



IX CAM
9º Congreso de Archivología del Mercosur

Organización del conocimiento: la clasificación facetada como acceso a los contenidos archivísticos.

Aplicación a un archivo fotográfico de empresa

Alessandro CHIARETTI

achiaretti@maas.ccr.it

Roma, 14 de septiembre 2011

Resumen

Durante el año 2011 el Centro Maas ha desarrollado una investigación basada en la aplicación de sistemas de organización del conocimiento, en particular de la lógica de la clasificación facetada (o taxonomía dinámica), como modalidad de acceso a sistemas de información archivística. Este método es un sistema abierto y de adaptación, además de ser ampliamente reconocido como herramienta de categorización, capaz de ofrecer búsquedas muy eficaces e intuitivas.

Se ilustra una nueva aplicación open source para la presentación de datos en la web que ofrece al usuario una navegación personalizada, basada en el paradigma de búsqueda facetada. La aplicación permite navegar en el conjunto de datos a través de innovativos y variados puntos de vista, por ejemplo, geográfico (con mapas), temporal (con líneas de tiempo) y visualización a través de tabla.

Se indican en conclusión las posibilidades de desarrollo de este método hacia la web semántica.

1. Introducción

El *gap* semántico entre el *user model* (conceptos) y el *system model* (palabras usadas para expresar los conceptos) (Sacco), o sea la utilización de distintos sinónimos para la descripción del mismo objeto, es uno de los principales problemas relacionados con la recuperación de la información a través de la búsqueda textual simple. A éste se pueden agregar también las dificultades del usuario no experto en el dominio en objeto para utilizar una terminología adecuada en la cadena de búsqueda; la cantidad de resultados no relevantes entregados (el llamado “rumor”); la cantidad de documentos relevantes que al contrario no son entregados (que pueden llegar al 80%); la general falta de asistencia a los usuarios en la formulación de la *query* (Sacco).

El recurso a la llamada búsqueda avanzada, a través de la utilización también de manera repetida y combinada de la búsqueda en específicos campos informativos y de los relativos diccionarios de los términos, puede ayudar a solucionar, cuando menos en parte,



IX CAM

9º Congreso de Archivología del Mercosur

estos problemas. Sin embargo es fácil intuir la complejidad de estos mecanismos para un usuario no experto, que difícilmente obtendrá grandes beneficios de estas soluciones, terminando por utilizar sólo una mínima parte de los auxilios a la búsqueda que se le ofrecen.

A pesar de que estos problemas son bastante conocidos, las imponentes campañas de informatización y de difusión en la red de datos que caracterizaron los últimos años, contribuyeron a crear la ilusoria convicción de que cualquier búsqueda en la web se puede satisfacer con una simple correspondencia lexical entre términos, sin necesidad de recurrir a específicos instrumentos de gestión de la información. La hodierna atención a las disciplinas en evolución como la gestión de los contenidos y la arquitectura de la información, y el favor siempre creciente de iniciativas ligadas al desarrollo de la web semántica, pone en cambio en evidencia como la tecnología puede representar un soporte eficaz a la gestión del conocimiento, que, sin embargo, no elimina el problema de la organización del mismo conocimiento (Gnoli y Marino y Rosati).

Los principios de organización e indización semántica, resultados de más de un siglo de investigaciones y experimentaciones en ámbitos tradicionales como la bibliotecología, están ofreciendo un aporte fundamental también en estas disciplinas emergentes y se ponen, más en general, como estratégicamente relevantes con respecto a la estructuración de la información en el ambiente digital.

Para indicar el conjunto de herramientas de organización y indización semántica, independientemente de su ámbito de aplicación, hoy se utiliza la expresión Sistemas de organización del conocimiento (en inglés *Knowledge organization systems* o KOS). Los sistemas de este tipo más difundidos, que se distinguen principalmente por el nivel de complejidad y calidad de relaciones entre objetos, son vocabularios controlados, tesauros, esquemas de clasificación y ontologías.

Objetivo del presente trabajo es demostrar como la interacción de los usuarios con los sitios web y sistemas de información que presentan sistemas de organización del conocimiento, resulta favorecida y más productiva con respecto a sistemas que utilizan exclusivamente las clásicas técnicas de *information retrieval*¹. Se van a ilustrar experiencias de utilización de sistemas de clasificación facetada al interior de sistemas de información referibles en particular al dominio archivístico, desarrolladas por el Centro Maas durante el 2011.

Se pondrá atención de manera particular a la aplicación realizada para el acceso a la sección fotográfica del sitio web del Archivo histórico IRI², una de las principales fuentes para el estudio de la historia de la empresa italiana.

¹ http://es.wikipedia.org/wiki/Recuperación_de_información

² Istituto per la ricostruzione industriale (Instituto para la reconstrucción industrial), ente público italiano constituido en 1933 y cerrado en 2002 (http://it.wikipedia.org/wiki/Istituto_per_la_Ricostruzione_Industriale).



2. Revisión bibliográfica

Los esquemas de clasificación, también llamados taxonomías, se pueden dividir en dos tipologías. La más difundida es seguramente la clasificación jerárquica-enumerativa, que en el ámbito de la bibliotecología enumera todas las posibles clases, conectándolas entre ellas a través de relaciones de tipo jerárquico. Como alternativa, el bibliotecólogo y matemático indio S.M. Ranganathan ha desarrollado durante los años treinta la clasificación facetada, llamada también clasificación analítico-sintética. En este sistema el 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).

Muchas veces se alude a las facetas también como a una clasificación multidimensional (Gatto), poniendo en evidencia como un sujeto se subdivide en distintas coordenadas, cada una de las cuales expresa un aspecto. En este sentido, “el sujeto compuesto puede ser imaginado como ubicado a la intersección de las coordenadas-facetas, en un espacio a tantas dimensiones cuantas son las facetas” (Gnoli, p. 10). Esto significa, por ejemplo, que al describir una persona, en base al contexto en que se lo hace, se puede utilizar la faceta “título de estudio”, “ciudad de nacimiento” o “orientación política”: por ejemplo la persona Alessandro Pertini nació en > Stella (faceta Ciudad de nacimiento) > y es > Socialista (faceta Orientación política).

Numerosos estudios (Gnoli y Marino y Rosati, pp. 63-96) (Marino) (Montero y Martín Rodríguez y Martín Rodríguez) (Veen) se dedican al análisis de las problemáticas relacionadas con la aplicación de la clasificación jerárquica a sistemas publicados en la web y a la comparación con el más eficaz enfoque facetado.

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, opuesto al único criterio jerárquico ofrecido de la clasificación jerárquico-enumerativa, hace que estos sistemas sean extremadamente adaptables a las diversas exigencias de los usuarios.

En los últimos años en ámbito digital 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 eficaces e intuitivas con respecto a los sistemas de clasificación tradicionales, y particularmente eficaces en ambiente digital (Murray).

Las ventajas del método analítico-sintético son indudables también desde el punto de vista del usuario de un sistema. En efecto, no siempre quien hace una búsqueda se pone



IX CAM

9º Congreso de Archivología del Mercosur

el simple objetivo de encontrar un objeto específico (*known-item seeking*); muchas veces tiene también la exigencia de adquirir mayores informaciones o recursos sobre un determinado argumento (*exhaustive seeking*), o bien no sabe exactamente lo que está buscando (*exploratory seeking*) (Rosenfeld y Morville, pp. 31-32) (Gnoli y Marino y Rosati, p. 108). Precisamente las relaciones dinámicas e imprevisibles anticipadamente que se crean entre los objetos clasificados en base a las mismas facetas favorecen la *information discovery* (Sacco 2006) y la navegación contextual resulta más refinada y flexible.

3. Metodología

El Centro Maas ha desarrollado una creciente sensibilidad en el desarrollo de las propias aplicaciones para sistemas de archivos, a los problemas relacionados con la gestión del conocimiento y a la recuperación de la información. Esta situación ha llevado, después de unos años durante los cuales se siguieron las evoluciones de distintos sistemas de gestión de la información basados en el método facetado, a la activación entre 2009 y 2010 de una concreta actividad de experimentación sobre estos temas.

Entre las propuestas más innovadoras de sistemas *open source* se ha individuado Exhibit³, un *framework* AJAX⁴ nacido en el ámbito del Proyecto SIMILE⁵ del Massachusetts Institute of Technology. Este producto permite la creación de complejas páginas web que contienen visualizaciones ricas y dinámicas de datos estructurados, accesibles a través de una navegación personalizada basada sobre el paradigma de la búsqueda facetada.

Uno de los aspectos más relevantes de Exhibit es su facilidad de gestión, de modo que basta un conocimiento básico del lenguaje HTML⁶, sin la necesidad de intervenciones de sujetos especializados como programadores web o administradores de sistemas y sin necesidad de instalar software de ningún tipo⁷.

Terminada esta fase de experimentación con resultados positivos, se procedió durante el 2011 al desarrollo de un nuevo software, siempre con características *open*, con el objetivo de retomar y ampliar las principales funcionalidades presentes en Exhibit, de superar algunos límites estructurales del software y hacerlo utilizable en distintos contextos.

4. Resultados y Discusión

³ <http://www.simile-widgets.org/exhibit/>.

⁴ <http://es.wikipedia.org/wiki/AJAX>

⁵ SIMILE (*Semantic Interoperability of Metadata and Information in unLike Environments*) se compone de varios proyectos, todos *open source*, que tienen como objetivo común suministrar instrumentos específicos en el contexto de tecnologías del *semantic web*. <http://simile.mit.edu/>.

⁶ <http://es.wikipedia.org/wiki/Html>

⁷ Como confirmación de simplicidad de uso de este instrumento, se destaca que SIMILE pone a disposición de los usuarios exhaustivas y simples instrucciones para la creación de aplicaciones Exhibit, bajo forma de tutorial y wiki. Véase http://simile-widgets.org/wiki/Getting_Started_with_Exhibit.



IX CAM

9º Congreso de Archivología del Mercosur

Considerando la facilidad de acceso a los datos y la posibilidad de aprovechar las potencialidades de la búsqueda facetada, los resultados obtenidos en la creación experimental de sistemas Exhibit, que permiten utilizar datos archivísticos, se pueden considerar sin dudas positivos.

En particular, los datos inseridos en Exhibit, codificados en formato JSON⁸, resultan interrogables usando las facetas, que pueden tener aspecto de listas o de *tag cloud*⁹, y a través de la selección de los valores presentes en las facetas, el usuario es guiado a través de un proceso iterativo de refinamiento y expansión de la búsqueda, que nunca dará resultados iguales a cero. Los resultados de la búsqueda, que se actualizan al mismo tiempo que las facetas seleccionadas, son accesibles a través de distintas modalidades: visualización de ficha individual detallada; visualización de tabla, en la que se puede decidir el orden de los datos; visualización sobre una línea del tiempo, donde los objetos son expuestos en base a las coordenadas temporales; visualización sobre un mapa, donde los objetos georreferenciados son expuestos en GoogleMaps¹⁰.

Como demostración de su simplicidad está también el aspecto totalmente *client side*, que sin embargo se revela también como el principal límite a la escalabilidad del *framework*, en cuanto esta característica arquitectural lo hace incapaz de tratar cantidades consistentes de datos. Debido a este motivo las aplicaciones Exhibit fueron realizadas para archivos de pequeñas dimensiones o para subsistemas de archivos más consistentes.

Precisamente el límite de la escalabilidad de Exhibit estuvo en el origen del desarrollo del nuevo software por el Centro Maas.

Las principales diferencias entre la nueva solución y Exhibit son relativas a la arquitectura del software y a las tipologías de datos utilizables. El paso desde la arquitectura *client side*, hasta la *client-server*, permite, en efecto, superar el problema de la escalabilidad y ofrece la posibilidad de gestionar también grandes cantidades de datos. La infraestructura server utilizada, basada en el versátil motor de búsqueda Apache Lucene¹¹ y sobre la componente Apache Solr¹² para la generación de los índices, permite la importación de distintas tipologías de datos, entre los cuales Xml, Html, Word, Pdf y bases de datos relacionales.

Los componentes de presentación de la aplicación han sido pensados para una reutilización modular, por este motivo el mismo *template* de presentación puede ser utilizado para el acceso a bases de datos distintas sin intervenciones de programación,

⁸ <http://es.wikipedia.org/wiki/Json>.

⁹ http://es.wikipedia.org/wiki/Tag_cloud

¹⁰ <http://maps.google.com/>

¹¹ <http://lucene.apache.org/>.

¹² <http://lucene.apache.org/solr/>.



IX CAM

9º Congreso de Archivología del Mercosur

de los documentos o sobre el contenido de cada faceta. Ha sido además introducida una línea de tiempo, a través de la cual es posible hacer búsquedas en los arcos temporales en que están comprendidas las fotografías, simplemente moviendo el cursor. Se ha considerado esta modalidad de interacción más comprensible por los usuarios con respecto a los más tradicionales cuadros de texto.

Para la visualización de los resultados se ofrecen distintas alternativas: una modalidad de tabla, en la que el usuario puede decidir qué columnas visualizar y en base a cuál criterio ordenar los resultados; una modalidad temporal, que permite la visualización de las fotos puestas en una línea del tiempo según el año de realización; una modalidad geográfica, en la que, a través de la georreferenciación efectuada en base al lugar de la foto, se presentan los resultados colocados en GoogleMaps; una visión detallada, donde se presentan fichas individuales completas de todos los datos descriptivos, que pueden ser ordenados según distintos criterios.

Estas posibilidades de interacción y visualización responden a la necesidad de mejorar la posibilidad del usuario de navegar entre los resultados, actividad que en los sistemas tradicionales frecuentemente se hace difícil por la presentación de listas planas y sin una clara organización.

Durante el desarrollo de la aplicación IRI se realizaron, personalizando el mismo ambiente, sistemas para otros archivos. Esta actividad permitió obtener una confirmación de la efectiva adaptabilidad del ambiente aplicable a contextos diferentes y también de la capacidad de gestión de grandes cantidades de datos, tratándose en un caso de una aplicación que gestiona una base de datos de más de 150.000 records¹⁵.

Además se desarrollaron distintas funciones adicionales necesarias para mejorar el uso del sistema en casos específicos. Las más relevantes son probablemente la introducción de facetas que gestionan una jerarquía en los datos, particularmente adaptas por ejemplo en el caso de informaciones de tipo geográfico (país > región > ciudad), y la inserción de una ulterior posibilidad de visualización temporal, que representa gráficamente la evolución de fenómenos temporales a través del uso de un *Timeplot*¹⁶.

El resultado más inmediato y relevante de esta solución es que el usuario queda en condiciones de aprovechar plenamente las sofisticadas posibilidades de búsqueda avanzada que el sistema le ofrece y, al mismo tiempo, prever con facilidad los resultados de la acción efectuada.

6. Conclusiones

¹⁵ Aún no pública, está actualmente disponible en <http://www.maas.ccr.it:8080/CPC>.

¹⁶ Esta componente es una incorporación del widget Timeplot de Simile (<http://www.simile-widgets.org/timeplot/>).



IX CAM

9º Congreso de Archivología del Mercosur

Analizando en general el resultado de la incorporación de las distintas aplicaciones a este nuevo ambiente, resultan evidentes las ventajas con respecto a los sistemas más tradicionales. 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 poder 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.

Resulta bastante simple establecer las principales direcciones en que la actual aplicación puede desarrollarse para aumentar las propias capacidades de gestión de sistemas complejos, no exclusivamente de tipo archivístico.

Una primera posibilidad de desarrollo, ya en parte explorada, se refiere a la introducción en la planta actual de modalidades de gestión, además de la clasificación facetada, de los otros sistemas de gestión del conocimiento, en particular tesauros¹⁷ y ontologías¹⁸. Las funcionalidades de una faceta de tipo jerárquico podrían ser desarrolladas para la gestión de verdaderos tesauros, mientras la gestión de datos de descripción del contexto histórico o institucional, expresados formalmente a través de ontologías, podría ser fructíferamente expuesta a través de la utilización integrada de facetas, líneas temporales y mapas geográficos. Estas implementaciones harían todavía más cercana la realización de la deseada web semántica (Berners-Lee).

Otro aspecto sobre el que son necesarias profundizaciones es el relativo a la capacidad del método facetado para representar también archivos con una estructura jerárquica muy profunda. En efecto, como ya expliqué, la base de datos objeto de esta intervención no presenta una relevante estructura jerárquica interna, situación bastante frecuente en los archivos fotográficos. Esta característica hace posible la presentación de objetos como entidades individuales independientes y sin distinción de tipología o nivel, exigencia que en cambio se hace más relevante en caso de representaciones de archivos fuertemente estructurados. A través de experimentaciones en curso se está tratando de superar este problema a través de distintos modos, como por ejemplo la inserción de una faceta que permita la selección del nivel de los objetos sobre los cuales hacer la búsqueda, la posibilidad de diferenciar gráficamente los objetos de distinto tipo o la exposición de la estructura jerárquica en la que se encuentra el objeto en distintos momentos de la navegación.

¹⁷ Un thesaurus es un vocabulario controlado en que se establecen relaciones de distintos tipos (jerárquico, de equivalencia, asociativas) entre los términos. <http://es.wikipedia.org/wiki/Tesauro>

¹⁸ Una ontología permite describir de manera formal el conocimiento que tenemos de un determinado dominio. Se definen distintas tipologías de relaciones entre clases e instancias individuales pertenecientes a las clases, creando así, más que un árbol jerárquico, una compleja red de relaciones. [http://es.wikipedia.org/wiki/Ontología_\(informática\)](http://es.wikipedia.org/wiki/Ontología_(informática))



IX CAM

9º Congreso de Archivología del Mercosur

Los problemas de uso y de correcto diseño de las interfaces son aspectos determinantes para facilitar la búsqueda y la utilización de la información y su transformación en conocimiento real. A este propósito una interesante perspectiva de desarrollo está seguramente representada por el proyecto *Elastic list*, que se presenta como una “fluida y rica interfaz para la navegación facetada” (Stefaner y Muller). Las características principales de este proyecto son la inmediata visualización gráfica de la relevancia, o sea del número de apariciones, de las varias voces en cada faceta, las transformaciones animadas a medida que se procede a los sucesivos refinamientos de la búsqueda y la visualización de los llamados gráficos *sparkline*¹⁹ para determinadas facetas.

7. Referencias bibliográficas

BERNERS-LEE, T. *Semantic Web Road map*. 1998. [en línea] [fecha de consulta: 1 septiembre 2011] Disponible en: <http://www.w3.org/DesignIssues/Semantic.html>

GATTO, E. A righe o a quadretti?... En *Le dimensioni dell'informazione. Attualità della classificazione a faccette tra biblioteconomia e Web, Giornata di studio AIB Piemonte-ISKO Italia. Torino, Biblioteca civica Geisser 12 giugno 2004*. BARAZIA, C. y GNOLI, C. comp. 2004. [en línea] [fecha de consulta: 1 septiembre 2011] Disponible en: <http://www.iskoi.org/doc/dimensioni4.htm> (ultima consultazione 20 agosto 2011).

GNOLI, C. *Classificazione a faccette*. Roma: Associazione italiana biblioteche, 2004. 44 p. ISBN - 88-7812-080-4

GNOLI, C. y MARINO, V. y ROSATI, L. *Organizzare la conoscenza. Dalle biblioteche all'architettura dell'informazione per il Web*. Milano: Tecniche Nuove, 2006. 211 p. ISBN - 88-481-1849-6

MARINO, V. *Classificazioni per il Web. I vantaggi dell'adozione di schemi a faccette*. AIB-WEB, 2004, [en línea] [fecha de consulta: 1 septiembre 2011] Disponible en: <http://www.aib.it/aib/contr/marino1.htm>

MONTERO, Y.H. y MARTÍN RODRÍGUEZ, F.J. y MARTÍN RODRÍGUEZ, O. *Clasificaciones facetadas y metadatos (I): Conceptos básicos*. 2003. [en línea] [fecha de consulta: 1 septiembre 2011] Disponible en: http://www.nosolousabilidad.com/articulos/clas_facetadas1.htm

MURRAY, P. *Faceted classification of information*. [en línea] [fecha de consulta: 1 septiembre 2011] Disponible en: <http://web.archive.org/web/20041204095504/http://www.kmconnection.com/DOC100100.htm>

ROSATI, L. *La classificazione a faccette fra Knowledge Management e Information Architecture (parte I)*. 2003. [en línea] [fecha de consulta: 1 septiembre 2011] Disponible en: http://www.itconsult.it/knowledge/articoli/pdf/itc_rosati_faccette_e_KM.pdf

ROSENFELD, L. y MORVILLE, P. *Information Architecture for the World Wide Web*, Second Edition. Milano: Hops, 2002. 416 p. ISBN - 88-8378-062-0

¹⁹ <http://en.wikipedia.org/wiki/Sparkline>



IX CAM

9º Congreso de Archivología del Mercosur

SACCO, G.M. *Guided Interactive Discovery of e-Government Services*. [en línea] [fecha de consulta: 1 septiembre 2011] Disponible en: <http://www.slideshare.net/guest66dc5f/giovanni-maria-sacco>

SACCO, G.M. Dynamic taxonomies and guided searches. *Journal of the American Society for Information Science and Technology*, 2006, vol. 57, no.6, 792-796 p. [fecha de consulta: 1 septiembre 2011] Disponible en: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.97.1510&rep=rep1&type=pdf>

STEFANER, M. y MULLER, B. *Elastic lists for facet browsers*. 2007. [en línea] [fecha de consulta: 1 septiembre 2011] Disponible en: <http://moritz.stefaner.eu/downloads/papers/Stefaner-2007-Elastic%20lists%20for%20facet%20browsers.pdf>

VEEN, J. *Faucet Facets: A Few Best Practices for Designing Multifaceted Navigation Systems*. 2 junio 2002. [en línea] [fecha de consulta: 1 septiembre 2011] <http://www.adaptivepath.com/ideas/e000034>