Recuperación retrospectiva de un archivo policíaco: el "*Casellario Politico Centrale*"[*]

Alessandro CHIARETTI - achiaretti@maas.ccr.it
Centro Maas srl, Roma, Italia - info@maas.ccr.it

1. Introducción

La utilización siempre más difundida de las tecnologías de la información amplía las posibilidades de comunicación y distribución de la información, pero también crea nuevos y frecuentes problemas de compatibilidad entre las distintas configuraciones de *hardware* y *software*, poniendo así limitaciones a las posibilidades de intercambiar y compartir las informaciones. Para evitar estos problemas, es necesario avanzar hacia la utilización de instrumentos que garanticen la llamada *platform indipendence*, o sea que las funciones realizadas por un computador sean lo más independientes posible del *hardware*, del sistema operativo y del *software* utilizado.

Los problemas relativos al intercambio de informaciones devienen un aspecto central en los procesos actuales de comunicación e involucran todos los ámbitos de aplicaciones, comprendido el archivístico, caracterizado por informaciones fuertemente estructuradas (Michetti, p. 7-8).

2. Objetivos y metodología

Mediante la presente exposición, trataré de mostrar algunas estrategias de utilización de estándares de descripción y tecnologías avanzadas que están en la base de las metodologías de recuperación retrospectiva de instrumentos de descripción archivística, experimentadas para superar las problemáticas asociadas a comunicar y compartir los recursos.

La expresión recuperación retrospectiva se refiere a la transposición de informaciones de cualquier soporte de origen a un soporte distinto. Ese paso mira a garantizar la integridad de los datos, su conservación en el tiempo y su reusabilidad, y permite atribuir a los mismos datos un valor adjunto potencial constituido por su inserción en un nuevo contexto de informaciones (Fig. 1) (Rendina, p. 87).

[*] *Paper presentado en el IV COINDEAR, 10-13 de abril 2012, San Bernardo, Chile.*

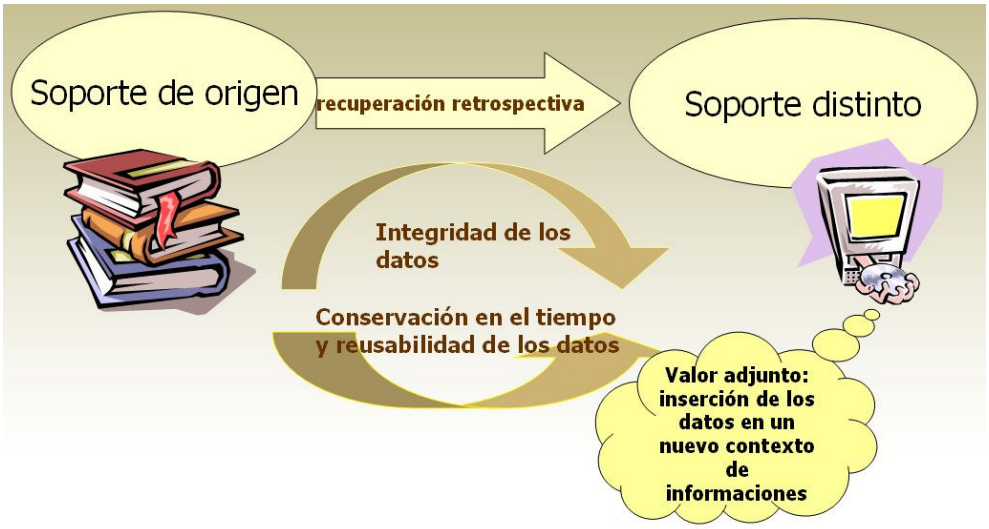

Fig. 1 – Modelo de Recuperación retrospectiva

Para que la transposición de datos estructurados de un soporte a otro se produzca sin perder informaciones, transportadas tanto directa como indirectamente por los datos (o sea deducibles de su contexto de informaciones originario), es necesario someterlos a una correcta operación de codificación descriptiva. Esto quiere decir que es necesario reconocer y describir todos los elementos significativos que conforman la estructura y el contenido de un "documento", a través de la introducción de códigos que sirven para identificar y circunscribir las estructuras de información que lo componen, señalando al mismo tiempo su naturaleza de manera explícita.

Cualquier operación de codificación tiene necesariamente que estar basada en tres actividades fundamentales: la individualización de las unidades mínimas del sistema a codificar; la elección del código al cual hacer referencia; la correlación entre los signos del código elegido y las unidades básicas del sistema sujeto a codificación.

El proceso de informatización de un texto necesita que sean explicitadas, en base a criterios formales adecuados, o sea comprensibles para el computador, todas esas informaciones transportadas a través de elementos distintos de los caracteres alfanuméricos. Esto es posible recurriendo a lenguajes formales, como los lenguajes de marcado (Orlandi, p. 26-54).

Uno de los instrumentos más eficaces en este ámbito es Xml (*eXtensible*

markup language)[1], un lenguaje no propietario, ya caracterizado por una enorme difusión y un amplísimo consenso, atento al perfil de la *platform indipendence*. Este lenguaje de codificación, basado en un marcado descriptivo, permite describir objetos estructurados jerárquicamente y crear una gramática formal, es decir una estructura lógica que refleja los componentes del documento y sus relaciones. Esta gramática, formada a través de *Document type definition* (Dtd) o *Xml Schema*, constituye la regla en base a la cual se analiza y valida cualquier instancia de documento asociada a ella.

Xml, enriquecido por soluciones y tecnologías paralelas, puede soportar la creciente necesidad de difundir documentos sofisticados y estructurados a través de la web. De hecho, una de las principales características de los documentos codificados a través de este lenguaje es el poder ser visualizados directamente a través de los *browser* comunes y, al mismo tiempo, poder ser gestionados e interrogados mediante un motor de búsqueda Xml. Además, con el auxilio de la tecnología paralela del Xsl[2], es posible obtener visualizaciones personalizadas de los mismos datos en distintos formatos de *output* (Xml, Html, Pdf etc.).

Como ulterior confirmación de la relevancia de Xml, se puede citar el segundo *Technology watch report*, publicado en el ámbito del proyecto europeo *Digicult*, que destaca como Xml y las tecnologías relacionadas pueden modificar profundamente la relación entre el sector de los bienes culturales y el mundo digital y mejorar los beneficios que los individuos y las empresas pueden obtener del patrimonio cultural (European commission).

Además del soporte de las nuevas tecnologías, que garantizan la integridad, la conservación a largo plazo y la reusabilidad de los datos, el otro aspecto fundamental del proceso de recuperación retrospectiva es el recurrir a un modelo estándar, al que hay que reconducir las clases de documentos para garantizar el acceso compartido y la interoperabilidad entre los distintos recursos digitalizados.

Considerando el ámbito archivístico, los modelos de referencia sin duda son

[1] Para mayores informaciones véanse las páginas dedicadas al desarrollo y a la implementación de Xml del W3C (http://www.w3.org/XML/).

[2] eXtensible Stylesheet Language (http://www.w3.org/standards/xml/transformation).

Ead[3] y Eac-Cpf[4] (formulaciones en lenguaje Xml de los estándares internacionales Isad e Isaar), a los cuales se agrega cada vez con mayor frecuencia Rdf-Owl[5], utilizado para la descripción de las informaciones de contexto.

Cuando la operación de recuperación se refiere a materiales en papel, las primeras fases del trabajo, que requieren bastante dedicación, son siempre la adquisición digital de los materiales, el reconocimiento óptico de los caracteres (Roc) y la codificación (más o menos automática) de los datos. A través de este proceso se proporciona a los datos una estructura que permitirá inserirlos provechosamente en el nuevo entorno.

Ocupándose directamente de materiales digitales, tampoco se pueden excluir dificultades en el proceso de recuperación, debidas principalmente a límites estructurales del *software* usado en origen o a datos incompletos. Además de este aspecto, no se puede subestimar la posibilidad de encontrarse con formatos de datos caracterizados por una cierta "opacidad", que puede dificultar una completa comprensión de su dimensión semántica, pero no sólo eso.

3. Resultados

Las varias fases de recuperación descritas en el presente trabajo se refieren a la base de datos del *Casellario Politico Centrale,* oficina dependiente del Ministerio del interior italiano, que tenía la tarea de administrar el fichero de los opositores políticos: anárquicos, republicanos, socialistas, pero también ociosos y vagabundos fueron objeto de una minuciosa actividad de vigilancia que alimentó un consistente archivo de más de 150.000 expedientes personales, conservado en el *Archivio Centrale dello Stato* de Roma.

La serie, con documentación comprendida principalmente entre 1894 y 1945, es una de las más consultadas y estudiadas del archivo. Las fichas de cada expediente, cada uno relativo a un "opositor", contienen, además de datos

[3] *Encoded Archivil Descripition*. El sitio oficial de Ead es http://www.loc.gov/ead/.

[4] *Encoded Archival Context*. El sitio oficial de Eac es http://eac.staatsbibliothek-berlin.de/.

[5] El *Resource Description Framework* (http://www.w3.org/TR/#tr_RDF) y el *Ontology Web Language* (http://www.w3.org/TR/owl-features/) son dos lenguajes de marcado sobre los que se basa la web semántica. A través de estas herramientas es posible representar y formalizar adecuadamente tanto las estructuras lógico-sintácticas como la semántica de los documentos. Owl, que técnicamente es una extensión del Rdf, es el lenguaje usado para describir de manera formal el conocimiento que se tiene de un determinado dominio, a través de las llamadas ontologías.

identificativos del expediente y de descripción sumaria de los documentos, informaciones de estado civil y domicilio (año y lugar de nacimiento, lugar de residencia, paternidad) e informaciones biográficas (trabajo y tendencia política) relativas a la persona.

La primera fase de la recuperación, llevada a cabo por el Centro Maas en colaboración con el *Archivio Centrale dello Stato* entre 2004 y 2005, se refirió a la transposición de los datos descriptivos de la serie de una base de datos en formato MSAccess a un formato Xml-Ead.

La operación de recuperación de los datos presentó dificultades debidas en su mayor parte a límites estructurales del *software* usado. En efecto, la base de datos fue creada anteriormente a través de la utilización de otro *software* que tenía viarias limitaciones (principalmente largueza limitada de los campos y campos no repetibles en la fila), que habían sido heredadas por la transposición en Access. Las informaciones descriptivas de los expedientes presentes en esta última versión, objeto de la recuperación, se encontraban entonces organizadas en dos tablas distintas, sin una clave unívoca de unión, y a veces había informaciones relacionadas con el mismo expediente colocadas en filas distintas. Después de una compleja actividad de análisis de los datos, fue posible unificar el conjunto de las informaciones en una única estructura en formato Xml. En un primer momento se codificaron los datos en base a una estructura creada para la ocasión (Dtd "local") que luego ha sido reconducida a la estructura Ead.

Una vez terminado el proceso de recuperación de los datos, se procedió a la creación de la aplicación web[6], que permite el acceso a las informaciones detalladas relativas a los expedientes a través de las tradicionales funciones de búsqueda textual. En particular, además de la búsqueda libre, se ofrece la posibilidad de hacer una compleja búsqueda avanzada también a través del auxilio de los diccionarios de los distintos campos informativos propuestos (denominación, lugar de nacimiento, lugar de residencia, tendencia política...).

La aplicación, basada sobre los componentes ExtraWay XML Engine[7] y

[6] Disponible en www.maas.ccr.it/cpc/. Funciona correctamente sólo con el browser Internet Explorer. Véase Fig. 2.

[7] http://www.3di.it/it/prodotti/tecnologia/extrawayxml.

Microsoft. NET[8], además de ofrecer la posibilidad de gestionar e imprimir los resultados de las búsquedas, permite al usuario visualizar directamente los datos en formato Xml-Ead.

Es necesario destacar como al término de esta operación de recuperación retrospectiva, los datos resultan estructurados de manera más coherente con respecto a la situación original y que, gracias a la codificación Xml-Ead, se puede alcanzar los objetivos de la fácil reutilización y de la conservación en el tiempo. Además de este aspecto, no hay que subestimar el valor agregado que ha proporcionado la elaboración del nuevo entorno de aplicaciones que, además de la publicación del sistema en la web, ha permitido una mejor y más sencilla posibilidad de interacción de los usuarios con los datos.

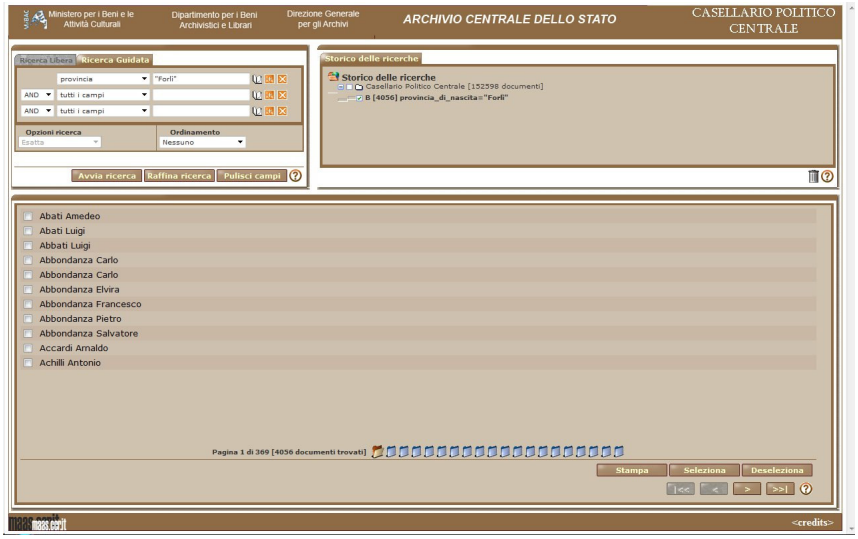

Fig. 2 – Aplicación versión 2005

La publicación del sistema lo expuso inevitablemente a la comparación continua con el estado de evolución de las tecnologías utilizadas en la web.

En los años siguientes, de hecho, se ha desarrollado en el ámbito informático una creciente sensibilidad por los aspectos de usabilidad de los sistemas y de los sitios web[9] (Hassan Montero). En particular, en base a los principios de la usabilidad, también la estructura general de un sitio o de una aplicación web reviste cierta importancia en favorecer el proceso de formación de un modelo mental del sistema por parte del usuario. A través de un diseño apropiado, es

[8] http://www.microsoft.com/net.

[9] En general, la usabilidad se preocupa de que el modelo mental de quién proyectó el *software* (*design model*), corresponda lo más posible al modelo mental del funcionamiento del *software* como se lo imagina el usuario final (*user model*).

posible ofrecer al usuario la oportunidad de prever con facilidad los resultados de una acción y facilitar la transformación de la información en conocimiento real.

Otro aspecto sobre el cual está surgiendo un notable interés es el relativo a la utilización en la web de los dichos sistemas de organización del conocimiento, o sea, los clásicos sistemas de indización semántica del ámbito bibliotecario, como tesauri, sistemas de clasificación y ontologías. Se considera que la interacción de los usuarios con sitios web y sistemas de información que utilizan sistemas de organización del conocimiento, resulta favorecida y más productiva con respecto a sistemas que utilizan exclusivamente las clásicas técnicas del *information retrieval* (Chiaretti)[10].

En particular, en este momento parece útil ilustrar la clasificación facetada, llamada también clasificación analítico-sintética, en base a la cual un objeto es descrito a través de un sistema de atributos mutuamente excluyentes, que representan cada uno un aspecto o una propiedad persistente del objeto. Tales atributos son llamados facetas y pueden ser utilizados como elementos de búsquedas individuales o en combinación con otros (Rosati).

El haber constatado que cada uno de nosotros busca de manera distinta, en base a las propias exigencias, y que el objetivo que guía nuestras búsquedas influye sobre la manera en que las efectuamos, nos lleva a apreciar el aspecto multidimensional de los sistemas facetados. La oferta de una multiplicidad de accesos a las informaciones basada en la posibilidad de elegir facetas distintas hace que estos sistemas sean extremadamente adaptables a las diversas exigencias de los usuarios.

En los últimos años en la web ha surgido un renovado interés en la potencialidad de la lógica facetada. En efecto, numerosos sitios adoptan, más o menos explícitamente, esta lógica, siendo ampliamente reconocida como herramienta de categorización capaz de ofrecer búsquedas más certeras e intuitivas, particularmente eficaces en ambiente digital (Murray) (Montero y Martín Rodríguez y Martín Rodríguez)[11].

---

[10] Para *Information retrieval* véase http://es.wikipedia.org/wiki/Recuperación_de_información.

[11] Como confirmación de este interés, también en el ámbito de las ciencias de la información, es útil citar el *Catalogo del Servizio bibliotecario nazionale italiano* (http://www.sbn.it/opacsbn/opac/iccu/free.jsp) o el Sistema Nacional de Bibliotecas Públicas chileno (http://www.bibliocatalogo.cl/) que proponen la utilización de las facetas como sistema

Con el objetivo de valorizar la importante cantidad de imágenes digitales adquiridas en el curso de los años y de inserir los datos en un nuevo entorno de aplicaciones, caracterizado principalmente por la utilización de sistemas de búsqueda semánticos, en sustitución de la complicada búsqueda avanzada textual, en el curso del 2011 se procedió a una nueva intervención de recuperación, con la que se obtuvo, como ulterior resultado positivo, un notable incremento de la usabilidad del sistema.

En lo que se refiere a la aplicación, se decidió abandonar totalmente el viejo entorno, basado en componentes propietarios, optando por instrumentos *open source* (Apache Lucene[12] y Apache Solr[13]). En cambio, en lo que concierne a los datos, se consideró todavía válida la elección hecha anteriormente, optando por la codificación Xml-Ead, que garantiza los requisitos de conservación en el tiempo y de interoperabilidad con otros sistemas informatizados.

La actividad se dirigió principalmente al desarrollo de la nueva aplicación[14]. Los principales aspectos novedosos de la aplicación pueden ser individualizados en la modalidad de búsqueda, a través de la utilización de las facetas, y en la exposición de los resultados.

Fig. 3 – Aplicación versión 2012, visualización *Cards*

para refinar la búsqueda.

[12] http://lucene.apache.org/.

[13] http://lucene.apache.org/solr/.

[14] La aplicación, aún no pública, está actualmente disponible en http://www.maas.ccr.it:8080/CPC/. Véase Fig. 3.

En lo que concierne a la búsqueda, a través de la selección de los valores presentes en las facetas, creadas en base al análisis de los campos informativos propuestos en la búsqueda avanzada de la vieja aplicación, el usuario es guiado en un proceso iterativo de refinamiento y expansión de la búsqueda, de modo de eliminar preliminarmente aquellas consultas que conduzcan a un resultado igual a cero.

Otro elemento novedoso ha sido introducido en lo concerniente a las búsquedas temporales.

Una línea de tiempo, a través de la cual es posible restringir los extremos cronológicos de las búsquedas simplemente moviendo un cursor, ha sido considerada como una modalidad de interacción más comprensible por los usuarios con respecto a los más tradicionales cuadros de texto. En este caso se realizaron dos líneas de tiempo, asociadas una a los extremos cronológicos de los expedientes y la otra a la fecha de nacimiento de los intestatarios de los mismos expedientes. En cambio, en relación a la posibilidad de búsqueda textual, además de un cuadro para la búsqueda simple, está presente la posibilidad de buscar los términos al interior de los valores presentados en cada faceta.

Para los resultados de la búsqueda, que se actualizan al mismo tiempo que las facetas seleccionadas, se ha tratado de evitar la presentación de listas planas, privas de organización a través de la oferta de modalidades innovadoras: una modalidad por *cards* individuales, que proponen una síntesis de los datos relativos al expediente personal; una presentación en tabla, en la que el usuario puede decidir cuales columnas visualizar y en base a qué criterio ordenar los resultados; una modalidad temporal, que permite la visualización gráfica de los datos sobre una línea de tiempo; una modalidad geográfica, en que, a través de la georreferenciación, efectuada en base al lugar de nacimiento del intestatario del expediente, se exponen los resultados posicionándolos en GoogleMaps (Fig. 4).

Para todas estas visualizaciones, seleccionando un expediente, es posible ver la ficha descriptiva completa de todos los datos relativos y, cuando estén presentes, también las imágenes digitales de los documentos contenidos en el expediente.

Además, para cada conjunto de resultados, se especifican las coordenadas

semánticas con respecto a las facetas del sistemas. Manteniendo el historial de las facetas y de los valores seleccionados se explicitan las relaciones, dinámicas y no anticipadamente previsibles, que se crean entre los conceptos (Sacco).

Fig. 4 – Aplicación versión 2012, visualización Mapa

4. Conclusiones

Analizando en su conjunto el resultado de la transposición de la base de datos del *Casellario Politico Centrale* en este nuevo entorno, resulta evidente el valor adjunto aportado: la exposición inmediata de todas las dimensiones informativas contempladas (representada por la clasificación de datos en facetas) y de sus efectivos valores, combinada con la posibilidad de reconducirlas fácilmente a planos espaciales o temporales de inmediata visualización gráfica, constituye una primera y fundamental fuente de comprensión completa del universo cognitivo ofrecido, así como un instrumento para encontrar rápida y eficazmente informaciones detalladas.

Se considera además útil poner en evidencia como, a partir de la codificación Xml aplicada a los datos en la primera intervención de recuperación, ha sido posible realizar, con tiempos y modalidades distintas, una serie de productos informáticos capaces de adaptarse a distintas exigencias y, por tanto, a distintos niveles de difusión y de fruición.

La misma metodología de trabajo, aplicada a distintos casos, ha confirmado el carácter central de la elección del formado Xml-Ead para la codificación de los

datos, en cuanto constituye un paso fundamental para garantizar la *platform indipendence* y la facilidad de reutilización de las informaciones en los sucesivos, e inevitables, procesos de recuperación.

A modo de conclusión, considero útil aludir a una reciente línea de desarrollo de las tecnologías de la web, que concierne a la web semántica (Berners-Lee 1998) y a los *Linked Open Data*[15] (Berners-Lee 2006) (Mazzo Uturriaga). La importancia de estas tecnologías, también en el ámbito de los bienes culturales, ha sido destacada en el *Final Report* del Library Linked Data Incubator Group del W3C (Library Linked Data). De hecho, a través de la adhesión a estos principios, y gracias al uso de los estándares de la web semántica, los datos se convertirían en recursos visibles y reusables en el universo de la web, también fuera del contexto original de utilización y completamente independientes de los *software* utilizados para su creación. De esta manera, se ofrecerá a los investigadores la posibilidad de conectar fuentes de distinto tipo y permitir nuevas conexiones entre sujetos, personas, organizaciones y lugares, con la finalidad de promover la investigación interdisciplinaria y enriquecer el conocimiento histórico. Parece definitivamente bastante claro que también los datos del *Casellario politico centrale* pueden encontrar una rentable colocación como *Linked Data*.

5. Referencias bibliográficas

BERNERS-LEE, T. *Semantic Web Road map.* 1998. [en línea] [fecha de consulta: 29 febrero 2012] Disponible en: http://www.w3.org/DesignIssues/Semantic.html

BERNERS-LEE, T. *Design Issues: Linked Data*, 2006. [en línea] [fecha de consulta: 29 febrero 2012] Disponible en: http://www.w3.org/DesignIssues/LinkedData.html

CHIARETTI, A. *Organización del conocimiento: la clasificación facetada como acceso a los contenidos archivísticos. Aplicación a un archivo fotográfico de empresa*, 2011. En IX Congreso de archivología del Mercosur, San Lorenzo, Paraguay, 16-18 nov. 2011. [en linea] [fecha de consulta: 29 febrero 2012] Disponible en: http://hdl.handle.net/10760/16394

[15] http://linkeddata.org/.

EUROPEAN COMMISSION, *Emerging Technologies for the Cultural and Scientific Heritage Sector, DigiCULT Technology Watch Report 2*, 2004, 216 p., ISBN 92-894-5276-5. [en línea] [fecha de consulta: 29 febrero 2012] Disponible en: http://www.digicult.info/downloads/twr_2_2004_final_low.pdf

HASSAN MONTERO, Y. Introducción a la Usabilidad. *No Solo Usabilidad*, 2002, n. 1, ISSN 1886-8592. [en línea] [fecha de consulta: 29 febrero 2012] Disponible en: http://www.nosolousabilidad.com/articulos/introduccion_usabilidad.htm

Library Linked Data Incubator Group Final Report. W3C Incubator Group Report 25 October 2011, 2011. [en línea] [fecha de consulta: 29 febrero 2012] Disponible en: http://www.w3.org/2005/Incubator/lld/XGR-lld-20111025/

MAZZO UTURRIAGA, R. *Linked Open Data: qué es y ejemplos en el mundo*, 20 octubre 2010. [en línea] [fecha de consulta: 29 febrero 2012] Disponible en: http://www.bcn.cl/de-que-se-habla/open-data-link-data

MICHETTI, G. Il linguaggio Sgml per la descrizione archivistica. *Archivi & Computer*, 2000, n. 1, 7-33 p.

MONTERO, Y.H. y MARTÍN RODRÍGUEZ, F.J. y MARTÍN RODRÍGUEZ, O. *Clasificaciones facetadas y metadatos (I): Conceptos basicos*, 2003. [en línea] [fecha de consulta: 29 febrero 2012] Disponible en: http://www.nosolousabilidad.com/articulos/clas_facetadas1.htm

MURRAY, P. *Faceted classification of information*. [en línea] [fecha de consulta: 29 febrero 2012] Disponible en: http://web.archive.org/web/20041204095504/http://www.kmconnection.com/DOC100100.htm

ORLANDI, T. *Informatica testuale. Teoria e prassi*, Bari: Laterza, 2010. 190 p. ISBN 978-88-420-9379-4

RENDINA, E. Strumenti di ricerca e trattamento informatico: la Guida generale degli Archivi di Stato italiani in formato XML. *Archivi & Computer*, 2003, n. 3, 85-96 p.

ROSATI, L. *La classificazione a faccette fra Knowledge Management e Information Architecture (parte I)*, 2003. [en línea] [fecha de consulta: 29 febrero 2012] Disponible en: http://www.itconsult.it/knowledge/articoli/pdf/itc_rosati_faccette_e_KM.pdf

SACCO, G.M. Dynamic taxonomies and guided searches. *Journal of the*

American Society for Information Science and Technology, 2006, vol. 57, n. 6, 792-796 p. [fecha de consulta: 29 febrero 2012] Disponible en: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.97.1510&rep=rep1&type=pdf