navigation reconfigurability, while preserving acceptable levels of user experience.
The rest of the paper is organized as follows. Section 2 describes the digital collection model. Section 3 addresses browsing in the presence of the kind of reconfigurable thesauri introduced by this model. Section 4 introduces some works related to our browsing approach. Finally, section 5 outlines the final conclusions and some lines of future work.

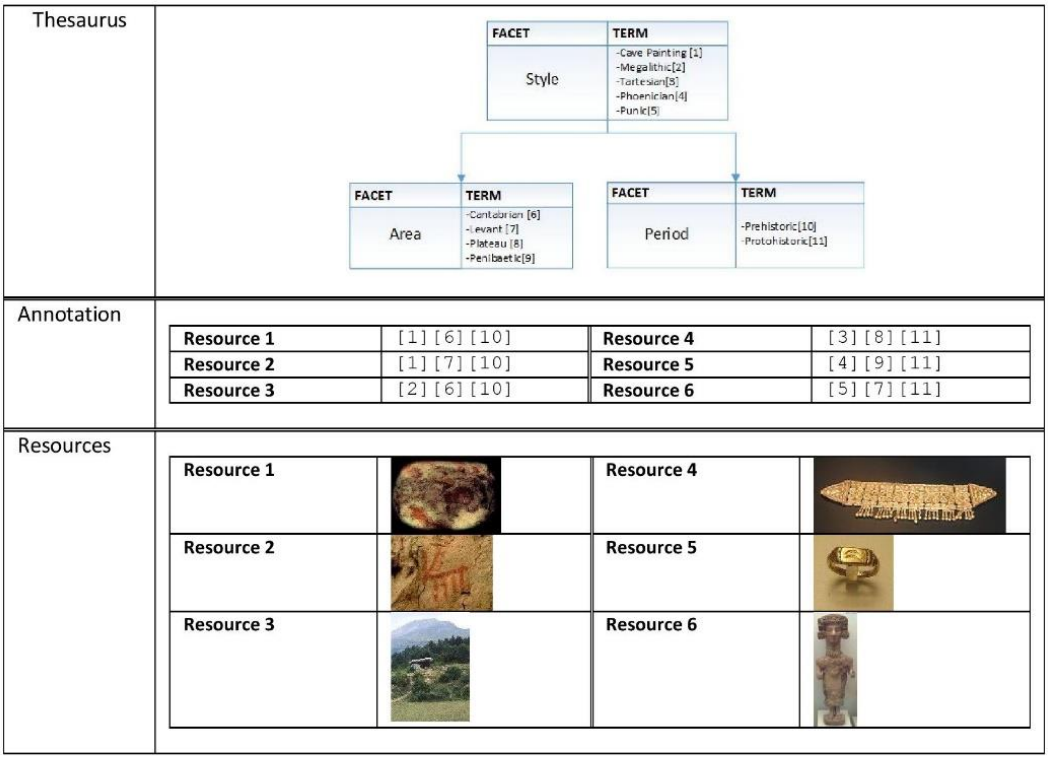


Fig. 1. A small digital collection

2. Digital Collections with Reconfigurable Faceted Thesauri

In this section we introduce our model of digital collection with reconfigurable thesauri. Subsection 2.1 describes the structure of these collections. Subsection 2.2 addresses thesauri reconfiguration.

Structure of the Collections

Our collections comprise the following parts (see Fig.1 for an example):

- On one hand, there are the *resources* in the collection. These resources are digital objects whose nature is no longer constrained by the model. Thus, these resources can be media

files (images, sound, video, etc.), external resources identified by their URIs, or entities of a more abstract nature (tuples of a table in a relational database, records in a bibliographical catalog, elements in an XML document, rows in a spreadsheet, etc.). For instance, the small collection depicted in Fig. 1 includes six image archives as resources, corresponding to photographs of artistic objects from the Prehistoric and Protohistoric artistic periods in Spain (Fig. 1 actually shows thumbnails of these images).

- On another hand, there is the *annotation* of the resources. This annotation consists of associating descriptive *terms* with resources. These terms are useful when cataloguing resources and, therefore, they enable future uses of the collection (navigation, search, etc.). Since each term has a unique identifier associated, annotating a resource consists of associating a set of term identifiers with such a resource. For instance, in Fig.1 resource number 1 has the terms identified by `[1][6]` and `[10]` associated.
- Finally, there is a faceted *thesaurus* that organizes the terms into facets and which arranges these facets hierarchically. For instance, the faceted thesaurus in Fig. 1 includes a root facet *Style*, representing the artistic style used, and two sub-facets: *Area* (representing the geographical area), and *Period* (representing the artistic period). Each facet includes representative terms related to this facet. Notice how each term consists of a descriptive name and the aforementioned unique identifier. In this way, the terms indicated below (`[1][6]` and `[10]`) actually refer to the terms *Cave Painting* in the facet *Style*, *Cantabrian* in the facet *Area*, and *Prehistoric* in the facet *Period*.

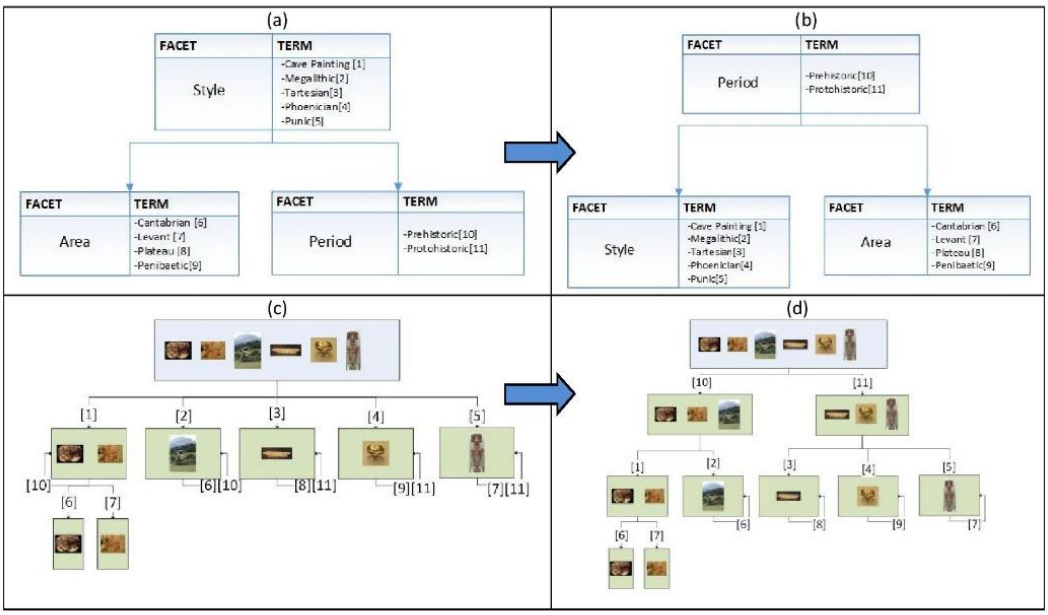


Fig. 2. (a) Original thesaurus in Fig. 1; (b) Reconfigured thesaurus; (c) Navigation map induced by (a); (d) Navigation map induced by (b)

Thesauri Reconfiguration

Our model lets users reconfigure thesauri by rearranging the hierarchical organization of facets in order to accommodate their experimental and/or exploratory needs. For instance, Fig. 2 shows an example concerning the collection in Fig. 1. Indeed, the thesaurus in Fig.2a (the original thesaurus in Fig. 1) reflects a structure primarily focused on the artistic style. Once this style has been set, it is possible to introduce either a geographical or an artistic period refinement. However, it may also be feasible to conceive of an alternative organization, with the artistic period as main focus, and with the geographical area and style as secondary features. This leads to the thesaurus in Fig. 2b, which has been obtained from the original one by altering the hierarchical facet arrangement.

Since the organization of a collection ultimately relies on its faceted thesaurus, by reconfiguring this thesaurus it is possible to implicitly reconfigure the structure of the entire collection, adapting it to different use scenarios as needed. This effect can be readily appreciated on the *navigation map* of a collection. Such a map is a directed graph in which:

- Nodes represent sets of resources, and arcs are labelled with terms used to narrow down the resources in the source nodes in order to yield the resources in the target ones (actually, all those resources in the source nodes annotated with the terms in the arcs).
- Structure is constrained by the facet hierarchy. In this way, root nodes can only be narrowed down with terms in root facets, and, if a node is produced by a term in a facet, it can only be narrowed down with terms in sub-facets of the mentioned facet.

Thus, reconfigurations in the thesaurus affect the entire navigation map. It is made apparent in Fig. 2c and Fig. 2d, which, respectively, outline navigation maps of the collection in Fig.1 before (Fig. 2c) and after (Fig. 2d) thesaurus reconfiguration.

3. Browsing with Reconfigurable Faceted Thesauri

As the previous section makes apparent, reconfigurations in the hierarchical structure of a thesaurus profoundly impact the structure of the overall collection. In particular, after reconfiguration, the collection's navigation map can be completely altered. This hampers the use of efficient implementations of faceted browsing (e.g., [7]), which are basically driven by the navigation map structure and therefore require pre-established and unmodifiable faceted thesauri (otherwise, the navigation map would have to be regenerated after each thesaurus' reconfiguration, which would be a costly task even for collections of moderate size, and which could seriously impact user experience). Thus, by allowing reconfigurability, it is necessary to switch to alternative representations, enabling *all* the possible navigations induced by *all* the possible reconfigurations of the faceted thesaurus. Subsection 3.1 characterizes the expected behavior as a finite state machine. Subsection 3.2 addresses some complexity issues associated with a naïve representation directly based on such a machine. Finally, subsection 3.3 discusses some indexing approaches that we have used to face these complexity issues.

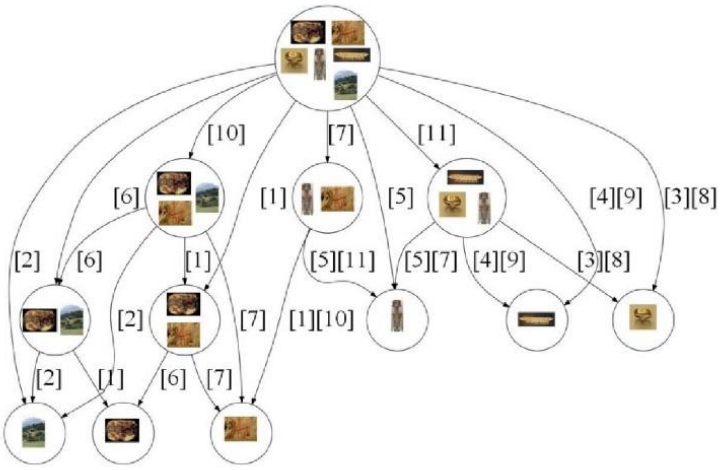


Fig. 3. Navigation automaton for the collection in Fig. 1

Navigation Automata

In order to characterize all the possible navigations induced by all the possible reconfigurations of a faceted thesaurus it is necessary to free terms from the faceted structure. Therefore, a plain set of terms must be considered and, in each interaction state, all the meaningful selection of terms must be applied. The result can be represented as a finite state machine that we will call a *navigation automaton*. This automaton will consist of *states* labelled by sets of resources, and *transitions* labelled by terms. More precisely:

- There will be an initial state labelled by all the resources in the collection.
- Given a state *S* labelled by a set of resources *R*, for each term *t* annotating some resource in *R* there will be a state *S'* labelled by all the resources in *R* annotated by *t*, as well as a transition from *S* to *S'* labelled by *t*.[20]

Fig. 3 shows the navigation automaton for collection in Fig. 1. Notice that the navigation automaton does not depend on the hierarchical organization of facets in the thesaurus, but only on the terms and on the resources in the collection. Therefore, it is not affected by reconfigurations in the thesaurus. Indeed, it can be thought of as the amalgam of all the possible navigation maps induced by all the possible reconfigurations of the collection thesaurus.

Since the navigation automaton embeds all the possible navigation maps, faceted browsing with respect to a particular thesaurus configuration can be formulated in a straightforward way, since there will be a direct correspondence among interaction states in the browsing process and states in the navigation automaton. In addition, in each interaction state will be a set of allowable facets to be explored. Indeed:

- The browsing process will start by considering the navigation automaton's initial state and the thesaurus' root faces as allowable ones.
- In each interaction state, the allowable facets will be used to constraint the possible terms to continue browsing. Once an allowed term is selected, the navigation automaton will be used to establish the new navigation state and the thesaurus to update the allowable facets.

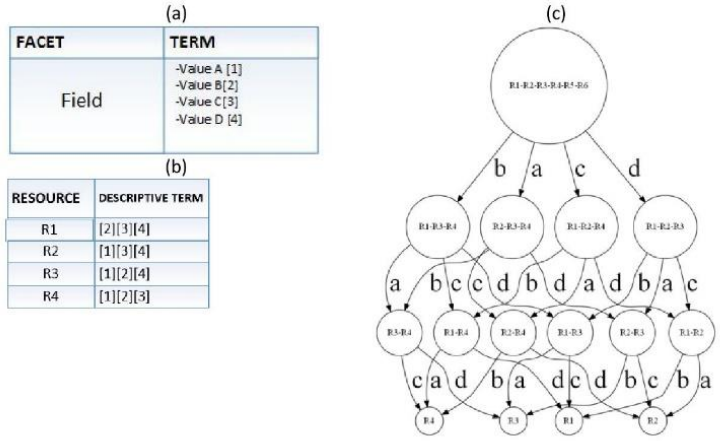


Fig. 4. Example of exponential explosion: (a) a simple thesaurus, (b) a simple associated collection, (c) the resulting navigation automaton

Complexity Issues

As indicated in the previous subsection, the explicit availability of the navigation automaton provides an elegant and efficient solution to faceted browsing in the presence of a reconfigurable thesaurus. Unfortunately, for collections with dense annotations there is the risk of facing unacceptably growing rates in the number of resulting states. It should not be surprising since we are attempting to represent all the possible ways of navigating in a single structure, regardless of the structure of the underlying thesaurus. In the worst case, the number of states can grow exponentially with respect to the number of resources. This extreme case, in which the number of states is 2^n-1 (with *n* the number of resources), arises, for instance, by distinguishing each pair of resource annotations in a single term (Fig. 4 shows an example with 4 resources and 4 terms).

While the extreme case presented can be somewhat artificial, it cannot be ignored if we want to deal with arbitrary evolving collections. For this purpose, it can be desirable to look for alternative indexing approaches to enable the dynamic recreation of the relevant parts of the navigation automaton during browsing while preserving required levels of user experience.

[20] Notice that *S* and *S'* can be the same -when all the resources in *R* are annotated by *t*.

Indexing Strategies

In order to deal with the complexity issues raised in the previous subsection, we have explored two different indexing strategies: *inverted indexes* and *navigation dendrograms*. Next paragraphs analyze these strategies and provides some empirical comparison results.

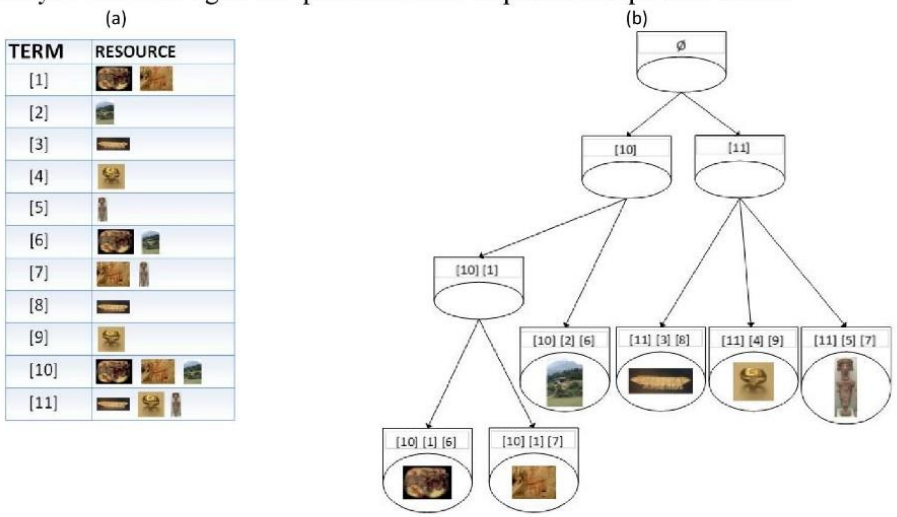


Fig. 5. (a) An inverted index for the collection in Fig. 1(a); (b) a navigation dendrogram for such a collection

Inverted Indexes

Inverted indexes are standard artifacts used for information retrieval [33]. Basically, for each description term, an inverted index associates the set of resources annotated with this term. Fig. 5a shows an example of inverted index for the collection in Fig. 1.

Inverted indexes can be used to recreate the browsing behavior described in the previous section in a straightforward way. Basically, it suffices to maintain interaction states consisting of the set of terms chosen and the set of allowable facets to be explored.

- The initial state is constituted by the empty set of terms and by the root facets.
- Given an interaction state the selection of the next term to be applied obeys the same constraints as with the navigation automaton: this term must be a term t included in some facet f among the allowable facets. Then, the new interaction state will consist of: (i) all the terms in the previous one plus the new term selected, (ii) all the sub-facets of f.

Concerning the resources filtered in each interaction state, these resources can be determined by considering the set of terms $\{t_1, .., t_n\}$ in such a state and by evaluating the conjunctive query $t_1 \wedge \ldots \wedge t_n$ using the inverted index.

The cost of evaluating the queries $t_1 \wedge \ldots \wedge t_n$ in each interaction state constitutes the main shortcoming of the approach. Indeed, it involves finding the intersections of the sets for $t_1 \ldots t_n$ in the inverted index. While there has been extensive research in performing these intersection operations efficiently [5], the cost is not negligible.[21] On the positive side is the availability of many mature implementations and frameworks that can be used in a straightforward way to support the technique. For instance, in our experiences, we used Lucene [17] for such a purpose.

(a)

Add resource *r* to dendrogram *d*:

CurrentNode = *d*'s root
Δ_{Res} = *r*'s terms

(b)

Find next interaction state of *s* given *t of facet f*:
notation: Given a dendrogram's node *n*:

- $n^{\uparrow}$: all the ancestors of *n* (including *n* itself)
- $n^{\downarrow}$: all the descendants of *n* (exluding n)

let *N* the set of dendrogram nodes in *s* {

[21] Notice that, although as indicated earlier, frameworks like Solr support faceted browsing in a straightforward and efficient manner by identifying paths in the thesaurus with terms, in our context these features are useless, since thesauri can be reconfigured anytime, thus invalidating this solution. So we are confined to explicitly evaluating conjunctive queries in each interaction state.

```
while there is some child n of CurrentNode such as
      n's filtering terms ⊆ ΔRes {
   CurrentNode = One of such child nodes
   ΔRes = ΔRes – CurrentNode's filtering terms
}

InsertionNode = CurrentNode
if (ΔRes ≠ ∅) {
  if there is some child n of InsertionNode such as
     n's filtering terms ∩ ΔRes ≠ ∅ {
       ChildNode = One of such child nodes
       ForkNode = create new node
       ΔFork = ΔRes ∩ ChildNode's filtering terms
       ΔChild = ChildNode's filtering terms - ΔFork
       ΔRes = ΔRes - ΔFork
       Set ForkNode's filtering terms to ΔFork
       Set ChildNodes'filtering terms to ΔChild
       change the arc InsertionNode → ChildNode to
                      InsertionNode → ForkNode
       add an arc ForkNode → ChildNode
       InsertionNode = ForkNode
       }
```

```
N' = ∅
foreach n in N {
  if there is n' in n↑ such as t is in the filtering terms of n' {
    N' = N' ∪ {n}
  }
  else let n↓t the nodes in n↓ with t in their filerting terms {
    N' = N' ∪ n↓t
  }
}

let F the set of subfacets of f {
 The next interaction state has N' as the set of dendrogram nodes
 and F as the set of allowable subfacets
}
```

(a. cont)

```
 if (ΔRes ≠ ∅)) {
 HostNode = create new node
 set HostNode's filtering terms to ΔRes
 set HostNode's own resources to ∅
 add an arc InsertionNode → HostNode
 InsertionNode = HostNode
 }
}
add r to InsertionNode's own resources
```

Fig. 6. (a) Pseudocode of the process for adding a resource to a navigation dendrogram; (b) Pseudocode to find the next interaction state during browsing

Navigation Dendrograms

In order to avoid the proliferation of intersection operations, which is characteristic of inverted indexes representations, we have envisioned a tree-shaped indexing scheme inspired by *dendrograms* in hierarchical clustering [13]. The resulting representations are called *navigation dendrograms*. Following hierarchical clustering principles, nodes in the dendrogram represent subsets of the overall resource set. In this way:

- The dendrogram's root represents the whole resource set.
- If a node represents a particular resource set, then each child node represents a partition of this set (i.e., child nodes represent mutually disjoint subsets of the parent's set).

The resource set associated to a node is not explicitly stored in this node. Instead, each resource is only hosted in one node (the resource's *host node*). Resources placed in a node are called the mentioned node's *own resources*. The overall resource set of a node is given by its own resources and by all the own resources of its descendants. Finally, in order to partition the resource space, each node has a set of *filtering terms* associated, so that all the own resources in the node and in all their descendants' must be annotated with these filtering terms (the node is said to *filter* those resources). Fig. 5b shows a navigation dendrogram for the collection in Fig. 1.

Initially, the dendrogram contains a single root node with an empty filtering set. Then, the dendrogram is incrementally constructed by sequentially adding resources, one resource at a time. Pseudocode in Fig. 6a details how a new resource is added to the dendrogram. Notice that, in the worst case, insertion of a resource involves the creation of two new nodes. In consequence, the number of nodes in the dendrogram is bound by $2R$ (with R the number of resources).

Concerning browsing, it is possible to conceive interaction states formed by:

- A set of *dendrogram nodes*, which are active in the interaction state.
- A set of *allowable facets*.

As in the earlier proposals, navigation firstly proceeds by selecting a term from one of the allowable facets. Then, the next interaction state can be obtained as sketched in Fig. 6b. Notice that, in order to speed up the computation shown in Fig. 6b, it is convenient to have direct access to the filtering terms for each node and its ancestors, as well as to have the node's descendants

classified by their filtering terms (for the sake of simplicity, details are not shown in the pseudocode in Fig. 6).
Once the interaction state is determined, resources selected in an interaction state can be lazily recovered by iterating the dendrogram nodes and their descendants' own resource sets.
Finally, it is worthwhile to notice how navigation dendrograms overcome the main shortcoming of inverted indexes: the need to explicitly carry out set intersections during browsing. On the negative side, the indexing process is substantially more complex than in the case of inverted index construction.

Experimental Results

In order to compare the two indexing strategies described, we implemented both on *Clavy*, an experimental system for managing digital collections with reconfigurable faceted thesauri-like schemata.[22] We also set up an experiment consisting of adding the resources in *Chasqui* [23],[23] a digital collection of 6283 digital resources on Precolombian American archeology, to *Clavy* and to simulate runs concerning browsing and schemata reconfiguration operations.
Each run was customized as follows. We interleaved resource insertion with browsing / reconfiguration rounds. Each insertion round consisted of 100 resource insertions (with the exception of the last one, in which the remaining resources where inserted). In turn, each browsing / reconfiguration round consisted of executing $0.1n$ browsing operations randomly interleaved with $0.01n$ reconfigurations (n being the number of resources inserted so far). Each browsing operation consisted, in turn, of selecting a feasible term, computing the next interaction state, and visiting all the resources filtered. Reconfiguration operations, then, consisted of feasible interchanges of two randomly selected facets,[24] followed by a browsing step.
Inverted indexes were managed using Lucene, while navigation dendrograms were managed using our own implementation (implemented in Java, as well as the Lucene framework). In both cases, in-memory indexes were used in order to avoid side effects of persistence, disturbing the experiment.

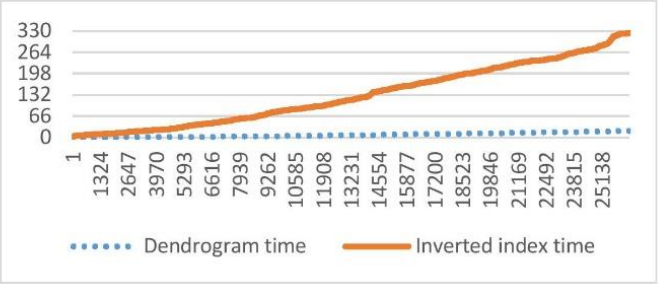


Fig. 7. Cumulative time of inverted indexes vs. dendrograms

Fig. 7 shows the results obtained from the two runs. The experiment was run on a PC with Windows 10, with a 3.4GHz Intel microprocessor, and with 8Gb of DDR3 RAM. The vertical axis corresponds to the number of operations carried out so far. The horizontal axis corresponds to cumulative time (in seconds). As is made apparent, the dendrogram-based approach clearly outperforms the inverted indexes (even though we are using a highly optimized framework, like Lucene, for inverted indexing vs. our own in-house experimental implementation for dendrograms).

[22] http://clavy.fdi.ucm.es/Clavy/
[23] http://oda-fec.org/ucm-chasqui
[24] By *feasible* we mean avoiding cycles in the resulting thesaurus.

4. Related Work

There are several faceted browsing systems that, like ours, envision the possibility that the user reconfigures the underlying facet hierarchy. For instance, *mSpace* [22] makes it possible to organize information spaces in plain sets of facets, which can be interactively arranged in lists (*slices*), representing linear hierarchies between selected facets. In [21] an implementation based on semantic web (RDF) technologies is proposed. In [25] the advantages and disadvantages of this implementation are analyzed and more conventional solutions based on relational databases are proposed. In turn, */facet* [11] is a multi-faceted browser driven by RDFS schemata and RDF data. Facet hierarchies are inferred from RDF class hierarchies, and the user can jump from one facet hierarchy to another while maintaining filtering constraints. In [11] some efficiency issues associated with the extensive use of semantic web infrastructures in */facet* are reported. Contrary to *mSpace*, in which facets lack organization and in which users confer (linear) structures on these facets by arranging them in slices, or to */facet*, in which RDFS schemata pre-establish faceted hierarchies and in which the user is allowed to jump between hierarchies, in our approach, facets are doted with a hierarchical structure (the faceted thesaurus) from the beginning and users organize this hierarchy according to their needs. It provides users with more guidance during the reconfiguration process than *mSpace*'s approach and more freedom than */facet*'s (where reconfigurability consists of dynamically pasting different pre-established hierarchies). In addition, we propose efficient indexing approaches, specifically tailored to our model instead of piggybacking on general-purpose semantic web or relational database solutions.

Our navigation automaton model is actually equivalent to lattice-based proposals to browse information spaces, as described in the seminal work of [6]. In these proposals, resources are tagged with keywords. The lattice organization induced consists of nodes characterized by sets of resources and sets of keywords related by a *Galois connection* (i.e., the set of keywords is the intersection of the resources' keywords and the set of resources consists of all the resources filtered by the keyword set). This organization is actually the main subject of the fertile theory of *formal concept analysis* [20], where resources are called *objects*, keywords are called *attributes*, objects tagged with attributes are called *formal contexts*, and lattice nodes are called *formal concepts*. Thus, states of our navigation automata can actually be identified with formal concepts, and automata themselves with explicit representations of concept lattices (with an explicit representation of the whole order relation and an explicit labelling of the arcs with transition information). In [15] the intrinsic complexity of formal concept analysis is examined and the problem of determining the size of concept lattices is proved to be a #P-complete one (i.e., harder than NP-complete). In consequence, complexity results in concept lattice theory are directly translatable to navigation automata (in fact, construction of section 3.2 was suggested by the proof of theorem 1 in [15]). In addition, there are several proposals on using concept lattices as the underlying indexing structures of digital collections (e.g., [8],[9],[10],[28]), which also can be affected by the worst-case complexity of formal concept analysis.

Inverted indexes have been extensively used to support faceted browsing. In [31] the basic technique, as well as subsequent enhancements, are illustrated with the use of Lucene. In [27] an alternative approach, based on relational databases, is presented. However, all these approaches are based on the assumption of pre-established and immutable faceted thesauri. As noticed in [1], if this assumption is left out, and therefore arbitrary multilevel exploratory search is allowed, inverted indexes can become costly due to the set operations involved. For small amounts of terms, multidimensional structures (as used in data-warehouse and data-mining scenarios) can be advantageous [14]. However, the performance of these multidimensional approaches can dramatically decrease when dimensionality increases. For this purpose, in [1] a technique called *tree striping* is described, which proposes subdividing the overall information space in disjoint sub-spaces, to apply standard inverted indexes or multidimensional indexing techniques to each resulting subspace and to use an efficient merging approach to aggregate the results. Nevertheless, and contrarily to our navigation dendrograms, both multi-dimensional

and tree striping techniques basically work with pre-established partitions of the information space.
Finally, it is worthwhile to notice that clustering techniques has been extensively used in social tagging systems (e.g., [12],[16],[19]) to enable the discovering of useful semantic relationships among tags in order to provide better guidance to users (e.g., by automatically discovering hierarchical structures of tags). Thus, clustering in these approaches is oriented to enhance users' browsing efficiency, while our navigation dendrograms are oriented to enhance system performance.

5. Conclusions and Future Work

Live digital collections, which involve active communities of specialized users (e.g., researchers or educators in a particular field), also require live organization schemata, which can be incrementally defined, refined and enhanced as collections evolve. In addition, in these scenarios users usually want systems to quickly respond to changes in the schemata, without waiting for costly and/or batch reorganization processes. In this paper we have addressed this problem of dynamic reconfigurability in the case of reconfigurable faceted thesauri, in which users can re-order facets in order to explore different and alternative ways of organizing the collections. Since facet hierarchy can be rearranged in unexpected ways, it is necessary to resort to a more free and exploratory browsing system. It has led us to model this system as a finite state machine, the *navigation automaton*, taking into account all the possible ways of navigating the collection, using terms selected from the facets. Unfortunately, we have also showed how, in some cases, the number of states in this automaton can increase exponentially with respect to the collection's size. In order to deal with this potential exponential factor we have explored two different indexing approaches: one based on standard inverted indexes (implemented in a robust and well-proven search framework: Lucene), and one inspired by hierarchical clustering techniques (the so-called *navigation dendrograms*). Some experiments with a real collection gave evidence of how the hierarchical clustering technique can outperform the inverted indexing one.
We are currently working on further optimizing our navigation dendrogram representation to leverage space requirements. Indeed, in order to provide efficient navigation we need to associate each node with the intersection and the union of all the terms annotating the resources under this node. In addition, for each union term we also need to store the descendant nodes that include such a term in their filtering sets. Fortunately, these sets present much regularity among nodes, which allows us to compress them by using tries and common node-set stores. Since the resulting structures provide time and space efficient representations for the nodes' intersection and union sets, all these optimizations enhance system performance, in addition to saving space. We are also looking for efficient ways to persist all this information, either by using standard relational databases or alternative NoSQL approaches (e.g., [3]), while causing a minimum impact on system performance. Once efficient persistence mechanisms are established we want to run more empirical evaluations also taking persistence into account. We also hope to enhance our model with support for arbitrary Boolean queries and for different ways of exploring the resources selected. These mechanisms will be based on the navigation automaton model (supported by our indexing proposals, and, in particular, by navigation dendrograms) in order to get as much efficiency as possible. Finally, we plan to perform more comprehensive tests of our model in the context of different Digital Humanities efforts carried out by some of the Humanities research groups with whom we cooperate. Among these efforts we can highlight, in addition to the aforementioned *Chasqui* collection, different digital collections curated by LEETHI, the UCM research group on European and Spanish Literatures, from Texts to Hypermedia (Mnemosine, a digital collection concerning rare texts from the Spanish Silver literature period[25], Ciberia, a digital collection concerning Spanish digital

[25] repositorios.fdi.ucm.es/mnemosine/

literature[26] and Tropos, a digital collection concerning creative digital writing for literature education[27]), as well as those concerning the Panamanian "El Caño" archeological site[28], compiled and curated by "El Caño" Foundation at Panama.

Acknowledgements

This work has been supported by the BBVA Foundation (research grant HUM14_251) and by the Spanish Ministry of Economy and Competitiveness (research grant TIN2014-52010-R).

References

1. Berchtold, S., Böhm, C., Keim, D-A., Kriegel, H-P., Xiaowei, X.: Optimal Multidimensional Query Processing Using Tree Striping. In: Proceedings of the 2nd International Conference on Data Warehousing and Knowledge Discovery, pp. 244-257. Springer, London, UK (2000)
2. Chengkai, L., Ning, Y., Senjuti, B-R., Lekhendro, L., Gautam, D.: Facetedpedia: Dynamic Generation of Query-Dependent Faceted Interfaces for Wikipedia. In: Proceedings of the 19th International World Wide Web Conference, pp. 651-660. ACM, Raleigh, NC, USA (2010)
3. Chodorow, K. MongoDB: The Definitive Guide. O'Reilly (2013)
4. Cigarrán-Recuero, J., Gayoso-Cabada, J., Rodríguez-Artacho M., Romero-López, D., Sarasa-Cabezuelo, A., Sierra, J-L.: Assessing Semantic Annotation Activities with Formal Concept Analysis. Expert Syst. with Applications 44(11), 5495-5508 (2014)
5. Culpepper, J-S., Moffat, A.: Efficient Set Intersection for Inverted Indexing. ACM Transations on Information Systems 29(1), article 1 (2010)
6. Godin, R., Saunders, G.: Lattice Model of Browsable Data Space. Information Sciences 40(2), 89-116 (1986)
7. Grainger, T., Potter, T.: Solr in Action. Manning Publications (2014)
8. Greene, G-J., Dunaiski, M., Fischer, B.: Browsing Publication Data using Tag Clouds over Concept Lattices Constructed by Key-Phrase Extraction. In: Proceedings of Russian and South African Workshop on Knowledge Discovery Techniques Based on Formal Concept Analysis, pp.10-22. CEUR, Stellenbosch, South Africa (2015)
9. Greene, G-J., Fischer, B.: Interactive Tag Cloud Visualization of Software Version Control Repositories. In: Proceedings of the third IEEE Working Conference on Software Visualization, pp. 56-65. IEEE, Raleight, NC, USA (2015)
10. Greene, G-J.: A Generic Framework for Concept-Based Exploration of Semi-Structured Software Engineering Data. In: Proceedings of the 30th IEEE/ACM International Conference on Automated Software Engineering, pp. 894-897. ACM, Lincoln, Nebraska, USA (2015)
11. Hildebrand, M., Ossenbruggen, J-v., Hardman, L.: /facet: A Browser for Heterogeneous Semantic Web Repositories. In: Proceedings of the 5th International Semantic Web Conference, pp. 272-285. Springer, Athens, GA, USA (2006)
12. Huang, J-W., Chen, K-Y., Chen, Y-C., Yang, K-N., Hwang, S., Huang, W-C.: A Novel Spatial Tag Cloud Using Multi-Level Clustering. Journal of Information Science and Engineering 30, 687-700 (2014)
13. Jain, A-K., Murty, M-N., Flynn, P-J.: Data Clustering: a Review. ACM Computing Surveys 31(3), 264-323 (1999)
14. Kriegel H.-P.: Performance Comparison of Index Structures for Multi-Key Retrieval. Proceedings of the ACM SIGMOD International Conference on Management of Data, pp. 186-196. ACM, Boston, MA (1984)
15. Kuznetsov, S. On computing the size of a lattice and related decision problems. Order 18(4), 313–321 (2001)

[26] repositorios.fdi.ucm.es/CIBERIA
[27] repositorios.fdi.ucm.es/Tropos/
[28] oda-fcc.org/nata

16. Li, R., Shenghua, B., Fei, B., Su, Z., Yu, Y. Towards Effective Browsing of Large Scale Social Annotations. In: Proceedings of 16th International World Wide Web Conference, pp. 943-952. ACM, Banff, Alberta, Canada (2007)
17. McCandless, M., Hatcher, E., Gospodnetic, O.: Lucene in Action, 2nd Edition. Manning Publications (2010)
18. Perugini, S.: Supporting Multiple Paths to Objects in Information Hierachies: Faceted Classification, Faceted Search, and Symbolic Links. Information Processing and Management 46(1), 22-43 (2010)
19. Radelaar, J.; Boor, A-J.; Vandic, D.; van Dam, J-W.; Fasinca, F. Improving search and exploration in tag spaces using automated tag clustering. Journal of Web Engineering 13(3-4), 277-301 (2014)
20. Sarmah, A-K., Hazarika, S-M., Sinha, S-K.: Formal Concept Analysis: Current Trends and Directions. Artificial Intelligence Review 44(1), 47-86. (2015)
21. Schraefel, M-C., Smith, D-A., Owens, A., Russell, A., Harris, C., Wilson, M.: The Evolving mSpace Platform: Leveraging the Semantic Web on the Trail of the Memex. In: Proceedings of the 16th Conference on Hypertext, pp. 174-183. ACM, Salzburg, Austria (2005)
22. Schraefel, M-C., Wilson, M., Russell, A., Smith, D-A.: MSPACE: Improving Information Access to Multimedia Domains with Multimodal Exploratory Search. Communications of the ACM 49(4), 47-49 (2006)
23. Sierra, J-L., Fernández-Valmayor, A., Guinea, M., Hernanz, H.: From Research Resources to Learning Objects: Process Model and Virtualization Experiences. Educational Technology & Society 9(3), 56-68 (2006)
24. Sierra, J-L., Fernández-Valmayor, A.: Tagging Learning Objects with Evolving Metadata Schemas. In: Proceedings of the 8th IEEE International Conference on Advanced Learning Technologies, pp. 829-833. IEEE. Santander, Spain (2008)
25. Smith, D-A., Owens, A., Schraefel, M-C., Sinclair, P., Max, P-A., Wilson, A., Rusell, A., Martinez, K., Lewis, P.: Challenges in Supporting Faceted Semantic Browsing of Multimedia Collections. In: Proceedings of the 2nd Int. Conference on Semantics and Digitial Media Technologies, pp. 280-283. Springer, Genoa, Italy (2007)
26. Tunkelang, D.: Faceted Search. Morgan & Claypool Publishers (2009)
27. Uddin, M-N., Janecek, P. The Implementation of Faceted Classification in Web Site Searching and Browsing. Online Information Review 31(2), 218-233 (2007)
28. Way, T., Eklund, P.: Social Tagging for Digital Libraries using Formal Concept Analysis. In: Proceedings of the 17th International Conference on Concept Lattices and their Applications, pp. 139-150. Sevilla, Spain (2010)
29. Wei, B., Liu, J., Zheng, Q.; A Survey of Faceted Search. Journal of Web Engineering 12(1-2), 41-64 (2013)
30. Yee, K-P., Swearingen, K., Li, K., Hearst, M.: Faceted Metadata for Image Search and Browsing. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 401-408. ACM, Fort Lauderdale, Florida, USA (2003)
31. Yitzhak, O-B., Golbandj, N., Har'El N. et al.: Beyond Basic Faceted Search. In: Proceedings of the 2008 International Conference on Web Search and Data Minining, pp. 33-44. ACM, Stanford, CA, USA (2008)
32. Zhang, Z., Li, W., Gurrin, C., Smeaton. A-F.: Faceted Navigation for Browsing Large Video Collection. In: Proceedings of the 22nd International Conference on Multimedia Modelling, pp. 412-417. Springer, Miami, USA (2016)
33. Zobel, J., Moffat, A.: Inverted Files for Text Search Engines. ACM Computing Surveys 33(2), article 6 (2006)

6.5 Multilevel browsing of folksonomy-based digital collections

Cita Completa:

Gayoso-Cabada, J., Rodríguez-Cerezo, D., & Sierra, J.-L. (2016). Multilevel Browsing of Folksonomy-Based Digital Collections. En Proceedings of the 17th Conference on Web Information Systems Engineering Part II, WISE 2016 (pp. 43-51). Lecture Notes in Computer Science 10042, Springer.

Resumen original de la contribución:

This paper describes how to extend the usual one-level tag selection navigation paradigm in folksonomy-based digital collections to a multilevel browsing one, according to which it is possible to incrementally narrow down the set of selected objects in a collection by sequentially adding more and more filtering tags. For this purpose, we present a browsing strategy based on finite automata. As well, we provide some experimental results concerning the application of the approach in Clavy, a system for managing digital collections with reconfigurable structures in digital humanities and educational settings.

Referencias Bibliográficas:

(Chodorow, 2013; Culpepper & Moffat, 2010; du Preez, 2015; Joaquín Gayoso-Cabada et al., 2016a; Godin et al., 1986; Greene, 2015; Hernandez, Falconer, Storey, Carini, & Sim, 2008; Jain et al., 1999; Koutrika, Zadeh, & Garcia-Molina, 2009; Kuznetsov, 2001; Leone, Geel, Müller, & Norrie, 2010; Mathes, 2004; McCandless et al., 2010; Peterson, 2006; Salton & McGill, 1986; Sarmah et al., 2015; José Luis Sierra et al., 2006; Wray & Eklund, 2010; Zobel & Moffat, 2006)

Multilevel Browsing of Folksonomy-Based Digital Collections

Joaquín Gayoso-Cabada, Daniel Rodríguez-Cerezo, and José-Luis Sierra(✉)

Fac. Informática, Universidad Complutense de Madrid, C/Prof. José García Santesmases 9, 28040 Madrid, Spain
{jgayoso,drcerezo,jlsierra}@fdi.ucm.es

Abstract. This paper describes how to extend the usual one-level tag selection navigation paradigm in folksonomy-based digital collections to a *multilevel browsing* one, according to which it is possible to incrementally narrow down the set of selected objects in a collection by sequentially adding more and more filtering tags. For this purpose, we present a browsing strategy based on finite automata. As well, we provide some experimental results concerning the application of the approach in *Clavy*, a system for managing digital collections with reconfigurable structures in digital humanities and educational settings.

Keywords: Multilevel browsing · Folksonomy · Indexing · Navigation automata

1 Introduction

Folksonomies are cataloguing schemes defined and applied collaboratively by communities of users. In this way, users not only apply folksonomies to organize digital resources, but also actively contribute to their creation and maintenance [12]. In this context, accommodating any but the simplest interaction models can become a substantial technical challenge.

An example of a particularly difficult-to-achieve interaction model is general, unconstrained, *multi-level browsing* [5]. In this setting, users sequentially select tags, and, in each stage, the set of objects tagged by all the selected tags is filtered. Even for collections of moderate size, computing these sets of objects can in some cases be too costly to be achieved within acceptable response times. By establishing predefined orders in which tags can be selected and by using these orders to create and maintain navigation trees, response times can be dramatically enhanced, but this rigid and aprioristic organization is contrary to the dynamic and agile nature of folksonomies, where tag sets are continuously changing. In this paper we address this interaction style in its most unconstrained and general form.

The rest of the paper is organized as follows. Section 2 introduces the basis of folksonomy-like organizations of digital collections. Section 3 introduces the multilevel browsing paradigm for this kind of collections and describes how to enable such a browsing style efficiently. Section 4 presents some related work. Finally, Sect. 5 outlines the final conclusions and some lines of future work.

W. Cellary et al. (Eds.): WISE 2016, Part II, LNCS 10042, pp. 43–51, 2016.
DOI: 10.1007/978-3-319-48743-4_4

2 Folksonomy-Based Digital Collections

Collections organized with folksonomies typically comprise the following parts (see Fig. 1 for an example):

- On one hand, there are the *resources* in the collection. For instance, the small collection depicted in Fig. 1 includes six image archives as resources, corresponding to photographs of artistic objects from the Prehistoric and Protohistoric artistic periods in Spain (Fig. 1 actually shows thumbnails of these images).
- On the other hand, there is the *annotation* of the resources. This annotation consists of associating descriptive *tags* with resources. These tags are useful when cataloguing resources and, therefore, they enable future uses of the collection (navigation, search, etc.). For instance, in Fig. 1, resource number 1 has the tags *Cave-Painting*, *Cantabrian* and *Prehistoric* associated.
- Finally, there is a *tag cloud* that groups all the tags that can be used to annotate the resources. Thus, the tag cloud shown in Fig. 1 groups all the tags that annotate resources in the collection. As usual, the size of tags in this cloud represents the presence (number of tagged resources) of the tag in the collection.

Consequently, the internal organization of this kind of collection is very similar in appearance to classic keyword-based systems [15]. However, what distinguishes these collections from classic keyword-based systems is the social and inductive nature in the creation of the cataloguing schemata (i.e., the tag clouds). Indeed, folksonomy-based systems actively involve user communities that add, modify, delete and tag resources, using existing tags or creating new ones as needed. In this way, tag clouds are not explicitly defined nor explicitly maintained, but emerge from the collaborative behavior of communities of practice [12]. While this somewhat uncontrolled and anarchic approach to tagging digital resources can additionally bring up some relevant concerns and critiques from a cataloguing point of view (e.g., existence of synonymous, irrelevant or very generic tags, etc.) [14], the fact is that these systems are extensively used in many scenarios (and especially in computer-mediated social ones) [3]. Therefore, in this paper we will not focus on the critiques and potential shortcomings of the approach, but on efficient ways of enabling sophisticated interaction strategies (multi-level browsing, in particular).

Folksonomy-like systems support a simple one-level browsing strategy in a straightforward way. According to this strategy, it is possible to select one tag in the tag cloud and recover all the resources tagged with said tag. Figure 2(a) illustrates this approach with the small collection from Fig. 1.

One-level browsing can be accomplished efficiently in a straightforward way by using and maintaining an *inverted index* [19], i.e., a data structure that provides a reference to the set of resources tagged by each tag and therefore directly links to the results for each selection. This simple and efficient implementation explains why most folksonomy-based systems include this interaction style as a primary browsing strategy. However, this style prevents more sophisticated exploratory behaviors involving two or more tags simultaneously. In the rest of the paper we will examine how to deal with more than one browsing level.

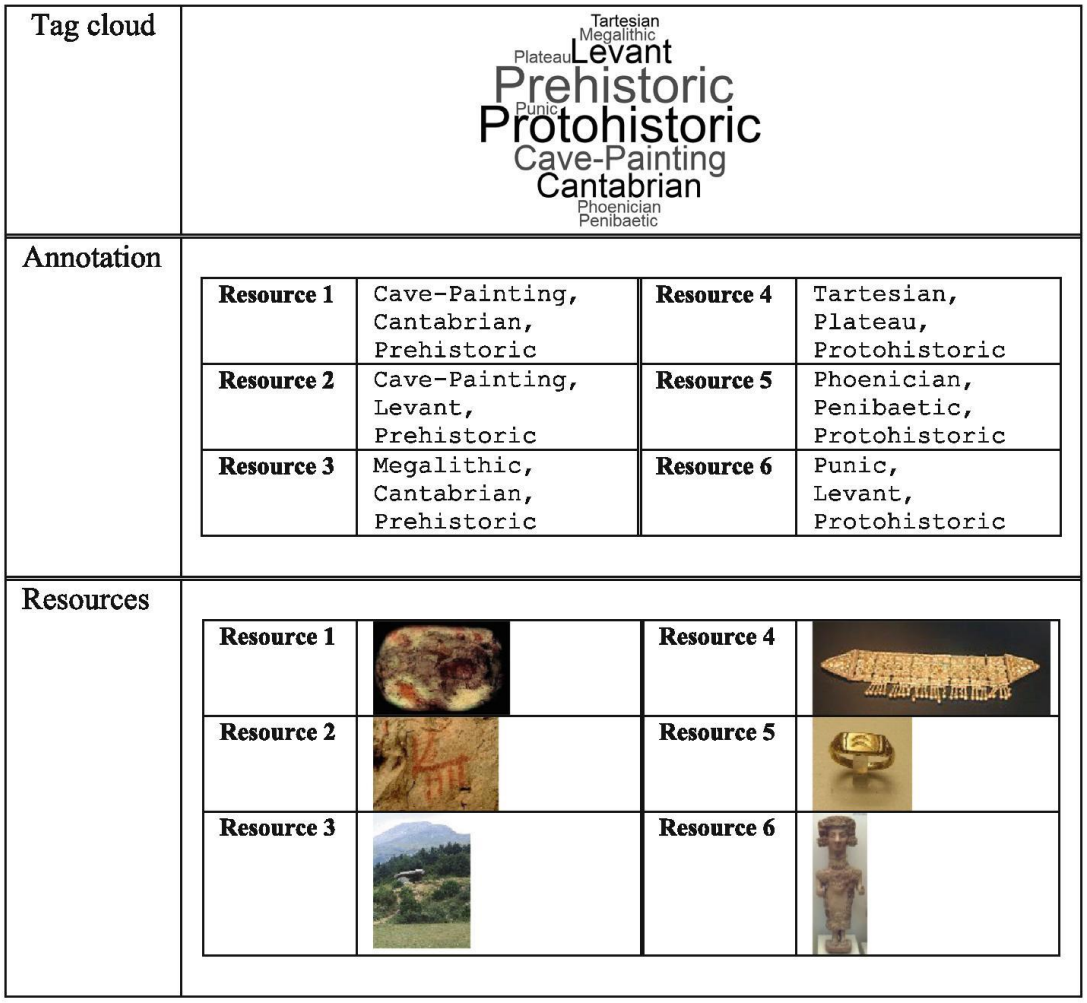


Fig. 1. A small digital collection

3 Multilevel Browsing in Folksonomy-Based Systems

This section addresses the multi-level browsing style in folksonomy-based systems. Subsect. 3.1 introduces the basic interaction behavior. Subsect. 3.2 characterizes this behavior as a finite state machine. Finally, Subsect. 3.3 gives some experimental results.

3.1 The Browsing Model

Conceptually, the extension from one-level to multi-level browsing in folksonomy-like systems is simple. Basically, when a tag is selected, not only is the set of resources narrowed down but also the tag cloud: the resulting tag cloud will be the one *induced* by the set of filtered resources **R**. Such a tag will contain all the tags annotating some resource in **R** with the exception of those tags annotating *all* the resources in **R** (since, in this case, the selection would not refine the set of resources). This makes it possible

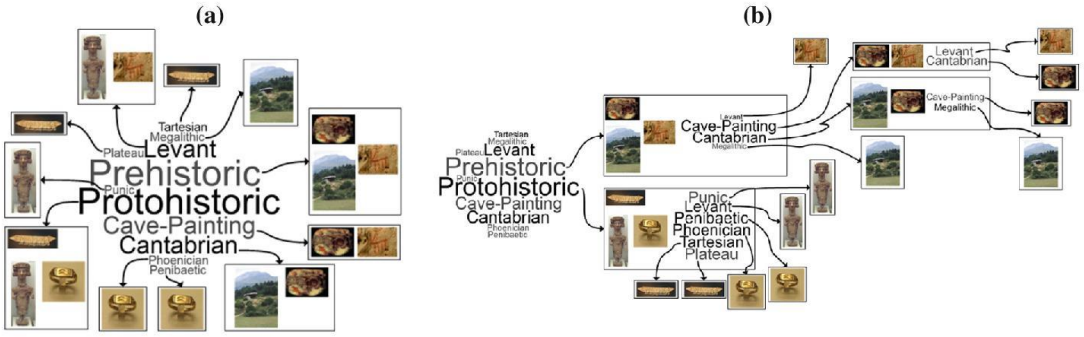


Fig. 2. Examples of (a) one-level browsing; (b) multi-level browsing

to carry out new selections successively on the narrowed tag clouds until a state containing an empty tag cloud is reached. The expected behavior is partially illustrated in Fig. 2(b), which shows the set of resources and the associated tag clouds after some browsing actions on the collection in Fig. 1.

As in the case of one-level browsing, multi-level browsing behavior can also be accomplished by using inverted indexes. However, now an evaluation of a conjunctive query in each interaction state is needed in order to determine the resources to be filtered. Although extensive research has been carried out on how to speed up these operations [2], in some cases the time inverted can negatively impact the user's interactive experience.

3.2 Navigation Automata

In order to accelerate multi-level browsing, it is necessary to have a suitable index structure. Ideally this structure should link to the set of resources selected by each meaningful set of tags t_1, …, t_n, in the same way, an inverted index directly provides the set of resources selected by a tag in the one-level approach. A way of providing such a structure is by using a finite state machine characterizing all the possible interactions and interaction states. This state machine will be called a *navigation automaton*. This automaton will consist of *states* labelled by sets of resources, and *transitions* labelled by tags (as an example, Fig. 3a shows the navigation automaton for the collection of Fig. 1). More precisely:

- There will be an initial state labelled by all the resources in the collection.
- Given a state **S** labelled by a set of resources **R**, for each tag **t** in the tag cloud induced by **R** there will be a state **S'** labelled by all the resources in **R** annotated by **t**, as well as a transition from **S** to **S'** labelled by **t**.

Since the navigation automaton contains all the possible ways of multi-level navigation, it can support multi-level browsing in a straightforward way. Unfortunately, in some cases the number of states in this automaton can grow very quickly (in the worst case, exponentially with respect to the number of resources). The most extreme case, in which the number of states is 2^n-1 (with n the number of resources), arises, for instance, by distinguishing each pair of resource annotations in a single tag. In order to

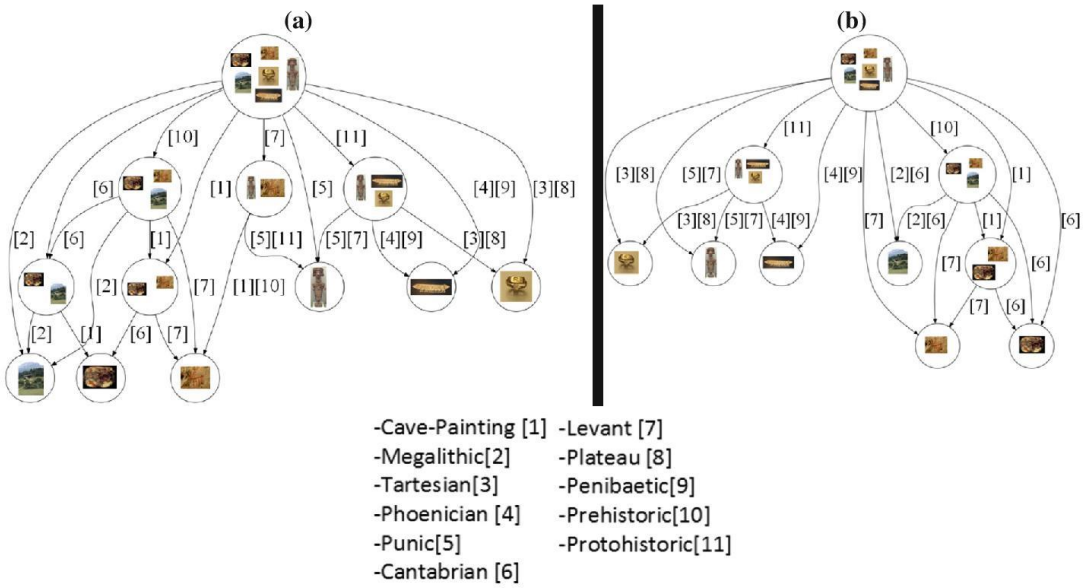


Fig. 3. (a) Navigation automaton for the collection in Fig. 1; (b) A non-deterministic version of the automaton in (a)

avoid this potential exponential factor in the explicit construction of navigation automata, it is possible to maintain non-deterministic versions of these automata, in such a way that only states representing disjoint partitions of their parent states are maintained. Figure 3b shows a feasible non-deterministic automaton equivalent to the one shown in Fig. 3a (it is worthwhile to point out that this solution may not be unique).

3.3 Experimental Evaluation

In order to evaluate our multilevel browsing approach, we have implemented it in *Clavy*, an experimental system for managing digital collections that lets users define organization schemata in a collaborative way.[1] In order to provide some structure to facilitate navigation, *Clavy* makes it possible to group tags in categories that are organized hierarchically. Nevertheless, this hierarchy is not pre-established, but can be edited by *Clavy* users at any time (see Fig. 4). Therefore, backstage, multi-level browsing support in *Clavy* must resort to the basic model described in Sect. 3, since the hierarchy is also subjected to continuous change and evolution. In addition to the automata-based browsing framework described in this paper, we have also implemented an inverted index-based solution in *Clavy*, using Lucene [13], a robust and highly optimized framework for implementing information retrieval applications.

In this context, we set up an experiment consisting of adding the resources in *Chasqui* [17],[2] a digital collection of 6283 digital resources on Pre-Columbian

[1] http://clavy.fdi.ucm.es/Clavy/.

[2] http://oda-fec.org/ucm-chasqui.

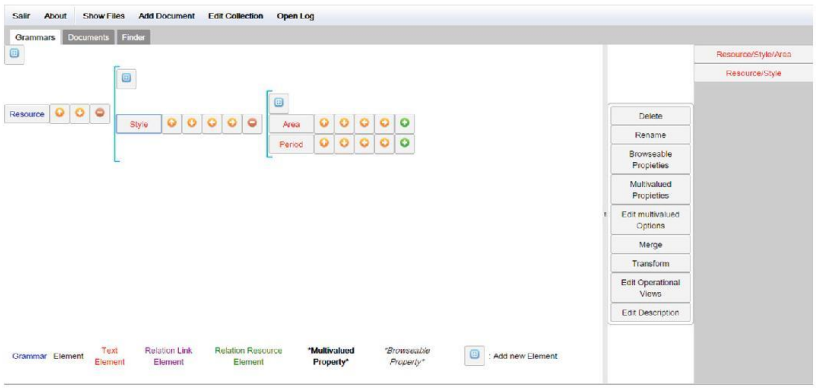


Fig. 4. Editing a hierarchy of tag categories with *Clavy*.

American archeology, to *Clavy* and simulating runs concerning hierarchy reconfiguration and browsing operations.

Each run was customized as follows. We interleaved resource insertion with hierarchy reconfiguration/browsing rounds. Each insertion round consisted of 100 resource insertions (with the exception of the last one, in which all the remaining resources where inserted). In turn, each browsing/reconfiguration round consisted of executing $0.1n$ browsing operations randomly interleaved with $0.01n$ reconfigurations (n being the number of resources inserted so far). Each browsing operation in turn consisted of selecting a feasible tag and computing the next set of active objects, or of establishing the initial state as the active one in case of unavailability of feasible tags; once the next interaction state was determined, all the filtered resources were visited. In both the cases of inverted indexes and automata, in-memory indexes were used in order to avoid the side effects of persistence that might disturb the experiment.

Figure 5 shows the results obtained from the two runs. The experiment was run on a PC with Windows 10, with a 3.4 GHz Intel microprocessor, and with 8 Gb of DDR3 RAM. The horizontal axis corresponds to the number of operations carried out so far. The vertical axis corresponds to cumulative time (in seconds). As is made apparent, the automata-based approach clearly outperforms inverted indexes (regardless of the fact that we are using a highly optimized framework, like Lucene, for inverted indexing vs. our own in-house experimental implementation for navigation automata).

4 Related Work

There are several systems that, like our proposal, implement several sorts of multi-level browsing onto folksonomy-based systems. Systems like the one described in [7, 9] are supported by inverted index approaches. Other systems, like that described in [11], are supported by extensible data adapters that interface between synchronized tag clouds and underlying database management systems. Instead of relying on inverted indexes and/or conventional database layers, our approach starts by characterizing the intrinsic behavior of multi-level browsing onto a folksonomy-like system in terms of navigation

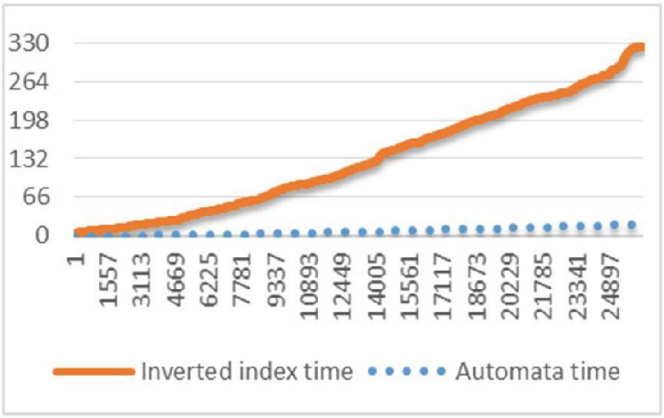


Fig. 5. Cumulative time of inverted indexes vs. automata

automata, and then tries to approximate this model with a non-deterministic version that provides reasonable time and space tradeoffs. In [4] we propose a representation of these non-deterministic automata inspired by *dendrograms* such as those used in hierarchical clustering settings [8].

Our navigation automata model is actually similar to lattice-based proposals to browse information spaces, as described in the seminal work of [5]. This organization is actually the main subject of the fertile theory of *formal concept analysis* [16]. Similarly, there are several proposals on using lattices as the underlying indexing structures for enabling multi-level browsing [6, 18]. However, all these approaches are limited by the intrinsic complexity of formal concept analysis [10]. This is why we have proposed a simpler but still practical approximation based on non-deterministic versions of navigation automata.

5 Conclusions and Future Work

Folksonomy-based digital collections are living entities in which not only digital resources, but also organization schemata, are subject to continuous change and evolution. This changing and evolving nature makes the accomplishment of sophisticated interaction paradigms particularly challenging. In this paper we have addressed the efficient inclusion of multilevel browsing strategies in these settings, in which sets of selected resources can be successively refined through the selection of sequences of tags. For this purpose we have modeled this behavior as a finite state machine, the *navigation automaton*, taking into account all the possible ways of navigating the collection by using tags. Unfortunately, we have also showed how, in some cases, the number of states in this automaton can increase exponentially with respect to the collection's size. In order to address this potential exponential factor we have proposed using non-deterministic versions of these automata. Some experiments with a real collection gave us evidence on how the automata-based technique can outperform more conventional and widely used ones, like those based on inverted indexes.

We are currently working on further optimizing our navigation automata representation. We are also looking for efficient ways to make all this information persistent, either by using standard relational databases or alternative NoSQL approaches. Finally, we also hope to include support for arbitrary Boolean queries and for alternative ways of exploring the resources selected.

Acknowledgements. This work has been supported by the BBVA Foundation (grant HUM14_251) and Spanish Ministry of Economy and Competitiveness (grant TIN2014-52010-R)

References

1. Chodorow, K.: MongoDB: The Definitive Guide. O'Reilly, Sebastopol (2013)
2. Culpepper, J-S., Moffat, A.: Efficient set intersection for inverted indexing. ACM Trans. Inf. Syst. **29**(1) (2010)
3. du Preez, M.: Taxonomies, folksonomies, ontologies: what are they and how do they support information retrieval? Indexer **33**(1), 29–37 (2015)
4. Gayoso-Cabada, J., Rodríguez-Cerezo, D., Sierra, J-L.: Browsing digital collections with reconfigurable faceted thesauri. In: 25th International Conference on Information Systems Development (ISD), Katowize, Poland (2016)
5. Godin, R., Saunders, G.: Lattice model of browsable data space. Inf. Sci. **40**(2), 89–116 (1986)
6. Greene, G-J.: A generic framework for concept-based exploration of semi-structured software engineering data. In: Proceedings of the 30th IEEE/ACM International Conference on Automated Software Engineering, pp. 894–897 (2015)
7. Hernandez, M-E., Falconer, S-M., Storey, M-A., Carini, S., Sim, I.: Synchronized tag clouds for exploring semi-structured clinical trial data. In: Proceedings of the 2008 Conference of the Center for Advanced Studies on Collaborative Research: Meeting of Minds (CASCON), article 4 (2008)
8. Jain, A.-K., Murty, M.-N., Flynn, P.-J.: Data clustering: a review. ACM Comput. Surv. **31** (3), 264–323 (1999)
9. Koutrika, G., Zadeh, Z-M., Garcia-Molina, H.: CourseCloud: summarizing and refining keyword searches over structured data. In: Proceedings of the 12th International Conference on Extending Database Technology (EDBT), pp. 1132–1135 (2009)
10. Kuznetsov, S.: On computing the size of a lattice and related decision problems. Order **18**(4), 313–321 (2001)
11. Leone, S., Geel, M., Müller, C., Norrie, M.C.: Exploiting tag clouds for database browsing and querying. In: Proper, E., Soffer, P. (eds.) CAiSE Forum 2010. LNBIP, vol. 72, pp. 15–28. Springer, Heidelberg (2011)
12. Mathes, A.: Folksonomies – cooperative classification and communication through shared metadata. Comput. Mediat. Commun. – LIS590CMC **47**(10), 1–13 (2004)
13. McCandless, M., Hatcher, E., Gospodnetic, O.: Lucene in Action, 2nd edn. Manning Publications, Greenwich (2010)
14. Peterson, E.: Beneath the metadata: some philosophical problems with folksonomy. D-Lib Mag. **12**(11) (2006)
15. Salton, G., McGill, M.J.: Introduction to Modern Information Retrieval. McGraw-Hill, Maidenherd (1986)

16. Sarmah, A.-K., Hazarika, S.-M., Sinha, S.-K.: Formal concept analysis: current trends and directions. Artif. Intell. Rev. **44**(1), 47–86 (2015)
17. Sierra, J.-L., Fernández-Valmayor, A., Guinea, M., Hernanz, H.: From research resources to learning objects: process model and virtualization experiences. Educ. Technol. Soc. **9**(3), 56–68 (2006)
18. Way, T., Eklund, P.: Social tagging for digital libraries using formal concept analysis. In: Proceedings of the 17th International Conference on Concept Lattices and their Applications (CLA) (2010)
19. Zobel, J., Moffat, A.: Inverted files for text search engines. ACM Comput. Surv. **33**(2) (2006). Article 6

6.6 Learning object repositories with dynamically reconfigurable metadata schemata

Cita Completa:

Gayoso-Cabada, J., Rodríguez-Cerezo, D., & Sierra, J.-L. (2016). Learning object repositories with dynamically reconfigurable metadata schemata. En Proceedings of the XVIII International Symposium on Computers in Education, SIIE 2016 (6 pags). IEEE Computer Society.

Resumen original de la contribución:

In this paper we describe a model for learning object repositories in which users have full control over metadata schemata. Thus, they can define new schemata and reconfigure existing ones in a collaborative fashion. In consequence, the repository must react to changes in schemata in a dynamic and responsive way. Since schemata enable operations like navigation and search, dynamic reconfigurability requires clever indexing strategies, resistant to changes in these schemata. For this purpose, we have used conventional inverted indexing approaches and have also devised a hierarchical clustering-based indexing model. By using Clavy, a system for managing learning object repositories in the field of the Humanities, we provide some experimental results that show how the hierarchical clustering-based model can outperform the more conventional inverted index-based solutions.

Referencias Bibliográficas:

(Ben-Yitzhak et al., 2008; Berchtold et al., 2000; Coombs, Renear, & DeRose, 1987; Culpepper & Moffat, 2010; Godin et al., 1986; Gonzalez-Barbone & Anido-Rifon, 2008, 2010; Greene, 2015; Hildebrand et al., 2006; Huang et al., 2014; «IEEE SA - 1484.12.1-2002 - IEEE Standard for Learning Object Metadata», s. f.; Jain et al., 1999; Kuznetsov, 2001; R. Li et al., 2007; McCandless et al., 2010; Polsani, 2006; Radelaar et al., 2014; Sarasa, Canabal, & Sacristán, 2008; Sarmah et al., 2015; schraefel et al., 2006; José Luis Sierra et al., 2006; Wray & Eklund, 2010; Zobel & Moffat, 2006)

Learning Object Repositories with Dynamically Reconfigurable Metadata Schemata

Joaquín Gayoso-Cabada Daniel Rodríguez-Cerezo José-Luis Sierra
Fac. Informática
Universidad Complutense de Madrid
Spain
{jgayoso,drcerezo,jlsierra}@fdi.ucm.es

***Abstract*—In this paper we describe a model for learning object repositories in which users have full control over metadata schemata. Thus, they can define new schemata and reconfigure existing ones in a collaborative fashion. In consequence, the repository must react to changes in schemata in a dynamic and responsive way. Since schemata enable operations like navigation and search, dynamic reconfigurability requires clever indexing strategies, resistant to changes in these schemata. For this purpose, we have used conventional inverted indexing approaches and have also devised a hierarchical clustering-based indexing model. By using *Clavy*, a system for managing learning object repositories in the field of the Humanities, we provide some experimental results that show how the hierarchical clustering-based model can outperform the more conventional inverted index-based solutions.**

Keywords—learning object repository, metadata schemata, dynamic reconfigurability, learning object indexing, browsing

I. INTRODUCTION

The dominant trend in the production of Learning Object (LO) repositories [15] follows a *top-down* approach, based on the heavy use of standards and recommendations (e.g., metadata standards like LOM [10], packaging proposals like IMS CP [21], SCORM [5] or IMS Common Cartridge [6], and interoperability proposals like IMS DRI[1] or OAI-PMH[2]).These standardization efforts make possible, for instance, the federation and interoperability of LO repositories in distributed networks (AGREGA [17] being a well-known example in the context of Spain).

However, the top-down approach is not particularly oriented to facilitating the inductive creation of domain-specific metadata schemata (i.e., the schemata that shape how LOs are described). This is a critical aspect in learning settings like the Humanities, in which metadata schemata must be frequently created, revised and modified in parallel to the creation of the repositories [20].

In order to facilitate the inductive construction and refinement of metadata schemata, in this paper we describe how to support a more *bottom-up* approach, according to which communities of users (e.g., instructors, researchers and students) collaborate in the construction of these schemata in addition to using them to describe learning materials. This collaboration involves not only defining new schemata and/or using existing ones, but also reconfiguring these schemata. In consequence, the repository must react to the changes in schemata accordingly. In addition, since schemata are typically reconfigured with experimental and/or exploratory purposes in mind, it is necessary to ensure that users don't need to wait for long periods until the schemata reconfigurations are reflected in the repository; on the contrary, ideally, they should be able to visualize the reconfiguration's effects immediately after changing the schemata. From a system architecture perspective, this is a particularly demanding requirement, since reconfigurations in schemata can affect the way in which the repository is browsed and / or searched. Thus, in this paper we introduce indexing strategies capable of complying with the harsh requirements posed by dynamic reconfigurability.

The rest of the paper is organized as follows. Section II introduces our model of repository with dynamically reconfigurable metadata schemata. Section III analyzes dynamic reconfigurability in these repositories. Section IV proposes some indexing approaches to enable dynamic reconfigurability and provides some comparative results. Section V analyzes some related works. Finally, section VI outlines the final conclusions and some lines of future work.

II. THE REPOSITORY MODEL

This section introduces our model of repository with dynamically reconfigurable metadata schemata. Subsection II.A describes the repository's structure, and subsections II.B, II.C, II.D and II.E their different parts (resources, metadata schemata, LOs, and navigation maps).

A. Structure of the repository

According to our model, repositories comprise the following parts:

- A set of *resources*. These resources are the atomic digital assets that integrate the LOs.
- A set of *metadata schemata*. These schemata characterize how to describe the types of objects that can integrate the repository.
- A set of LOs. These LOs aggregate resources and simpler LOs in educationally-meaningful clusters.
- A *navigation map*. This map makes it possible to navigate the repository by using the structures imposed on LOs by metadata schemata.

Fig. 1 sketches an example of repository structured according to our model (it is a repository concerning artistic

[1] www.imsglobal.org/digitalrepositories
[2] www.openarchives.org/OAI/openarchivesprotocol.html

978-1-5090-4596-9/16/$31.00 ©2016 IEEE

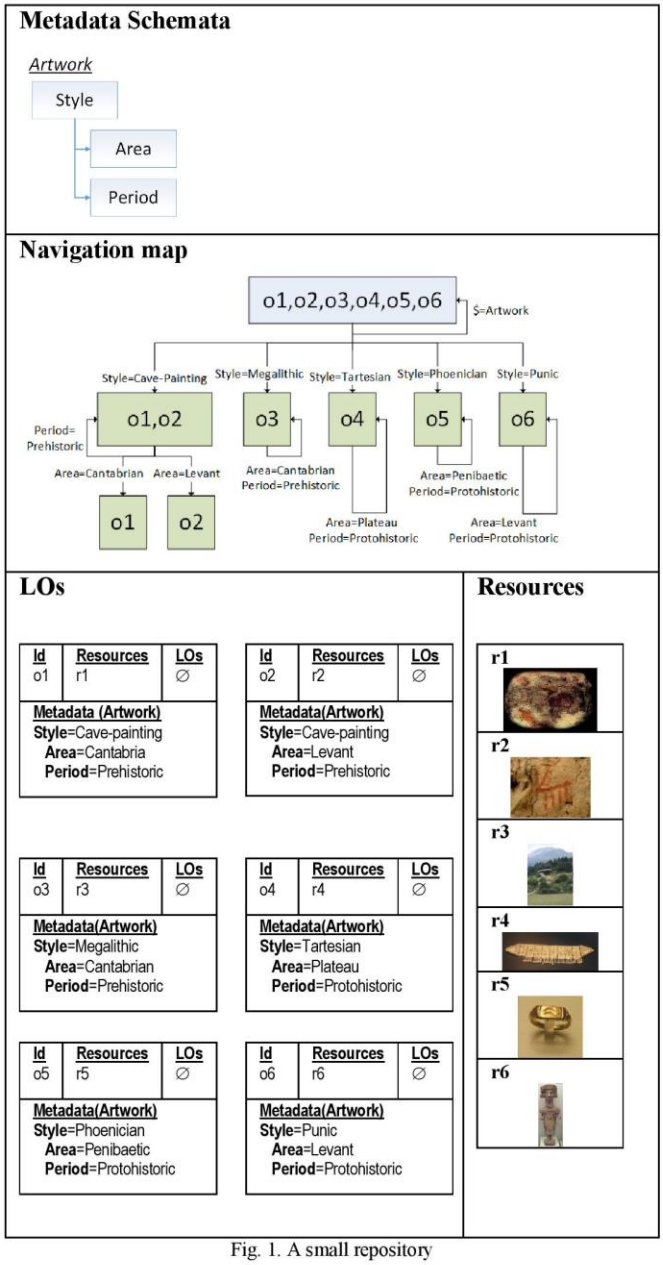


Fig. 1. A small repository

objects from the Prehistoric and Protohistoric artistic periods in Spain).

B. Resources

Resources in our model can be any digital entity with educational value. Therefore, resources can be archives of different types (images, sound or video archives, electronic documents, e-books, etc.), external resources identified by a URL, or even entities of more abstract nature (tuples of a table in a relational database, records in a bibliographical catalog, elements in an XML document, rows in a spreadsheet, etc). Each resource has a unique identifier associated, which is useful when referring to the resource from LOs.

For instance, the repository in Fig. 1 includes six image archives as resources, corresponding to photographs of different artistic objects (Fig. 1 actually shows thumbnails of these images).

C. Metadata Schemata

Metadata schemata are the cornerstone of the repositories. In our proposal, users can freely create new schemata and edit existing ones[3]. In this way, it is necessary to adopt a schemata model both general and *agnostic* enough to accommodate a great variety of users' expressive needs. For this purpose, our model is inspired by generalized markup languages (e.g., SGML or XML) [2]. In this way, each schema, in addition to having a unique name, is a hierarchical arrangement of *elements*. Each element is characterized by a descriptive name, and can be of one of the following two types:

- *Description element*. These elements introduce descriptive values.
- *Structural element*. These elements do not introduce values, but they are useful in creating intermediate structures.

Thus, by providing suitable hierarchies of structural and description elements, it is possible to mimic the description capabilities of common metadata schemata (e.g., LOM).

For instance, the repository of Fig. 1 includes one single schema, named *artwork*, oriented to providing a simplified description of an artistic object in terms of its artistic *style*, and, within this cultural style, in terms of the geographical *area* and the cultural *period*.

D. Learning Objects

Concerning LOs, they comprise the following parts:

- A (possibly empty) set of references to resources (references are made by id).
- A (possibly empty) set of references to other LOs.
- A *metadata document*. This is a tree-like structure conforming one metadata schema. For this purpose, suitable values are assigned to the description elements (this assignment does not need to be complete: by default, values will be initialized to $\perp$).

The repository in Fig. 1 includes one LO for each resource included in the repository (notice, however, that this one-to-one correspondence between resources and LOs cannot necessarily be extrapolated to other repositories). For each LO there is a metadata document indicating the artistic style, geographical area and cultural period associated to the LO.

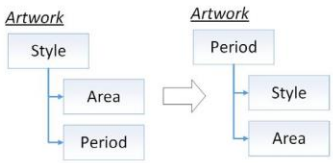
E. Navigation map

Finally, the navigation map is a directed graph in which:

- Nodes represent sets of LOs, and arcs are labelled with *element-value* pairs used to narrow down the LOs: an

[3] In concrete implementations it is possible to restrict editions to privileged users (e.g., instructors), as well as to introduce a more complex permission system.

(a)



(b)

o1 **Metadata (Artwork)** **Period**=Prehistoric **Style**=Cave-painting **Area**=Cantabrian	**o2** **Metadata (Artwork)** **Period**=Prehistoric **Style**=Cave-painting **Area**=Levant	**o3** **Metadata (Artwork)** **Period**=Prehistoric **Style**=Megalithic **Area**=Cantabrian
o4 **Metadata (Artwork)** **Period**=Protohistoric **Style**=Tartesian **Area**=Plateau	**o5** **Metadata (Artwork)** **Period**=Protohistoric **Style**=Phoenician **Area**=Penibaetic	**o6** **Metadata (Artwork)** **Period**=Protohistoric **Style**=Punic **Area**=Levant

(c)

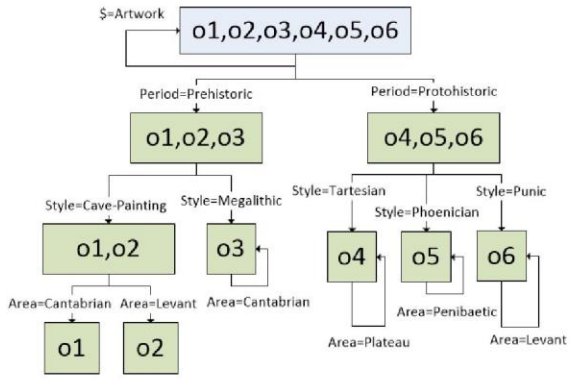


Fig. 2. (a) A reconfiguration of the schema of the repository in Fig. 1; (b) Effect of the reconfiguration in the metadata documents; (c) Effect in the navigation map

arc's target node will contain only those LOs exhibiting the *element – value* pair in the source node.

– The structure of the map is constrained by the schemata hierarchies. In this way, nodes can only be narrowed down with *element – value* pairs comprising child elements of elements present in incoming arcs.
– There is also a root node, which represents the overall set of LOs. It can be narrowed down by a special element $, whose values are the different schemata names, and whose child elements are the schemata root elements.

Fig. 1 also shows a navigation map for the repository. Notice how each path in this map is constrained by the schemata structure (in this way browsing starts by selecting a value for the artistic style, and then continues by selecting a value either for the geographical area or for the artistic period).

III. RECONFIGURABILITY

In this section we address the concern of dynamically reconfiguring the metadata schemata of a repository. Subsection III.A analyzes how this reconfiguration is carried out and its effects in the different parts of the repository. Subsection III.B describes how to avoid such effects in LO representation. Subsection III.C describes, in turn, how to deal with navigation.

A. Reconfigurable Metadata Schemata

Our model lets users reconfigure metadata schemata by rearranging the hierarchical organization of elements. For instance, Fig. 2a shows an example concerning the repository in Fig. 1, which prioritizes the artistic period as the primary classification focus instead of the cultural style (as in the example in Fig. 1).

Since the organization of a repository ultimately relies on its schemata, by reconfiguring these schemata the repository's overall structure is also reconfigured. More precisely:

– The metadata documents for each LO must be changed to reflect the new hierarchical organization of elements. As an example, this effect is made apparent in Fig. 2b.
– The navigation map is also deeply affected by the reconfiguration. For instance, Fig. 2c shows how, after reconfiguring the schema of the repository in Fig. 1, the navigation map is also altered to reflect the change in focus represented by the reconfiguration (entering by *period* and refining by *style* or by *area* instead of entering by *style* and refining by *period* or by *area*).

B. Reconfigurable metadata documents

In order to address the effect of schemata reconfigurations on metadata documents, it is necessary to find document representations resistant to reorganizations of element hierarchies. Fortunately, since all the metadata documents conforming a particular schema share a common structure (indeed, that represented by the schema), the solution in this case is easy: documents can be represented as tables by assigning values to elements in the schemata instead of the whole hierarchical structure. Fig. 3a exemplifies this representation for the repository in Fig. 1. Notice that these tables remain invariable whatever the reorganizations carried out in the element hierarchies may be. In addition, the additional cost incurred by the representation is negligible: one indirection level. Indeed, structure recovery is a simple matter of traversing the corresponding metadata schema and of querying the table for each traversed element.

C. Reconfigurable navigation maps

The reconfiguration of the navigation map is a substantially more convoluted matter. Indeed, as Fig. 2 makes apparent, a simple reconfiguration in a metadata schema may involve a complete reconfiguration of the underlying navigation map. Therefore, it is necessary to look for alternatives to the explicit representation of such a map.

(a)

o1 Metadata(Artwork)

Style	Cave-Painting
Period	Prehistoric
Area	Cantabrian

o2 Metadata(Artwork)

Style	Cave-Painting
Period	Prehistoric
Area	Levant

o3 Metadata(Artwork)

Style	Megalithic
Period	Prehistoric
Area	Cantabrian

o4 Metadata(Artwork)

Style	Tartesian
Period	Protohistoric
Area	Plateau

o5 Metadata(Artwork)

Style	Phoenician
Period	Protohistoric
Area	Penibaetic

o6 Metadata(Artwork)

Style	Punic
Period	Protohistoric
Area	Levant

(b)

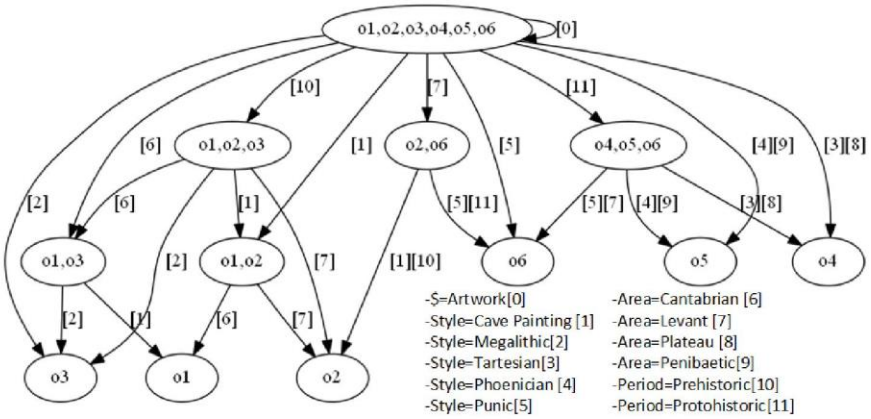


Fig. 3. (a) Tabular representations of the metadata documents in the repository in Fig. 1; (b) navigation automaton for the repository in Fig. 1

Ideally, it would be convenient to provide a structure capable of representing *all* the possible navigations induced by *all* the possible reconfigurations of the schemata in a compact and unified way. For this purpose, it is necessary to free *element-value* pairs from the hierarchical organizations induced by these schemata. Therefore, a plain set of *element-value* pairs must be considered and, in each interaction state in the navigation process, the applicability of all the meaningful selections must be envisioned. The result can be represented as a finite state machine, which we will call a *navigation automaton*. This automaton will consist of *states* labelled by sets of LOs, and *transitions* labelled by element-value pairs.

More precisely:

- There will be an initial state labelled by all the LOs in the repository.
- Given a state S labelled by a set of LOs $\boldsymbol{O}$, for each element-value pair e=v in the metadata document of some LO in $\boldsymbol{O}$ there will be a state S' labelled by *all* the LOs in $\boldsymbol{O}$ with e=v in their metadata documents, as well as a transition from S to S' labelled by e=v.[4]

Fig. 3b shows the navigation automaton for the repository in Fig. 1. Notice that the navigation automaton does not depend on the hierarchical organization of elements in the schemata, but only on the element-value pairs in the metadata documents. Therefore, it is not affected by reconfigurations in the schemata.

Unfortunately, although the explicit availability of the navigation automaton provides an efficient and elegant solution to navigation in the presence of reconfigurable schemata, in some cases the number of states in this automaton can grow very fast (in the worst case, exponentially with respect to the repository's size). This fact can be confirmed by identifying states in navigation automata with *formal concepts* in *concept lattices* (such as these are understood in *formal concept analysis* [18])[5]. The most extreme case, in which the number of states is 2^n-1 (with n the number of LOs), arises, for instance, by distinguishing each pair of metadata documents in a single *element-value* pair[6].

This worst-case exponential grow ratio conforms a theoretical barrier that can hinder the explicit representation of the navigation automaton, especially in live and open scenarios such as those faced by a general-purpose LO repository. Therefore, it may be recommendable to look for alternative indexing approaches.

IV. INDEXING APPROACHES

This section introduces two indexing approaches to enable the dynamic recreation of navigation automata: *inverted indexes* (subsection IV.A) and *navigation dendrograms* (subsection IV.B). Subsection IV.C provides some experimental results comparing the two approaches.

A. Inverted indexes

Inverted indexes are standard artifacts used for information retrieval [24]. Basically, for each element-value pair, an inverted index associates the set of LOs by including such a pair in its metadata document. Fig. 4a shows an example of inverted index for the repository in Fig. 1.

Notice that this kind of inverted index can be used to determine the set of objects selected in each navigation path

[4] Notice that S and S' can be the same -when all the LOs in $\boldsymbol{O}$ have e=v in their metadata documents.

[5] Indeed, navigation automata can actually be thought of as an explicit representation of concept lattices. As indicated in [12], the problem of determining the size of concept lattices is proved to be a #P-complete one (i.e., harder than NP-complete). Thus, the exponential factor underlying the intrinsic complexity of the problem can hinder the direct applicability of the technique on repositories of moderate or large sizes.

[6] This construction is actually suggested by the proof of theorem 1 in [12]

by intersecting the sets associated with the element-value pairs traversed. The cost of evaluating the intersection operations cited constitutes the main shortcoming of the approach. While there has been extensive research in performing these intersection operations efficiently [3], the cost is not negligible. On the positive side is the availability of many mature implementations and frameworks that can be used in a straightforward way to support the technique. For instance, in our experiences, we used Lucene [14] for such a purpose.

(a)

Pair/Sid	RESOURCE
[0]	o1,o2,o3,o4,o5,o6
[1]	o1,o2
[2]	o3
[3]	o4
[4]	o5
[5]	o6
[6]	o1,o3
[7]	o2,o6
[8]	o4
[9]	o5
[10]	o1,o2,o3
[11]	o4,o5,o6

(b)

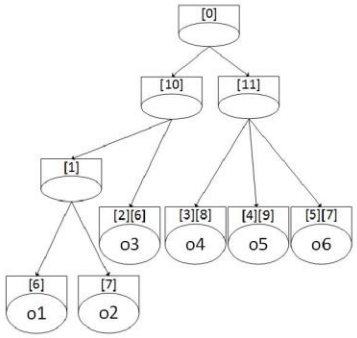


Fig. 4. (a) An inverted index for the repository in Fig. 1; (b) A navigation dendrogram (References [0],[1],etc. are defined in Fig. 3b)

B. Navigation dendrograms

In order to avoid the proliferation of intersection operations, which is characteristic of inverted index representations, we have designed a tree-shaped indexing scheme inspired by *dendrograms* in hierarchical clustering [11]. The resulting structures are called *navigation dendrograms*.

Nodes in a navigation dendrogram represent subsets of the overall LO set. The LO set associated to a node is not explicitly stored in this node. Instead, each LO is only hosted in one node (the LO's *host node*). LOs placed in a node are called the mentioned node's *own* LOs. The overall LO set of a node is given by its own LOs and by all the own LOs of its descendants. Finally, in order to partition the LO space, each node has a set of *filtering* element-value pairs associated, so that all the own LOs in the node and in all their descendants' must include these filtering pairs in their metadata documents.

Navigation dendrograms can be built to contain as many as $2K$ nodes (K being the number of LOs in the repository). In addition, navigation can be articulated by maintaining a set of the dendrogram's nodes. Then, when an element-value pair is selected, this set is refined as follows:

- Nodes containing the selected pair in their filtering sets or having an ancestor meeting such a condition are preserved.
- Nodes having any descendant containing the selected pair in their filtering set are replaced by all the descendants meeting such a condition.
- Any other node is discarded.

By maintaining all the information required to carry out this refinement in the nodes (i.e., filtering pairs of node ancestors, and references to descendants per filtering pairs) this process can be carried out very efficiently. Indeed, the resulting structure is a non-deterministic version of the navigation automaton that explicitly avoids the aforementioned potential exponential factor.

Fig. 4b shows an example of a navigation dendrogram for the repository in Fig. 1.

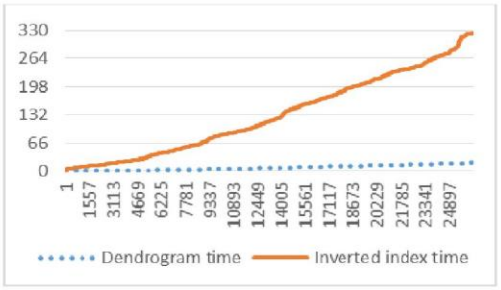


Fig. 5. Accumulated time of inverted indexes vs. dendrograms

C. Experimental evaluation

In order to compare the two approaches described, we implemented both on *Clavy*, an experimental system for managing LO repositories with reconfigurable metadata schemata.[7]

We also set up an experiment consisting of adding the LOs in *Chasqui* [20], a repository of 6283 LOs on Precolombian American archeology, to *Clavy* and simulating runs concerning navigation and schemata reconfiguration operations. Each run interleaved 100 LO insertions with $0.1n$ navigation operations randomly interleaved with $0.01n$ reconfigurations (n being the number of LOs inserted so far). Each navigation operation in turn consisted of selecting a feasible element-value pair, computing the next interaction state, and visiting all the LOs filtered. Reconfiguration operations, then, consisted of feasible interchanges of two randomly selected elements,[8] followed by a navigation step. Inverted indexes were managed using Lucene, while navigation dendrograms were managed using our own implementation. In both cases, in-memory indexes were used to avoid the side effects of persistence which would disturb the experiment.

Fig. 5 shows the results obtained from the two runs (an experiment run on a PC with Windows 10, with a 3.4GHz Intel microprocessor, and with 8Gb of DDR3 RAM). The horizontal axis corresponds to the number of operations carried out so far. The vertical axis corresponds to accumulated time (in seconds). As is made apparent, the dendrogram-based approach clearly outperforms the inverted indexes (even though we are using a highly optimized framework, like Lucene, for inverted indexing vs. our own in-house experimental implementation for dendrograms).

[7] http://clavy.fdi.ucm.es

[8] By *feasible* we mean avoiding cycles in the resulting schema.

V. RELATED WORK

Our proposal is similar to browsing systems for browsing information spaces that, like ours, envision the possibility that the user can reconfigure the underlying metadata schemata (e.g., [8][19]). However, these systems are typically supported by general-purpose semantic web or relational database solutions instead of by model-specific indexing approaches.

A seminal work on using concept lattices to organize and navigate information spaces is [4]. Some recent systems using concept lattices as their underlying indexing structure are [7][22]. However, all these approaches face the theoretical limit imposed by the intrinsic complexity of formal concept analysis. This is why we proposed a simpler but still practical approximation based on navigation dendrograms.

Inverted indexes have been extensively used to support hierarchical navigation (e.g., guided by faceted thesauri). Works like [23] describe efficient approaches to enable this navigation. However, all these approaches are based on the assumption of pre-established and immutable schemata. As pointed out in [1], if this assumption is left out, inverted indexes can become costly due to the set operations involved.

Finally, it is worthwhile to notice that clustering techniques have been extensively used in open metadata schemata (i.e., folksonomy-like systems) to enable the discovery of useful semantic relationships among terms in order to provide better guidance to users (e.g., [9][13][16]). Thus, clustering in these approaches is oriented to enhancing users' navigation efficiency, while our navigation dendrograms are oriented to improving the internal efficiency of the supporting software.

VI. CONCLUSIONS AND FUTURE WORK

In this paper we have addressed the problem of dynamic reconfigurability in LO repositories. Since metadata schemata can be rearranged in unexpected ways, it is necessary to use internal representation mechanisms resistant to these changes. In the case of metadata documents we have shown how a tabular representation of the assignment of values to elements in the schemata suffices. However, dealing with the navigation system is substantially more cumbersome. We have shown how a concept lattice-like representation (which we have called a *navigation automaton*) can elegantly address this concern. However, this representation exhibits a potential exponential factor that, at least in theory, hinders its applicability (especially in live and open settings, in which schemata evolution cannot be envisioned beforehand). For this purpose, we have proposed alternative indexing approaches (one based on inverted indexes, and another one based on dendrograms). We have also provided some evidence of how dendrograms can outperform inverted indexes.

We are currently working on optimizing and making our representations persistent. In addition, we want to further study the practical grow ratio of the navigation automaton in real-world scenarios, to support arbitrary Boolean queries, and to run more empirical evaluations.

VII. ACKNOWLEDGEMENTS

This work has been supported by the BBVA Foundation (grant HUM14_251) and Spanish Ministry of Economy and Competitiveness (grant TIN2014-52010-R)

VIII. REFERENCES

[1] Berchtold, S., Böhm, C., Keim, D-A., Kriegel, H-P., Xiaowei, X.: Optimal Multidimensional Query Processing Using Tree Striping. DaWaK'00, 244-257. 2000

[2] Coombs, J. H., Renear, A. H., DeRose, S. J. Markup Systems and the Future of Scholarly Text Processing. Communications of the ACM, 30 (11), 933-947. 1987

[3] Culpepper, J-S., Moffat, A.: Efficient Set Intersection for Inverted Indexing. ACM Transactions on Inf. Systems 29(1), article 1 (2010)

[4] Godin, R., Saunders, G. Lattice Model of Browsable Data Space. Information Sciences 40(2), 89-116. 1986

[5] González-Barbone, V., Anido-Rifón, L.E. Creating the first SCORM object. Computers & Education 51(4): 1634-1647. 2008

[6] Gonzalez-Barbone, V., Anido-Rifón, L.E. From SCORM to Common Cartridge: A step forward. Computers & Education 54(1): 88-102. 2010

[7] Greene, G-J. A Generic Framework for Concept-Based Exploration of Semi-Structured Software Engineering Data. ASE'15, 894-897. 2015

[8] Hildebrand, M., Ossenbruggen, J-v., Hardman, L.: /facet: A Browser for Heterogeneous Semantic Web Repositories. WWW'06, 272-285. 2006

[9] Huang, J-W., Chen, K-Y., Chen, Y-C., Yang, K-N., Hwang, S., Huang, W-C. A Novel Spatial Tag Cloud Using Multi-Level Clustering. Journal of Information Science and Engineering 30, 687-700. 2014

[10] IEEE Standard 1484.12.1-2002. 2002. IEEE Standard for Learning Object Metadata

[11] Jain, A-K., Murty, M-N., Flynn, P-J.: Data Clustering: a Review. ACM Computing Surveys 31(3), 264-323. 1999

[12] Kuznetsov, S. On computing the size of a lattice and related decision problems. Order 18(4), 313–321. 2001

[13] Li, R., Shenghua, B., Fei, B., Su, Z., Yu, Y. Towards Effective Browsing of Large Scale Social Annotations. WWW'07, pp. 943-952. 2007

[14] McCandless, M., Hatcher, E., Gospodnetic, O.: Lucene in Action, 2nd Edition. Manning Publications. 2010

[15] Polsani, P. Use and Abuse of Reusable Learning Objects. JODI 3(4).2003

[16] Radelaar, J., Boor, A-J., Vandic, D., van Dam, J-W., Fasinca, F. Improving search and exploration in tag spaces using automated tag clustering. Journal of Web Engineering 13(3-4), 277-301. 2014

[17] Sarasa-Cabezuelo, A., Canabal-Barreiro, J-M., Sacristán-Heras, J-C. Agrega - Spanish Education Community Federation of Repositories Of Learning Objects. eLearning 2008: 47-50

[18] Sarmah, A-K., Hazarika, S-M., Sinha, S-K.: Formal Concept Analysis: Current Trends and Directions. Art. Int. Review 44(1), 47-86. 2015

[19] Schraefel, M-C., Wilson, M., Russell, A., Smith, D-A.: MSPACE: Improving Information Access to Multimedia Domains with Multimodal Exploratory Search. Communications of the ACM 49(4), 47-49. 2006

[20] Sierra, J.L., Fernández-Valmayor, A., Guinea, M., Hernanz, G. From Research Resources to Learning Objects: Process Model and Virtualizacion Experiences. Ed. Tech. & Society 9(3), 56-68. 2006

[21] Sierra, J.L., Moreno-Ger, P., Martínez-Ortiz, I., Fernández-Manjón, B. A highly modular and extensible architecture for an integrated IMS-based authoring system: the <e-Aula> experience. Software Practice and Experience 37(4): 441-461. 2007

[22] Way, T.; Eklund, P. Social Tagging for Digital Libraries using Formal Concept Analysis. CLA'10. 2010

[23] Yitzhak, O-B., Golbandj, N., Har'El N. et al. Beyond Basic Faceted Search. WSDM'08, 33-44. 2008

[24] Zobel, J., Moffat, A.: Inverted Files for Text Search Engines. ACM Computing Surveys 33(2), article 6. 2006

Referencias

Aho, A. V., Lam, M. S., Sethi, R., & Ullman, J. D. (2006). *Compilers: Principles, Techniques, and Tools* (2nd edition). Boston: Addison Wesley.

Aitchison, J., Gilchrist, A., & Bawden, D. (2000). *Thesaurus construction and use: a practical manual*. Psychology Press. Recuperado a partir de https://books.google.es/books?hl=es&lr=&id=p46zdHVkubMC&oi=fnd&pg=PP1&dq=Thesaurus+Construction+and+Use:+A+Practical+Manual&ots=vWdd5IzIx9&sig=xb7yKVF2sQDJWcPpmqvEuWo7Tns

Ambroziak, J. (2002). *Data indexing technique.* Google Patents. Recuperado a partir de https://www.google.com/patents/US6460047

ANSI/NISO, N. I. S. O. (US). (2005). *Guidelines for the construction, format, and management of monolingual controlled vocabularies*. NISO Press.

Arnaiz Barrero, J. E. (2008). ¡Chasqui: Repositorio de objetos virtuales independientes del dominio. Recuperado 11 de febrero de 2017, a partir de http://eprints.ucm.es/9214/

Arnold, D., Balkan, L., Humphreys, R. L., Meijer, S., & Sadler, L. (1994). *Machine translation: An introductory guide*. Recuperado a partir de https://pdfs.semanticscholar.org/4fe1/54f4682bb3b7b8bd13fb4d0015cb941d04dd.pdf

Aroyo, L., & Dicheva, D. (2004). The new challenges for e-learning: The educational semantic web. *Educational Technology & Society*, *7*(4), 59–69.

Azouaou, F., & Desmoulins, C. (2006). MemoNote, a context-aware annotation tool for teachers. En *Information Technology Based Higher Education and Training, 2006. ITHET'06. 7th International Conference on* (pp. 621–628). IEEE. Recuperado a partir de http://ieeexplore.ieee.org/abstract/document/4141686/

Baader, F. (2003). *The description logic handbook: Theory, implementation and applications*. Cambridge university press. Recuperado a partir de https://books.google.es/books?hl=es&lr=&id=e6_hJtM07qwC&oi=fnd&pg=PR9&dq=Baader,+F.,+Calvanese,+D.,+McGuinness,+D.+L.,+Nardi,+D.,+%26+Paterl-Schneider,+P.+F.+(Eds.).+The+description+logic+handbook:+theory,+implementation+and+applications+(2nd+ed..+Cambridge+University+Press.&ots=WjC2BUx-H8&sig=UFRS4nLk_yxBc27Azl2H0XVe7Yk

Baeza-Yates, R., Ribeiro-Neto, B., & others. (1999). *Modern information retrieval* (Vol. 463). ACM press New York. Recuperado a partir de ftp://mail.im.tku.edu.tw/seke/slide/baeza-yates/chap10_user_interfaces_and_visualization-modern_ir.pdf

Bainbridge, D., Ke, K.-Y., & Witten, I. H. (2006). Document level interoperability for collection creators. En *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries, 2006. JCDL '06* (pp. 105–106). Chapel Hill, NC, USA: ACM. https://doi.org/10.1145/1141753.1141773

Bao, S., Zhou, Y., & He, S. (2005). Research on pressure component design ontology building based on knowledge sharing and reusing. En *Computer Supported Cooperative Work in Design, 2005. Proceedings of the Ninth International Conference on* (Vol. 2, pp. 1183–1187). IEEE. Recuperado a partir de http://ieeexplore.ieee.org/abstract/document/1504264/

Barroso, I., Azevedo, M., & Ribeiro, C. (2009). Thematic digital libraries at the University of Porto: Metadata integration over a repository infrastructure. En *Research and Advanced Technology for Digital Libraries* (pp. 392–395). Springer. Recuperado a partir de http://link.springer.com/chapter/10.1007/978-3-642-04346-8_42

Baumeister, J., Reutelshoefer, J., Haupt, F., & Nadrowski, K. (2008). Capture and refactoring in knowledge wikis coping with the knowledge soup. Recuperado a partir de http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.142.6714

Baumeister, J., Reutelshoefer, J., & Puppe, F. (2011). KnowWE: a Semantic Wiki for knowledge engineering. *Applied Intelligence*, *35*(3), 323–344.

Baumgartner, R., Flesca, S., & Gottlob, G. (2001). Visual web information extraction with lixto. En *VLDB* (Vol. 1, pp. 119–128). Recuperado a partir de http://www.vldb.org/conf/2001/P119.pdf

Baumgartner, R., Frölich, O., & Gottlob, G. (2007). The Lixto systems applications in business intelligence and semantic Web. *The Semantic Web: Research and Applications*, 16–26.

Bechhofer, S. (2009). OWL: Web ontology language. En *Encyclopedia of Database Systems* (pp. 2008–2009). Springer. Recuperado a partir de http://link.springer.com/10.1007/978-0-387-39940-9_1073

Bechhofer, S., Carr, L., Goble, C., Kampa, S., & Miles-Board, T. (2002). The semantics of semantic annotation. En *OTM Confederated International Conferences« On the Move to Meaningful Internet Systems»* (pp. 1152–1167). Springer. Recuperado a partir de http://link.springer.com/chapter/10.1007/3-540-36124-3_73

Referencias

Bekaert, J., Liu, X., Van de Sompel, H., Lagoze, C., Payette, S., & Warner, S. (2006). Pathways core: a data model for cross-repository services. En *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries* (pp. 368–368). Chapel Hill, NC, USA: ACM. https://doi.org/10.1145/1141753.1141863

Bendaoud, R., Toussaint, Y., & Napoli, A. (2008). Pactole: A methodology and a system for semi-automatically enriching an ontology from a collection of texts. En *International Conference on Conceptual Structures* (pp. 203–216). Springer. Recuperado a partir de http://link.springer.com/chapter/10.1007/978-3-540-70596-3_14

Benedetti, B., & Masci, M. E. (2005). A New Project for the Pompeii's Superintendence Website: a Case Study. Perspectives for the Integration and On Line Publication of Digital Resources in an Institutional Repository.

Ben-Yitzhak, O., Golbandi, N., Har'El, N., Lempel, R., Neumann, A., Ofek-Koifman, S., … Yogev, S. (2008). Beyond basic faceted search. En *Proceedings of the 2008 International Conference on Web Search and Data Mining* (pp. 33–44). ACM. Recuperado a partir de http://dl.acm.org/citation.cfm?id=1341539

Berchtold, S., Böhm, C., Keim, D. A., Kriegel, H.-P., & Xu, X. (2000). Optimal multidimensional query processing using tree striping. En *International Conference on Data Warehousing and Knowledge Discovery* (pp. 244–257). Springer. Recuperado a partir de http://link.springer.com/chapter/10.1007/3-540-44466-1_24

Berry, D. (Ed.). (2012). *Understanding Digital Humanities* (2012 edition). Houndmills, Basingstoke, Hampshire ; New York: Palgrave Macmillan.

Berry, D. M. (2012). Introduction: Understanding the digital humanities. En *Understanding digital humanities* (pp. 1–20). Springer. Recuperado a partir de http://link.springer.com/chapter/10.1057/9780230371934_1

Bluhm, M., Getting, B., Hayft, M., & Walz, S. (2006). *Electronic document repository management and access system*. Google Patents. Recuperado a partir de https://www.google.com/patents/US7085755

BNE, B. N. de E. (2011a, enero 20). Historia de la Catalogación. Biblioteca Nacional de España. Recuperado 15 de febrero de 2017, a partir de http://www.bne.es/es/Inicio/Perfiles/Bibliotecarios/Procesos-tecnicos/NormasInternacionales/ReglasDeCatalogacion/HistoriaDeLaCatalogacion/

BNE, B. N. de E. (2011b, enero 21). Descripción Bibliográfica Internacional Normalizada (ISBD). Recuperado 15 de febrero de 2017, a partir de http://www.bne.es/es/Inicio/Perfiles/Bibliotecarios/Procesos-tecnicos/NormasInternacionales/ISBD/

BNE, B. N. de E. (2013). Manual de indización de Encabezamientos de Materia. Recuperado 15 de febrero de 2017, a partir de http://www.bne.es/es/Micrositios/Publicaciones/MEMBNE/

BNE, B. N. de E. (2014, octubre 30). Normas y estándares de catalogación. Recuperado 15 de febrero de 2017, a partir de http://www.bne.es/es/Inicio/Perfiles/Bibliotecarios/Procesos-tecnicos/NormasInternacionales/

BNE, B. N. de E. (2015, marzo 24). Políticas de la BNE. Recuperado 15 de febrero de 2017, a partir de http://www.bne.es/es/Inicio/Perfiles/Bibliotecarios/Procesos-tecnicos/NormativaBNE/

BNE, B. N. de E. (2016, abril 11). La BNE adoptará RDA como estándar de catalogación. Recuperado 15 de febrero de 2017, a partir de http://www.bne.es/es/AreaPrensa/noticias2016/1104-BNE-adoptara-RDA-como-estandar-de-catalogacion.html

Borges, J. L. (1944). Ficciones: La Biblioteca de Babel. *Alianza Editorial, Argentina.*

Brachman, R., & Levesque, H. (2004). *Knowledge Representation and Reasoning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Bradley, N. (2001). *The XML Companion* (3 edition). Addison-Wesley Professional.

Brewster, C., & O'Hara, K. (2004). Knowledge representation with ontologies: the present and future. *IEEE Intelligent Systems*, *19*(1), 72–81.

Brin, S. (1998). Extracting patterns and relations from the world wide web. En *International Workshop on The World Wide Web and Databases* (pp. 172–183). Springer. Recuperado a partir de http://link.springer.com/chapter/10.1007/10704656_11

Bueno, M. G. (2004). El proyecto Chasqui. En *En apoyo del aprendizaje en la universidad : hacia el espacio europeo de educación superior, 2004, ISBN 84-7491-774-3, págs. 228-233* (pp. 228-233). Editorial Complutense. Recuperado a partir de https://dialnet.unirioja.es/servlet/articulo?codigo=1220327

Calhoun, K. (2013). *Digital Libraries*. London: Facet Publishing.

Caplan, P. (2008). Repository to repository transfer of enriched archival information packages. *D-Lib Magazine*, *14*(11/12).

Caplan, P. (2010). DAITSS, an OAIS-based Preservation Repository. En *Proceedings of the 2010 Roadmap for Digital Preservation Interoperability Framework Workshop* (p. 17:1–17:4). Gaithersburg, MD, USA: ACM. https://doi.org/10.1145/2039274.2039291

Referencias

Caplan, P., Kehoe, W., & Pawletko, J. (2010). Towards Interoperable Preservation Repositories (TIPR). En *Proceedings of the 2010 Roadmap for Digital Preservation Interoperability Framework Workshop* (p. 16:1–16:4). Gaithersburg, MD, USA: ACM. https://doi.org/10.1145/2039274.2039290

Carpineto, C., & Romano, G. (2004). *Concept Data Analysis: Theory and Applications*. John Wiley & Sons.

Carr, L. (2009). *Eprints Architecture Development: on the Road to v3.2 - EPrints Files*. University of Southampton. Recuperado a partir de http://files.eprints.org/442/

Carrión Gútiez, M. (1988). Manual de bibliotecas. *Revista Española de Documentación Científica*, *11*(1), 106.

Cauffman, L. S., Thompson, J. N., & Cauffman, J. M. (1994). *Billing system with data indexing*. Google Patents. Recuperado a partir de https://www.google.com/patents/US5325290

Chan, L. M. (2007). *Cataloging and Classification: An Introduction* (3 edition). The Scarecrow Press, Inc.

Chi, Y.-L., Hsu, T.-Y., & Yang, W.-P. (2005). Building ontological knowledge bases for sharing knowledge in digital archive. En *2005 International Conference on Machine Learning and Cybernetics* (Vol. 4, p. 2261-2266 Vol. 4). https://doi.org/10.1109/ICMLC.2005.1527321

Chodorow, K. (2013). *MongoDB: The Definitive Guide: Powerful and Scalable Data Storage*. O'Reilly Media, Inc.

Chowdhury, G. (2010). *Introduction to Modern Information Retrieval, Third Edition* (3rd ed.). Facet Publishing.

Cigarrán, J. M., Gonzalo, J., Peñas, A., & Verdejo, F. (2004). Browsing search results via formal concept analysis: Automatic selection of attributes. En *International Conference on Formal Concept Analysis* (pp. 74–87). Springer. Recuperado a partir de http://link.springer.com/chapter/10.1007/978-3-540-24651-0_8

Cigarrán, J. M., Peñas, A., Gonzalo, J., & Verdejo, F. (2005). Automatic selection of noun phrases as document descriptors in an FCA-based information retrieval system. En *International Conference on Formal Concept Analysis* (pp. 49–63). Springer. Recuperado a partir de http://link.springer.com/10.1007%2F978-3-540-32262-7_4

Cigarrán-Recuero, J., Gayoso-Cabada, J., Rodríguez-Artacho, M., Romero-López, M.-D., Sarasa-Cabezuelo, A., & Sierra, J.-L. (2014). Assessing semantic annotation activities with formal concept analysis. *Expert Systems with Applications*, *41*(11), 5495–5508.

Cimiano, P., Handschuh, S., & Staab, S. (2004). Towards the self-annotating web. En *Proceedings of the 13th international conference on World Wide Web* (pp. 462–471). ACM. Recuperado a partir de http://dl.acm.org/citation.cfm?id=988735

Cimiano, P., Hotho, A., & Staab, S. (2005). Learning concept hierarchies from text corpora using formal concept analysis. *J. Artif. Intell. Res.(JAIR)*, *24*(1), 305–339.

Cimiano, P., Hotho, A., Stumme, G., & Tane, J. (2004). Conceptual knowledge processing with formal concept analysis and ontologies. En *International Conference on Formal Concept Analysis* (pp. 189–207). Springer. Recuperado a partir de http://link.springer.com/chapter/10.1007/978-3-540-24651-0_18

Cole, R. J., & Eklund, P. W. (1996). Application of formal concept analysis to information retrieval using a hierarchically structured thesaurus. En *Proc. Fourth International Conference on Conceptual Structures (to appear), Sydney, Australia*. Recuperado a partir de https://www.researchgate.net/profile/Peter_Eklund/publication/2323576_Application_of_Formal_Concept_Analysis_to_Information_Retrieval_using_a_Hierarchically_Structured_Thesaurus/links/0912f50b31630b733d000000.pdf

Coombs, J. H., Renear, A. H., & DeRose, S. J. (1987). Markup systems and the future of scholarly text processing. *Communications of the ACM*, *30*(11), 933–947.

Coyle, K., & Hillmann, D. (2007). Resource Description and Access (RDA): Cataloging Rules for the 20th Century. *D-Lib Magazine*, *13*(1/2). https://doi.org/10.1045/january2007-coyle

Culpepper, J. S., & Moffat, A. (2010). Efficient set intersection for inverted indexing. *ACM Transactions on Information Systems (TOIS)*, *29*(1), 1.

Da Rocha, T. R., Willrich, R., Fileto, R., & Tazi, S. (2009). Supporting collaborative learning activities with a digital library and annotations. En *Education and technology for a better world* (pp. 349–358). Springer. Recuperado a partir de http://link.springer.com/chapter/10.1007/978-3-642-03115-1_37

Dasiopoulou, S., Giannakidou, E., Litos, G., Malasioti, P., & Kompatsiaris, Y. (2011). A survey of semantic image and video annotation tools. En *Knowledge-driven multimedia information extraction and ontology evolution* (pp. 196–239). Springer. Recuperado a partir de http://link.springer.com/10.1007%2F978-3-642-20795-2_8

Davey, B. A., & Priestley, H. A. (2002). *Introduction to Lattices and Order* (2 edition). Cambridge, UK ; New York, NY: Cambridge University Press.

Referencias

DCMI, D. C. M. I. (2012). Dublin core metadata element set, version 1.1. Recuperado a partir de http://www.dublincore.org/documents/dces/

DCMI, D.-L. working group. (2000, agosto 6). DC-Library Application Profile (DC-Lib AP). Recuperado 15 de febrero de 2017, a partir de http://dublincore.org/documents/2001/10/12/library-application-profile/

de Souza, K. X. S., Davis, J., & de Medeiros Evangelista, S. R. (2006). Aligning ontologies, evaluating concept similarities and visualizing results. En *Journal on Data Semantics V* (pp. 211–236). Springer. Recuperado a partir de http://link.springer.com/chapter/10.1007/11617808_8

Dean, M., Schreiber, G., van Harmelen, F., Hendler, J., Horrocks, I., McGuinness, D. L., … Stein, L. A. (2003). *OWL web ontology language reference. W3C Working Draft*. March.

Devedzic, V., Jovanovic, J., & Gasevic, D. (2007). The pragmatics of current e-learning standards. *IEEE Internet Computing*, *11*(3). Recuperado a partir de http://ieeexplore.ieee.org/abstract/document/4196171/

Di Donato, F., Morbidoni, C., Fonda, S., Piccioli, A., Grassi, M., & Nucci, M. (2013). Semantic annotation with Pundit: a case study and a practical demonstration. En *Proceedings of the 1st international workshop on collaborative annotations in shared environment: metadata, vocabularies and techniques in the digital humanities* (p. 16). ACM. Recuperado a partir de http://dl.acm.org/citation.cfm?id=2517995

Dill, S., Eiron, N., Gibson, D., Gruhl, D., Guha, R., Jhingran, A., … others. (2003). A case for automated large-scale semantic annotation. *Web Semantics: Science, Services and Agents on the World Wide Web*, *1*(1), 115–132.

Dill, S., Eiron, N., Gibson, D., Gruhl, D., Guha, R., Jhingran, A., … others. (2003). SemTag and Seeker: Bootstrapping the semantic web via automated semantic annotation. En *Proceedings of the 12th international conference on World Wide Web* (pp. 178–186). ACM. Recuperado a partir de http://dl.acm.org/citation.cfm?id=775178

Dingli, A., Ciravegna, F., & Wilks, Y. (2003). Automatic semantic annotation using unsupervised information extraction and integration. En *Proceedings of SemAnnot 2003 Workshop*. Recuperado a partir de http://ceur-ws.org/Vol-101/Alexiei_Dingli-et-al.pdf

DRI, I. (2003). *IMS Digital Repositories specification V1. 0*.

du Preez, M. (2015). Taxonomies, folksonomies, ontologies: what are they and how do they support information retrieval? *The Indexer*, *33*(1), 29–37.

DuBois, A. (2003). *Close Reading: The Reader*. Duke University Press. Recuperado a partir de https://books.google.es/books?hl=es&lr=&id=YBbHmtKFCugC&oi=fnd&pg=PP11&dq=Lentricchia,+F.,+%26+Dubois,+A.+(2003).+Close+reading:+the+reader.+Durham,+N.+C.:+Duke+University+Press.&ots=QqrorsYHMy&sig=P6W6O7d9LZTokYbLD6n2XHnPr_o

DuraSpace, T. F. D. T. (2008). Tutorial 1 - Introduction to Fedora - Fedora Create - DuraSpace Wiki. Recuperado 5 de febrero de 2017, a partir de https://wiki.duraspace.org/display/FEDORACREATE/Tutorial+1+-+Introduction+to+Fedora

Etzioni, O., Cafarella, M., Downey, D., Popescu, A.-M., Shaked, T., Soderland, S., … Yates, A. (2005). Unsupervised named-entity extraction from the web: An experimental study. *Artificial intelligence*, *165*(1), 91–134.

Evjen, B., Sharkey, K., Thangarathinam, T., Kay, M., Vernet, A., & Ferguson, S. (2007). *Professional XML* (1 edition). Indianapolis, IN: Wrox.

Fan, L., & Xiao, T. (2007). *An Automatic Method for Ontology Mapping. B. Apolloni et al.(Eds.): KES/WIRN, Part III, LNAI 4694, 661-669*. Springer.

Fernández-Pampillón Cesteros, A. (2012, abril). El proyecto OdA: «Objetos de Aprendizaje en el campus virtual». Recuperado 1 de noviembre de 2016, a partir de http://eprints.sim.ucm.es/14924/

Fernández-Valmayor Crespo, A., Fernández-Pampillón Cesteros, A., & Varadero Software Factory, V. S. F. (2013, febrero). Guía de Gestión del repositorio de Objetos Digitales OdA. Recuperado 1 de noviembre de 2016, a partir de http://eprints.sim.ucm.es/20263/

Fernández-Valmayor Crespo, A., Guinea Bueno, M., Navarro Martín, A., & Sierra Rodríguez, J. L. (2005). Integración de investigación y docencia en el campus virtual: el sistema Chasqui. En A. Fernández-Pampillón Cesteros & J. Merino Granizo (Eds.), *II Jornada Campus Virtual UCM: cómo integrar investigación y docencia en el CV-UCM* (pp. 336-348). Madrid: Editorial Complutense. Recuperado a partir de http://eprints.sim.ucm.es/5808/

Fisher, D., & Frey, N. (2012). Close Reading In Elementary Schools. *The Reading Teacher*, *66*(3), 179-188. https://doi.org/10.1002/TRTR.01117

Formica, A. (2006). Ontology-based concept similarity in formal concept analysis. *Information sciences*, *176*(18), 2624–2641.

Fraternali, P., Rossi, G., & Sánchez-Figueroa, F. (2010). Rich Internet Applications. *IEEE Internet Computing*, *14*(3), 9-12. https://doi.org/10.1109/MIC.2010.76

Referencias

Friesen, N. (2004). The International Learning Object Metadata Survey. *International Review of Research in Open and Distance Learning*, *5*(3), n3.

Fritz, D. A., & Fritz, R. J. (2003). *MARC21 for Everyone: a practical guide*. American Library Association. Recuperado a partir de https://books.google.es/books?hl=es&lr=&id=1-3itZFuZ4MC&oi=fnd&pg=PA1&dq=MARC21+for+Everyone:+a+practical+guide&ots=1QrINzxkRc&sig=L18kqNxzB3NWuXiQiV24voLR8uk

Gamallo, P., Lopes, G. P., & Agustini, A. (2007). Inducing classes of terms from text. En *International Conference on Text, Speech and Dialogue* (pp. 31–38). Springer. Recuperado a partir de http://link.springer.com/chapter/10.1007/978-3-540-74628-7_7

Ganter, B., & Wille, R. (1999). *Formal Concept Analysis: Mathematical Foundation*. Springer-Verlag New York Incorporated.

Garrido Arilla, M. R. (1996). *Teoría e historia de la catalogación de documentos*. Síntesis. Recuperado a partir de http://infocuib.laborales.unam.mx/~ec08s02b/archivos/data/1/23.pdf

Gayo, J. E. L., De Pablos, P. O., & Lovelle, J. M. C. (2010). WESONet: Applying semantic web technologies and collaborative tagging to multimedia web information systems. *Computers in Human Behavior*, *26*(2), 205–209.

Gayoso Cabada, J. (2012). Modelo para la anotación colaborativa de textos literarios digitalizados. Recuperado a partir de http://eprints.ucm.es/16747/

Gayoso Cabada, J., Fernández-Pampillón Cesteros, A. M., & Sierra Rodríguez, J. L. (2015a). Exportador/Actualizador de datos OdA-XLS. Guía Técnica. Recuperado 1 de noviembre de 2016, a partir de http://eprints.ucm.es/33011/

Gayoso Cabada, J., Fernández-Pampillón Cesteros, A. M., & Sierra Rodríguez, J. L. (2015b). Generador de informes HTML con OdA-Clavy. Guía Técnica. Recuperado 1 de noviembre de 2016, a partir de http://eprints.sim.ucm.es/33012/

Gayoso-Cabada, Joaquín, Rodríguez-Cerezo, D., & Sierra, J.-L. (2016a). Browsing Digital Collections with Reconfigurable Faceted Thesauri. En *International Conference on Information Systems Development (ISD)* (pp. 378-389). Recuperado a partir de http://aisel.aisnet.org/isd2014/proceedings2016/CogScience/5

Gayoso-Cabada, Joaquín, Rodríguez-Cerezo, D., & Sierra, J.-L. (2016b). Learning object repositories with dynamically reconfigurable metadata schemata. En *Proceedings of the XVI International Symposium on Computers in Education, SIIE 2016* (p. 6). IEEE Computer Society. https://doi.org/10.1109/SIIE.2016.7751848

Gayoso-Cabada, Joaquín, Rodríguez-Cerezo, D., & Sierra, J.-L. (2016c). Multilevel Browsing of Folksonomy-Based Digital Collections. En *Proceedings of the 17th Conference on Web Information Systems Engineering Part II, WISE 2016* (pp. 43-51). Lecture Notes in Computer Science 10042, Springer. https://doi.org/10.1007/978-3-319-48743-4_4

Gayoso-Cabada, Joaquin, Ruiz, C., Pablo-Nuñez, L., Sarasa-Cabezuelo, A., Goicoechea-de-Jorge, M., Sanz-Cabrerizo, A., & Sierra-Rodriguez, J.-L. (2012). A flexible model for the collaborative annotation of digitized literary works. En *Proceedings of the 2012 Digital Humanities Conference* (pp. 190–193). Recuperado a partir de http://www.dh2012.uni-hamburg.de/conference/programme/abstracts/a-flexible-model-for-the-collaborative-annotation-of-digitized-literary-works/

Gayoso-Cabada, Joaquín, Sanz-Cabrerizo, A., & Sierra, J.-L. (2013). @Note: An Electronic Tool for Academic Readings. En *Proceedings of the 1st International Workshop on Collaborative Annotations in Shared Environment: Metadata, Vocabularies and Techniques in the Digital Humanities* (p. 17:1–17:4). Florence, Italy: ACM. https://doi.org/10.1145/2517978.2517996

Gerolimos, M., Papadourakis, G., Nikitakis, M., & Sitas, A. (2011). Cataloging conventional and digital objects: new tools with old names or old names to new tools? Recuperado a partir de https://repository.edulll.gr/edulll/handle/10795/682

Giannopoulos, G., Bikakis, N., Dalamagas, T., & Sellis, T. (2010). GoNTogle: a tool for semantic annotation and search. En *Extended Semantic Web Conference* (pp. 376–380). Springer. Recuperado a partir de http://link.springer.com/10.1007%2F978-3-642-13489-0_27

Gigee, G. (2006). *MARC and MARCXML*.

Godin, R., Saunders, E., & Gecsei, J. (1986). Lattice model of browsable data spaces. *Information Sciences*, *40*(2), 89–116.

Goldsmith, B., & Knudson, F. (2006). Looking Back, Looking Forward: A Metadata Standard for LANL's aDORe Repository. En *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 272–273). New York, NY, USA: ACM. https://doi.org/10.1145/1141753.1141814

Gonzalez-Barbone, V., & Anido-Rifon, L. (2008). Creating the first SCORM object. *Computers & Education*, *51*(4), 1634–1647.

Referencias

Gonzalez-Barbone, V., & Anido-Rifon, L. (2010). From SCORM to Common Cartridge: A step forward. *Computers & Education*, *54*(1), 88–102.

Graefe, G. (2006). B-tree indexes for high update rates. *ACM Sigmod Record*, *35*(1), 39–44.

Grainger, T., Potter, T., & Seeley, Y. (2014). *Solr in action*. Manning Cherry Hill. Recuperado a partir de http://toc.dreamtechpress.com/toc_978-93-5119-435-4.pdf

Greaves, M. (2004). *The DARPA agent markup language homepage*.

Greene, G. J. (2015). A generic framework for concept-based exploration of semi-Structured Software Engineering Data. En *Automated Software Engineering (ASE), 2015 30th IEEE/ACM International Conference on* (pp. 894–897). IEEE. Recuperado a partir de http://ieeexplore.ieee.org/abstract/document/7372087/

Greene, G. J., Dunaiski, M., Fischer, B., Ilvovsky, D., & Kuznetsov, S. O. (2015). Browsing publication data using tag clouds over concept lattices constructed by key-phrase extraction. En *Proceedings of Russian and South African Workshop on Knowledge Discovery Techniques Based on Formal Concept Analysis* (pp. 10–22). Recuperado a partir de https://pdfs.semanticscholar.org/329c/a8ab508467d86d6678d7a7bcc9778e228a3f.pdf

Greene, G. J., & Fischer, B. (2015). Interactive tag cloud visualization of software version control repositories. En *Software Visualization (VISSOFT), 2015 IEEE 3rd Working Conference on* (pp. 56–65). IEEE. Recuperado a partir de http://ieeexplore.ieee.org/abstract/document/7332415/

Guenther, R., & Radebaugh, J. (2006). Standards Showcase: MODS, METS, MARCXML. *Recuperado Septiembre*, *25*, 2007.

Guermeur, D., & Unruh, A. (2010). *Google App Engine Java and GWT Application Development*. Packt Publishing Ltd. Recuperado a partir de https://books.google.es/books?hl=es&lr=&id=9fXccQMwwuQC&oi=fnd&pg=PT2&dq=Unruh,+A.+(2010+).+Google+App+Engine+Java+and+GWT+Application+Development.+Birmingham:+Packt+Publishing.&ots=NCCA4GGm8a&sig=RskM7HjlOtf9bDQJocTTYnc7zko

Handschuh, S., & Staab, S. (2002). Authoring and annotation of web pages in CREAM. En *Proceedings of the 11th international conference on World Wide Web* (pp. 462–473). ACM. Recuperado a partir de http://dl.acm.org/citation.cfm?id=511506

Handschuh, S., Staab, S., & Ciravegna, F. (2002). S-CREAM—semi-automatic creation of metadata. En *International Conference on Knowledge Engineering and Knowledge Management* (pp. 358–372). Springer. Recuperado a partir de http://link.springer.com/10.1007%2F3-540-45810-7_32

Heath, B. P., McArthur, D. J., McClelland, M. K., & Vetter, R. J. (2005). Metadata lessons from the iLumina digital library. *Communications of the ACM*, *48*(7), 68–74.

Hedden, H. (2008). Controlled vocabularies, thesauri, and taxonomies. *The Indexer*, *26*(1), 33–34.

Heinz, S., & Zobel, J. (2003). Efficient Single-pass Index Construction for Text Databases. *J. Am. Soc. Inf. Sci. Technol.*, *54*(8), 713–729. https://doi.org/10.1002/asi.10268

Hepp, M. (2007). Possible ontologies: How reality constrains the development of relevant ontologies. *IEEE Internet Computing*, *11*(1). Recuperado a partir de http://ieeexplore.ieee.org/abstract/document/4061129/

Hernandez, M.-E., Falconer, S. M., Storey, M.-A., Carini, S., & Sim, I. (2008). Synchronized tag clouds for exploring semi-structured clinical trial data. En *Proceedings of the 2008 conference of the center for advanced studies on collaborative research: meeting of minds* (p. 4). ACM. Recuperado a partir de http://dl.acm.org/citation.cfm?id=1463794

Hildebrand, M., van Ossenbruggen, J., & Hardman, L. (2006). /facet: A browser for heterogeneous semantic web repositories. En *International Semantic Web Conference* (pp. 272–285). Springer. Recuperado a partir de http://link.springer.com/10.1007%2F11926078_20

Hirst, G. (2009). Ontology and the lexicon. En *Handbook on ontologies* (pp. 269–292). Springer. Recuperado a partir de http://link.springer.com/chapter/10.1007/978-3-540-92673-3_12

Hitzler, P., Krotzsch, M., & Rudolph, S. (2009). *Foundations of semantic web technologies*. CRC Press. Recuperado a partir de https://books.google.com/books?hl=es&lr=&id=BdzL24RqcGIC&oi=fnd&pg=PP1&dq=Foundations+of+semantic+web++technologies&ots=EfFSRthu_M&sig=W9Ih2FBwCSXDJpqMV8C65cXkD60

Hogue, A., & Karger, D. (2005). Thresher: automating the unwrapping of semantic content from the World Wide Web. En *Proceedings of the 14th international conference on World Wide Web* (pp. 86–95). ACM. Recuperado a partir de http://dl.acm.org/citation.cfm?id=1060762

Horrocks, I. (2008). Ontologies and the Semantic Web. *Commun. ACM*, *51*(12), 58–67. https://doi.org/10.1145/1409360.1409377

Referencias

Huang, J.-W., Chen, K.-Y., Chen, Y.-C., Yang, K.-N., Hwang, I.-S., Huang, W.-C., & others. (2014). A Novel Spatial Tag Cloud Using Multi-Level Clustering. *J. Inf. Sci. Eng.*, *30*(3), 687–700.

Hunter, J., & Gerber, A. (2010). Harvesting community annotations on 3D models of museum artefacts to enhance knowledge, discovery and re-use. *Journal of Cultural Heritage*, *11*(1), 81–90.

Huynh, D., Karger, D., & Quan, D. (2002). Haystack: A platform for creating, organizing and visualizing information using RDF. En *Proceedings of the 3rd International Conference on Semantic Web-Volume 55* (pp. 76–87). CEUR-WS. org. Recuperado a partir de http://dl.acm.org/citation.cfm?id=3000579

Huynh, D., Mazzocchi, S., & Karger, D. (2005). Piggy bank: Experience the semantic web inside your web browser. En *International Semantic Web Conference* (pp. 413–430). Springer. Recuperado a partir de http://link.springer.com/10.1007%2F11574620_31

IEEE, I. L. T. S. C. (2001). *IEEE LOM working draft 6.1*.

IEEE, I. L. T. S. C. IEEE Standard for Learning Object Metadata (2002). Recuperado a partir de https://standards.ieee.org/findstds/standard/1484.12.1-2002.html

IEEE SA - 1484.12.1-2002 - IEEE Standard for Learning Object Metadata. (s. f.). Recuperado 10 de marzo de 2017, a partir de https://standards.ieee.org/findstds/standard/1484.12.1-2002.html

IFLA, F. I. de A. de B. y B. (2011). International Standard Bibliographic Description. Recuperado 15 de febrero de 2017, a partir de http://www.ifla.org/publications/international-standard-bibliographic-description

IFLA, F. I. de A. de B. y B., & BNE, B. N. de E. (2014). Área 0 : Forma del contenido y tipo de medio. Política de uso en la Biblioteca Nacional de España. Recuperado a partir de http://travesia.mcu.es/portalnb/jspui/handle/10421/2299

ISO 3166/MA, I. 3166 M. A. (1997). ISO 3166 - Country codes. Recuperado 16 de febrero de 2017, a partir de http://www.iso.org/iso/home/standards/country_codes.htm

ISO, E. (2013). 472: 2013. *Plastics-Vocabulary (ISO*, *472*.

ISO/TC 37/SC 1. (2014). ISO 1087-1:2000 - Terminology work -- Vocabulary -- Part 1: Theory and application. Recuperado 16 de febrero de 2017, a partir de http://www.iso.org/iso/catalogue_detail.htm?csnumber=20057

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. *ACM computing surveys (CSUR)*, *31*(3), 264–323.

Jiang, G., & Chute, C. G. (2009). Auditing the semantic completeness of SNOMED CT using formal concept analysis. *Journal of the American Medical Informatics Association*, *16*(1), 89–102.

Jiang, G., Ogasawara, K., Endoh, A., & Sakurai, T. (2003). Context-based ontology building support in clinical domains using formal concept analysis. *International journal of medical informatics*, *71*(1), 71–81.

Jiang, G., Pathak, J., & Chute, C. G. (2009). Formalizing ICD coding rules using formal concept analysis. *Journal of Biomedical Informatics*, *42*(3), 504–517.

Jones, R. (2007). Giving birth to next generation repositories. *International Journal of Information Management*, *27*(3), 154–158.

JSC, J. S. C. (2014). Joint Steering Committee for Development of RDA: RDA. Recuperado 15 de febrero de 2017, a partir de http://www.rda-jsc.org/archivedsite/rda.html#background

Jurkiewicz, J., & Nowiński, A. (2011). Detailed Presentation versus Ease of Search—Towards the Universal Format of Bibliographic Metadata. Case Study of Dealing with Different Metadata Kinds during Import to Virtual Library of Science. En *Research Conference on Metadata and Semantic Research* (pp. 186–193). Springer. Recuperado a partir de http://link.springer.com/chapter/10.1007/978-3-642-24731-6_19

Kahn, R., & Wilensky, R. (2006). A framework for distributed digital object services. *International Journal on Digital Libraries*, *6*(2), 115-123. https://doi.org/10.1007/s00799-005-0128-x

Katifori, A., Halatsis, C., Lepouras, G., Vassilakis, C., & Giannopoulou, E. (2007). Ontology visualization methods—a survey. *ACM Computing Surveys (CSUR)*, *39*(4), 10.

Keyser, P. D. (2007). *Indexing: From Thesauri to the Semantic Web*. Chandos Publishing (Oxford), Limited.

Kim, D.-S., Hwang, S.-H., & Kim, H.-G. (2007). Concept analysis of OWL ontology based on the context family model. En *Convergence Information Technology, 2007. International Conference on* (pp. 896–901). IEEE. Recuperado a partir de http://ieeexplore.ieee.org/abstract/document/4420373/

Kimball, R., & Caserta, J. (2004). *The Data Warehouse?ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data* (1 edition). Indianapolis, IN: Wiley.

Referencias

Kiu, C.-C., & Lee, C.-S. (2008). Ontological knowledge management through hybrid unsupervised clustering techniques. En *Asia-Pacific Web Conference* (pp. 499–510). Springer. Recuperado a partir de http://link.springer.com/chapter/10.1007/978-3-540-78849-2_50

Kiyavitskaya, N., Zeni, N., Cordy, J. R., Mich, L., & Mylopoulos, J. (2009). Cerno: Light-weight tool support for semantic annotation of textual documents. *Data & Knowledge Engineering*, *68*(12), 1470–1492.

Kiyavitskaya, N., Zeni, N., Mich, L., Cordy, J. R., & Mylopoulos, J. (2007). Annotating accommodation advertisements using cerno. *Information and Communication Technologies in Tourism 2007*, 389–400.

Kogut, P. A., & Holmes III, W. S. (2001). AeroDAML: Applying Information Extraction to Generate DAML Annotations from Web Pages. En *Semannot@ K-CAP 2001*. Recuperado a partir de http://km.aifb.kit.edu/ws/semannot2001/positionpapers/AeroDAML3.pdf

Koivunen, M.-R. (2005). Annotea and semantic web supported collaboration. En *Invited talk at workshop on user aspects of the semantic web (User-SWeb) at European semantic web conference* (pp. 5–16). Recuperado a partir de http://ceur-ws.org/Vol-137/01_koivunen_final.pdf

Koutrika, G., Zadeh, Z. M., & Garcia-Molina, H. (2009). Coursecloud: summarizing and refining keyword searches over structured data. En *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology* (pp. 1132–1135). ACM. Recuperado a partir de http://dl.acm.org/citation.cfm?id=1516496

Kriegel, H.-P. (1984). Performance comparison of index structures for multi-key retrieval. En *ACM SIGMOD Record* (Vol. 14, pp. 186–196). ACM. Recuperado a partir de http://dl.acm.org/citation.cfm?id=602284

Kroeger, A. (2013). The road to BIBFRAME: the evolution of the idea of bibliographic transition into a post-MARC Future. *Cataloging & classification quarterly*, *51*(8), 873–890.

Krötzsch, M., Hitzler, P., & Zhang, G.-Q. (2005). Morphisms in context. En *International Conference on Conceptual Structures* (pp. 223–237). Springer. Recuperado a partir de http://link.springer.com/10.1007%2F11524564_15

Krötzsch, M., Vrandečić, D., & Völkel, M. (2006). Semantic mediawiki. En *International semantic web conference* (pp. 935–942). Springer. Recuperado a partir de http://link.springer.com/10.1007%2F11926078_68

Kurilovas, E., Kubilinskiene, S., & Dagiene, V. (2014). Web 3.0–Based personalisation of learning objects in virtual learning environments. *Computers in Human Behavior*, *30*, 654–662.

Kushmerick, N. (2000). Wrapper induction: Efficiency and expressiveness. *Artificial Intelligence*, *118*(1-2), 15–68.

Kuznetsov, S. O. (2001). On computing the size of a lattice and related decision problems. *Order*, *18*(4), 313–321.

Lagoze, C., Lynch, C., Waters, D., Van De Sompel, H., & Hey, T. (2006). Augmenting interoperability across scholarly repositories (pp. 85–85). Presentado en Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL'06, Chapel Hill, NC, USA: IEEE. https://doi.org/10.1145/1141753.1141768

Lamarca Lapuente, M. J. (2006). *Hipertexto: el nuevo concepto de documento en la cultura de la imagen* (tesis). Recuperado a partir de http://www.hipertexto.info/

Lancaster, F. W. (1972). Vocabulary control for information retrieval. Recuperado a partir de http://eric.ed.gov/?id=ED075999

Lazarinis, F., & Fotis Lazarinis. (2015). *Cataloguing and Classification: An Introduction to AACR2, RDA, DDC, LCC, LCSH and MARC 21 Standards.* Elsevier.

LC, L. of C., & NDMSO, N. D. and M. S. O. (1999). MARC 21 Format for Bibliographic Data: Table of Contents (Network Development and MARC Standards Office, Library of Congress). Recuperado 2 de diciembre de 2016, a partir de https://www.loc.gov/marc/bibliographic/

Lee, W., & Sugimoto, S. (2006). Toward core subject vocabularies for community-oriented subject gateways. *International Journal of Metadata, Semantics and Ontologies*, *1*(3), 167–175.

Leone, S., Geel, M., Müller, C., & Norrie, M. C. (2010). Exploiting tag clouds for database browsing and querying. En *Forum at the Conference on Advanced Information Systems Engineering (CAiSE)* (pp. 15–28). Springer. Recuperado a partir de http://link.springer.com/chapter/10.1007/978-3-642-17722-4_2

Lester, N., Zobel, J., & Williams, H. E. (2004). In-place Versus Re-build Versus Re-merge: Index Maintenance Strategies for Text Retrieval Systems. En *Proceedings of the 27th Australasian Conference on Computer Science - Volume 26* (pp. 15–23). Darlinghurst, Australia, Australia: Australian Computer Society, Inc. Recuperado a partir de http://dl.acm.org/citation.cfm?id=979922.979925

Lewis, D. D., & Spärck Jones, K. (1996). Natural language processing for information retrieval. *Communications of the ACM*, *39*(1), 92–101.

Referencias

Li, C., Yan, N., Roy, S. B., Lisham, L., & Das, G. (2010). Facetedpedia: dynamic generation of query-dependent faceted interfaces for wikipedia. En *Proceedings of the 19th international conference on World wide web* (pp. 651–660). ACM. Recuperado a partir de http://dl.acm.org/citation.cfm?id=1772757

Li, R., Bao, S., Yu, Y., Fei, B., & Su, Z. (2007). Towards effective browsing of large scale social annotations. En *Proceedings of the 16th international conference on World Wide Web* (pp. 943–952). ACM. Recuperado a partir de http://dl.acm.org/citation.cfm?id=1242700

Lim, L., Wang, M., Padmanabhan, S., Vitter, J. S., & Agarwal, R. (2003). Dynamic maintenance of web indexes using landmarks. En *Proceedings of the 12th international conference on World Wide Web* (pp. 102–111). Budapest, Hungary: ACM. Recuperado a partir de http://dl.acm.org/citation.cfm?id=775167

Lipscomb, C. E. (2000). Medical subject headings (MeSH). *Bulletin of the Medical Library Association*, *88*(3), 265.

LMF, L. M. F. (2008). *ISO code number for LMF is ISO-24613: 2008*.

Lynch, C., Parastatidis, S., Jacobs, N., Van de Sompel, H., & Lagoze, C. (2007). The OAI-ORE effort: progress, challenges, synergies. En *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital libraries* (pp. 80–80). Vancouver, BC, Canada: ACM. Recuperado a partir de http://dl.acm.org/citation.cfm?id=1255190

Mahdi Taheri, S., & Hariri, N. (2012). A comparative study on the indexing and ranking of the content objects including the MARCXML and Dublin Core's metadata elements by general search engines. *The Electronic Library*, *30*(4), 480–491.

Malik, S. K., Prakash, N., & Rizvi, S. A. M. (2010). Semantic annotation framework for intelligent information retrieval using KIM architecture. *International Journal of Web & Semantic Technology (IJWest)*, *1*(4), 12–26.

Martínez de Sousa, J. (2004). *Diccionario de bibliología y ciencias afines* (3.a ed.). Recuperado a partir de https://dialnet.unirioja.es/servlet/libro?codigo=321483

Mathes, A. (2004). *Folksonomies-cooperative classification and communication through shared metadata*. December. Recuperado a partir de http://firstmonday.org/ojs/index.php/fm/article/downloadSuppFile/4994/1186

Maynard, D. (2003). Multi-source and multilingual information extraction. *Expert Update*, *6*(3), 11–16.

McCandless, M., Hatcher, E., & Gospodnetić, O. (2010). *Lucene in Action*. Manning.

Méndez, E. (2006). Dublin Core, metadatos y vocabularios. *El profesional de la información*, *15*(2), 84–86.

Mu, X. (2010). Towards effective video annotation: An approach to automatically link notes with video content. *Computers & Education*, *55*(4), 1752–1763.

Nasir Uddin, M., & Janecek, P. (2007). The implementation of faceted classification in web site searching and browsing. *Online information review*, *31*(2), 218–233.

Neven, F., & Duval, E. (2002). Reusable learning objects: a survey of LOM-based repositories. En *Proceedings of the tenth ACM international conference on Multimedia* (pp. 291–294). Juan les Pins, France: ACM. Recuperado a partir de http://dl.acm.org/citation.cfm?id=641067

Nilsson, M. (2008). Harmonization of metadata standards. Recuperado a partir de http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.567.5444

Nilsson, M., Baker, T., & Johnston, P. (2008). Interoperability Levels for Dublin Core Metadata. Recuperado 5 de marzo de 2017, a partir de http://dublincore.org/documents/2008/11/03/interoperability-levels/

Noruzi, A. (2006). Folksonomies:(un) controlled vocabulary? *Knowledge organization*, *33*(4), 199–203.

Noy, N. F., Sintek, M., Decker, S., Crubézy, M., Fergerson, R. W., & Musen, M. A. (2001). Creating semantic web contents with protege-2000. *IEEE intelligent systems*, *16*(2), 60–71.

Oliveira, P., & Rocha, J. (2013). Semantic annotation tools survey. En *Computational Intelligence and Data Mining (CIDM), 2013 IEEE Symposium on* (pp. 301–307). IEEE. Recuperado a partir de http://ieeexplore.ieee.org/abstract/document/6597251/

Pal, J. K., & others. (2016). Resolving the confusion over metadata-creation in digital archives. *Annals of Library and Information Studies (ALIS)*, *63*(2), 110–116.

Park, J., & Tosaka, Y. (2010). Metadata creation practices in digital repositories and collections: Schemata, selection criteria, and interoperability. *Information Technology and Libraries*, *29*(3), 104.

Peltier, M., Bézivin, J., & Guillaume, G. (s. f.). *MTRANS: A general framework, based on XSLT, for model transformations*.

Perugini, S. (2010). Supporting multiple paths to objects in information hierarchies: Faceted classification, faceted search, and symbolic links. *Information processing & management*, *46*(1), 22–43.

Referencias

Peterson, E. (2006). Beneath the metadata: Some philosophical problems with folksonomy. *D-Lib Magazine*, *12*(11). Recuperado a partir de http://www.citeulike.org/group/2924/article/950399

Poelmans, J., Ignatov, D. I., Kuznetsov, S. O., & Dedene, G. (2013). Formal concept analysis in knowledge processing: A survey on applications. *Expert systems with applications*, *40*(16), 6538–6560.

Polsani, P. R. (2006). Use and abuse of reusable learning objects. *Journal of Digital information*, *3*(4). Recuperado a partir de https://journals.tdl.org/jodi/index.php/jodi/article/view/89

Ponnekanti, N. (2003). *Database system with methodology for online index rebuild*. Google Patents. Recuperado a partir de https://www.google.com/patents/US6591269

Ponnekanti, N., & Kodavalla, H. (2000). Online index rebuild. En *ACM SIGMOD Record* (Vol. 29, pp. 529–538). ACM. Recuperado a partir de http://dl.acm.org/citation.cfm?id=335462

Popov, B., Kiryakov, A., Kirilov, A., Manov, D., Ognyanoff, D., & Goranov, M. (2003). KIM–semantic annotation platform. En *International Semantic Web Conference* (pp. 834–849). Springer. Recuperado a partir de http://link.springer.com/chapter/10.1007/978-3-540-39718-2_53

Popov, B., Kiryakov, A., Ognyanoff, D., Manov, D., & Kirilov, A. (2004). KIM–a semantic platform for information extraction and retrieval. *Natural language engineering*, *10*(3-4), 375–392.

Proudfoot, R. (2005). The White Rose Consortium ePrints Repository: creating a shared institutional repository for the Universities of Leeds, Sheffield and York. *Aliss Quarterly*, 19–23.

Radelaar, J., Boor, A.-J., Vandic, D., Van Dam, J.-W., & Fasincar, F. (2014). Improving search and exploration in tag spaces using automated tag clustering. *Journal of Web Engineering*, *13*(3-4), 277–301.

RAE, R. A. E. (2013). Diccionario de la Real Academia Española (DRAE). *la Red*.

Rani, S., Goodkin, J., Cobb, J., Habing, T., Urban, R., Eke, J., & Pearce-Moses, R. (2006). Technical Architecture Overview: Tools for Acquisition, Packaging and Ingest of Web Objects into Multiple Repositories. En *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 360–360). Chapel Hill, NC, USA: ACM. https://doi.org/10.1145/1141753.1141855

Reed, S. L., & Lenat, D. B. (2002). Mapping ontologies into Cyc. En *AAAI 2002 Conference Workshop on Ontologies For The Semantic Web* (pp. 1–6). Recuperado a partir de http://www.aaai.org/Papers/Workshops/2002/WS-02-11/WS02-11-010.pdf

Reeve, L., & Han, H. (2005). Survey of semantic annotation platforms. En *Proceedings of the 2005 ACM symposium on Applied computing* (pp. 1634–1638). ACM. Recuperado a partir de http://dl.acm.org/citation.cfm?id=1067049

Richards, D. (2004). Addressing the ontology acquisition bottleneck through reverse ontological engineering. *Knowledge and Information Systems*, *6*(4), 402–427.

Richards, D. (2006). Ad-Hoc and personal ontologies: a prototyping approach to ontology engineering. En *Pacific Rim Knowledge Acquisition Workshop* (pp. 13–24). Springer. Recuperado a partir de http://link.springer.com/chapter/10.1007/11961239_2

Richardson, L., & Ruby, S. (2008). *RESTful web services*. O'Reilly Media, Inc. Recuperado a partir de https://books.google.es/books?hl=es&lr=&id=XUaErakHsoAC&oi=fnd&pg=PP1&dq=Richardson,+L.,+and+S.+Ruby+(2007).+Restful+web+services.+Beijing:+O%E2%80%99Reilly.&ots=5kjpAnlGoC&sig=v4_b0_PtoEMqcz_OOV83jAQ_gro

Rodríguez Bravo, B. (2011). *Apuntes sobre representación y organización de la información*. Recuperado a partir de https://dialnet.unirioja.es/servlet/libro?codigo=604525

Rodríguez, E. E. (2007). Edición preliminar de la ISBD consolidada. *Proc. of III Encuentro Internacional de Catalogadores, Buenos Aires*, 8.

Romero López, D. (2013). Mnemosine: Hacia una Biblioteca Digital de Textos Raros y Olvidados en la Edad de Plata (1868-1936). Recuperado a partir de https://www.mysciencework.com/publication/show/567eba6238bb5593bf10773ab565de63

Romero López, D. (2014). Hacia la Smartlibrary: Mnemosine, una biblioteca digital de textos literarios raros y olvidados de la Edad de Plata (1868-1936) 1. Fase I. *Janus: estudios sobre el Siglo de Oro*, (1), 411–422.

Rudolph, S. (2004). *Exploring Relational Structures Via FLE. KE Wolff et al.(Eds.): ICCS, LNAI 3127, 196-212*. Springer.

Rudolph, Sebastian. (2008). Acquiring generalized domain-range restrictions. En *International Conference on Formal Concept Analysis* (pp. 32–45). Springer. Recuperado a partir de http://link.springer.com/chapter/10.1007/978-3-540-78137-0_3

Referencias

Rudolph, Sebastian, Völker, J., & Hitzler, P. (2007). Supporting lexical ontology learning by relational exploration. En *International Conference on Conceptual Structures* (pp. 488–491). Springer. Recuperado a partir de http://link.springer.com/chapter/10.1007/978-3-540-73681-3_41

Ruiz, C., Gayoso-Cabada, J., Sarasa-Cabezuelo, A., Pablo-Nuñez, L., Sanz-Cabrerizo, A., & Sierra-Rodriguez, J.-L. (2012). Web-services API in@ Note. En *Proc. of the INTEREDITION Symposium on Scholarly Digital Editions, Tools and Infrastructure*. Recuperado a partir de http://www.interedition.eu/wp-content/bestanden/2012/03/8_1.pdf

Rumbaugh, J., Jacobson, I., & Booch, G. (2005). *The Unified Modeling Language Reference Manual (2nd Edition)*. Pearson Higher Education.

Salem, R. (2009). Complex data integration into an active XML repository. En *Proceedings of the International Conference on Management of Emergent Digital EcoSystems* (p. 76). ACM.

Salton, G., & McGill, M. J. (1986). Introduction to modern information retrieval. Recuperado a partir de http://www.citeulike.org/group/1808/article/821224

Salvador Oliván, J. A. (2008). *Recuperación de la Información*. Buenos Aires: Alfagrama.

Sarasa, A., Canabal, J. M., & Sacristán, J. C. (2008). Agrega-Spanish Education Community Federation Of Repositories Of Learning Objects. En *e-Learning* (pp. 47–50).

Sarasa, A., Canabal, J. M., Sacristán, J. C., & Jiménez, R. (2008). Uso de IMS VDEX en Agrega. En *X Simposio Internacional de Informática Educativa SIIE 2008* (pp. 119-124). Ediciones Universidad de Salamanca. Recuperado a partir de https://dialnet.unirioja.es/servlet/articulo?codigo=2874769

Sarasa-Cabezuelo, A., & Sierra, J.-L. (2013a). Grammar-driven development of JSON processing applications. En *Computer Science and Information Systems (FedCSIS), 2013 Federated Conference on* (pp. 1557–1564). IEEE. Recuperado a partir de http://ieeexplore.ieee.org/abstract/document/6644224/

Sarasa-Cabezuelo, A., & Sierra, J.-L. (2013b). The grammatical approach: A syntax-directed declarative specification method for XML processing tasks. *Computer Standards & Interfaces*, *35*(1), 114–131.

Sarasa-Cabezuelo, A., & Sierra, J.-L. (2015). A Syntax-Directed Model Transformation Framework Based on Attribute Grammars. En *International Symposium on Languages, Applications and Technologies* (pp. 145–152). Springer. Recuperado a partir de http://link.springer.com/chapter/10.1007/978-3-319-27653-3_14

Sarasa-Cabezuelo, A., Temprado-Battad, B., Rodríguez-Cerezo, D., & Sierra, J.-L. (2012). Building XML-driven application generators with compiler construction tools. *Computer Science and Information Systems*, *9*(2), 485–504.

Sarmah, A. K., Hazarika, S. M., & Sinha, S. K. (2015). Formal concept analysis: current trends and directions. *Artificial Intelligence Review*, *44*(1), 47–86.

schraefel, m. c., Wilson, M., Russell, A., & Smith, D. A. (2006). mSpace: Improving Information Access to Multimedia Domains with Multimodal Exploratory Search. *Commun. ACM*, *49*(4), 47–49. https://doi.org/10.1145/1121949.1121980

Schroeter, R., Hunter, J., Guerin, J., Khan, I., & Henderson, M. (2006). A synchronous multimedia annotation system for secure collaboratories. En *e-Science and Grid Computing, 2006. e-Science'06. Second IEEE International Conference on* (pp. 41–41). IEEE. Recuperado a partir de http://ieeexplore.ieee.org/abstract/document/4031014/

Schwertner, N., & Chavez, R. (2005). An approach to modeling content for digital repositories. En *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 373–373). Denver, CO, USA: IEEE. https://doi.org/10.1145/1065385.1065476

Sertkaya, B. (2009). Ontocomp: A protege plugin for completing owl ontologies. En *European Semantic Web Conference* (pp. 898–902). Springer. Recuperado a partir de http://link.springer.com/chapter/10.1007/978-3-642-02121-3_78

Sierra, J. L., & Fernández-Valmayor, A. (2008). Tagging Learning Objects with Evolving Metadata Schemas. En *2008 Eighth IEEE International Conference on Advanced Learning Technologies* (pp. 829-833). https://doi.org/10.1109/ICALT.2008.129

Sierra, J. L., Fernandez-Valmayor, A., Guinea, M., Hernanz, H., & Navarro, A. (2005). Building repositories of learning objects in specialized domains: the Chasqui approach. En *Fifth IEEE International Conference on Advanced Learning Technologies (ICALT'05)* (pp. 225-229). https://doi.org/10.1109/ICALT.2005.77

Sierra, J.-L., & Fernández-Valmayor, A. (2008). Tagging learning objects with evolving metadata schemas. En *Advanced Learning Technologies, 2008. ICALT'08. Eighth IEEE International Conference on* (pp. 829–833). IEEE. Recuperado a partir de http://ieeexplore.ieee.org/abstract/document/4561844/

Sierra, José Luis, & Fernández-Valmayor, A. (2006). A Heritage Dissemination Approach for the Production and Maintenance of Repositories of Learning Objects1. En *En Proceedings of the 8th. International Symposium on Computers in Education SIIE* (Vol. 6). Recuperado a partir de http://www.e-ucm.es/drafts/57.pdf

Referencias

Sierra, José Luis, Fernández-Valmayor, A., Guinea, M., & Hernanz, H. (2006). From Research Resources to Learning Objects: Process Model and Virtualization Experiences. *Journal of Educational Technology & Society*, *9*(3), 56-68.

Šimko, J., Tvarožek, M., & Bieliková, M. (2013). Human computation: Image metadata acquisition based on a single-player annotation game. *International Journal of Human-Computer Studies*, *71*(10), 933–945.

Slype, G. van, Hípola, P., & Moya Anegón, F. (1991). *Los lenguajes de indización: concepción, construcción y utilización en los sistemas documentales*. Fundación Germán Sánchez Ruipérez; Pirámide. Recuperado a partir de http://eprints.rclis.org/18372/1/Los_lenguajes_de_indizacion.pdf

Smith, Daniel A., Owens, A., Russell, A., Harris, C., Wilson, M., & others. (2005). The evolving mSpace platform: leveraging the Semantic Web on the Trail of the Memex. En *Proceedings of the sixteenth ACM conference on Hypertext and hypermedia* (pp. 174–183). ACM. Recuperado a partir de http://dl.acm.org/citation.cfm?id=1083391

Smith, Daniel Alexander, Owens, A., Sinclair, P., André, P., Wilson, M. L., Russell, A., … others. (2007). Challenges in supporting faceted semantic browsing of multimedia collections. En *International Conference on Semantic and Digital Media Technologies* (pp. 280–283). Springer. Recuperado a partir de http://link.springer.com/10.1007%2F978-3-540-77051-0_34

Smith, M., Barton, M., Branschofsky, M., McClellan, G., Walker, J. H., Bass, M., … Tansley, R. (2003). DSpace: An Open Source Dynamic Digital Repository. *D-Lib Magazine*, *9*(1). https://doi.org/10.1045/january2003-smith

Soon, K., & Kuhn, W. (2004). Formalizing user actions for ontologies. En *International Conference on Geographic Information Science* (pp. 299–312). Springer. Recuperado a partir de http://link.springer.com/chapter/10.1007/978-3-540-30231-5_20

Staples, T., Wayland, R., & Payette, S. (2003). The Fedora Project: An Open-source Digital Object Repository Management System. *D-Lib Magazine*, *9*(4). https://doi.org/10.1045/april2003-staples

Stumme, G., & Maedche, A. (2001). FCA-Merge: Bottom-up merging of ontologies. En *IJCAI* (Vol. 1, pp. 225–230). Recuperado a partir de http://www.academia.edu/download/30556812/fca01.pdf

Suleman, H., & Edward, A. (2002). Designing protocols in support of digital library componentization. En *International Conference on Theory and Practice of Digital Libraries* (pp. 568–582). Springer. Recuperado a partir de http://link.springer.com/chapter/10.1007/3-540-45747-X_43

Tabata, K., & Mitsumori, S. (2002). An assertion-based information-probe system: Document-skeleton and glossary-skeleton approach. *Information Knowledge Systems Management*, *3*(2-4), 123–152.

Tarrant, D., O'Steen, B., Brody, T., Hitchcock, S., Jefferies, N., & Carr, L. (2009). Using OAI-ORE to transform digital repositories into interoperable storage and services applications. *Code4Lib Journal*, *6*. Recuperado a partir de http://journal.code4lib.org/articles/1062?utm_source=feedburner&utm_medium=feed&utm_campaign=Feed:+c4lj+(The+Code4Lib+Journal)

Tazi, S., Al-Tawki, Y., & Drira, K. (2003). Editing pedagogical intentions for document reuse. *4th IEEE Technology Based Higher Education and Training*, 274–278.

Tello, A. L. (2001). Ontologías en la Web semántica. *Jornadas de Ingeniería Web*, *1*. Recuperado a partir de https://www.researchgate.net/profile/A_Lozano-Tello/publication/254438615_Ontologas_en_la_Web_Semntica/links/0f31753cccccc056e0000000.pdf

Ternier, S., Duval, E., Massart, D., Campi, A., Guinea, S., & Ceri, S. (2008). Interoperability for searching learning object repositories: the ProLearn query language. *D-Lib Magazine*, *14*(1), 1.

Tiropanis, T., Davis, H., Millard, D., & Weal, M. (2009). Semantic technologies for learning and teaching in the Web 2.0 era. *IEEE Intelligent Systems*, *24*(6). Recuperado a partir de http://ieeexplore.ieee.org/abstract/document/5372202/

Tunkelang, D. (2009). *Faceted Search*. Morgan and Claypool Publishers.

Van Assche, F., Anido-Rifon, L., Campbell, L. M., & Willem, M. (2003). *Controlled Vocabularies for Learning Object Metadata. Typology, impact analysis, guidelines and a web based Vocabularies Registry*. June.

Van de Sompel, H., Lagoze, C., Bekaert, J., Liu, X., Payette, S., & Warner, S. (2006). An interoperable fabric for scholarly value chains. *D-Lib Magazine*, *12*(10), 1082–9873.

Van Looy, J., & Baetens, J. (2003). *Close reading new media: Analyzing electronic literature* (Vol. 16). Leuven University Press. Recuperado a partir de https://books.google.es/books?hl=es&lr=&id=4aXd7ZJJk3oC&oi=fnd&pg=PA7&dq=+Close+reading+new+media:+Analyzing+electronic+literature&ots=SlhjGdm6db&sig=__UjVFZ-mUGG4yAzk9b3Ai_JhVE

Vargas-Vera, M., Moreale, E., Stutt, A., Motta, E., & Ciravegna, F. (2007). MnM: Semi-Automatic Ontology Population from Text. En R. Sharman, R. Kishore, & R. Ramesh (Eds.), *Ontologies* (pp. 373-402). Springer US. https://doi.org/10.1007/978-0-387-37022-4_13

Referencias

Vargas-Vera, M., Motta, E., Domingue, J., Lanzoni, M., Stutt, A., & Ciravegna, F. (2002). MnM: Ontology driven semi-automatic and automatic support for semantic markup. En *International Conference on Knowledge Engineering and Knowledge Management* (pp. 379–391). Springer. Recuperado a partir de http://link.springer.com/10.1007/3-540-45810-7_34

Völker, J., & Rudolph, S. (2008). Lexico-logical acquisition of OWL DL axioms. En *International Conference on Formal Concept Analysis* (pp. 62–77). Springer. Recuperado a partir de http://link.springer.com/chapter/10.1007/978-3-540-78137-0_5

Wang, H., Isenor, A., & Graybeal, J. (2011). Harmonization of Metadata Standards. Recuperado 5 de marzo de 2017, a partir de https://marinemetadata.org/guides/mdatastandards/crosswalks/harmonization

Wille, R. (1992). Concept lattices and conceptual knowledge systems. *Computers & mathematics with applications*, *23*(6-9), 493–515.

Wille, R. (2009). Restructuring lattice theory: An approach based on hierarchies of concepts. En *International Conference on Formal Concept Analysis* (pp. 314–339). Springer. Recuperado a partir de http://link.springer.com/chapter/10.1007/978-3-642-01815-2_23

Wray, T., & Eklund, P. W. (2010). Social tagging for digital libraries using formal concept analysis. Recuperado a partir de http://ro.uow.edu.au/infopapers/1504/

Xu, H., & Xiao, D. (2009). Building information specification ontology for computer network management based on formal concept analysis. En *Information and Automation, 2009. ICIA'09. International Conference on* (pp. 312–317). IEEE. Recuperado a partir de http://ieeexplore.ieee.org/abstract/document/5204941/

Xu, W., Li, W., Wu, M., Li, W., & Yuan, C. (2006). Deriving event relevance from the ontology constructed with formal concept analysis. En *International Conference on Intelligent Text Processing and Computational Linguistics* (pp. 480–489). Springer. Recuperado a partir de http://link.springer.com/chapter/10.1007/11671299_50

Yang, D. (2010). *Java persistence with JPA*. Outskirts Press. Recuperado a partir de http://dl.acm.org/citation.cfm?id=1841782

Yee, K.-P., Swearingen, K., Li, K., & Hearst, M. (2003). Faceted metadata for image search and browsing. En *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 401–408). ACM. Recuperado a partir de http://dl.acm.org/citation.cfm?id=642681

Zhang, Z., Li, W., Gurrin, C., & Smeaton, A. F. (2016). Faceted Navigation for Browsing Large Video Collection. En *International Conference on Multimedia Modeling* (pp. 412–417). Springer. Recuperado a partir de http://link.springer.com/chapter/10.1007/978-3-319-27674-8_42

Zhao, Y., Halang, W., & Wang, X. (2007). Rough ontology mapping in E-business integration. En *E-Service Intelligence* (pp. 75–93). Springer. Recuperado a partir de http://link.springer.com/content/pdf/10.1007/978-3-540-37017-8_3.pdf

Zheng, B., Zhang, W., & Feng, X. F. B. (2013). A survey of faceted search. *Journal of Web engineering*, *12*(1&2), 041–064.

Zobel, J., & Moffat, A. (2006). Inverted files for text search engines. *ACM computing surveys (CSUR)*, *38*(2), 6.