De la recuperación de documentos a la recuperación de información en los archivos: Estudio de las técnicas de recuperación que aparecen en la especificación MOREQ.

RAQUEL GÓMEZ DÍAZ - RAQUEL BRINGAS GONZÁLEZ
F. Traducción y Documentación.
Universidad de Salamanca

RESUMEN.

La rápida evolución tecnológica actual está empujando a los archiveros a tomar conciencia de la utilidad de las nuevas tecnologías, así como de las técnicas de recuperación automatizada de información en el desarrollo de su trabajo.

Comenzaremos analizando rápidamente este entorno de cambio que incluye factores como la evolución de los soportes, la forma de trabajar o las nuevas exigencias de los usuarios, para enmarcar el campo de desarrollo de esta comunicación.

Posteriormente y al hilo de la reciente

"presentación en sociedad"de la revisión de la traducción del *Modelo de requisitos para la gestión de documentos electrónicos de archivo,* (en adelante la especificación MoReq), trataremos de enmarcar la recuperación automatizada de información y documentos en el terreno de la archivística. El estudio está encaminado a la clarificación de los conceptos relativos a este tema, recogidos en el capítulo octavo de la especificación MoReq.

1. LOS ARCHIVOS ANTE UN ENTORNO DE CAMBIO.

Podemos afirmar que el desarrollo tecnológico ha influido de forma determinante en las tareas que constituyen la vida diaria del archivo. La evolución que se ha tenido que llevar a cabo ha generado nuevas formas de trabajar, e inevitablemente, "nuevos, mejores, distintos" archiveros, que se encuentran ante el reto de tener que adaptarse a los nuevos formatos de los documentos y a "nuevos formatos" de usuarios. Junto al investigador erudito que frecuentaba los archivos, ahora también se acercan a nuestros centros otro tipo de usuarios no tan especializados que recurren a nosotros incluso desde un entorno virtual, gracias a tecnologías de la comunicación que favorecen la consulta, uso e intercambio de información dispersa y remota. Parece que lo que no está en la red no existe, por tanto, los archivos, celosos siempre de su contenido, tendrán que hacerse visibles en este nuevo entorno si quieren dar servicio a nuevas y viejas demandas.

Hoy prácticamente ningún archivero se atreve a decir que los ordenadores no mejoran o facilitan su trabajo. Hemos pasado del papel a la tecnología Blu-ray en pocos años incrementando exponencialmente en la capacidad de almacenamiento y si bien en un principio los ordenadores fueron utilizados como herramientas auxiliares, tras la aparición de los documentos electrónicos y de los nuevos soportes se han convertido en útiles imprescindibles.

La evolución interior se ve empujada y acompañada por la evolución externa, por usuarios que exigen a los archivos mayores prestaciones y la optimización de los recursos disponibles. La manera en que desde los archivos damos respuesta a la demanda informativa requiere nuevas técnicas y herramientas de gestión que proporcionen soluciones ágiles.

Una de las técnicas más relevantes en el entorno documental cuya utilidad se ha visto potenciada con el uso de las herramientas informáticas, es la recuperación de información. Al hilo de la reciente "presentación en sociedad"de la revisión de la traducción del *Modelo de requisitos para la gestión de documentos electrónicos de ar-*

chivo[1] (en adelante especificación MoReq), trataremos de enmarcar la recuperación automatizada de información y documentos en el terreno de la archivística. La comunicación está encaminada a la clarificación de los conceptos relativos a este tema recogidos en el capítulo octavo de dicha especificación.

2. LA RECUPERACIÓN EN LOS ARCHIVOS.

Garantizar derechos y deberes es la razón de ser de la mayor parte del trabajo archivístico. Custodiamos, conservamos, clasificamos, ordenamos y describimos documentación con el fin de facilitar el acceso a la misma, siendo la localización y recuperación de los documentos o de la información contenida en ellos, tal y como reseñamos en el apartado anterior, una de las actividades que más ha variado gracias a las nuevas tecnologías.

La manera tradicional de acceder al contenido de los documentos de archivo es a través de los instrumentos de descripción. Las guías nos proporcionan datos acerca de los fondos, los inventarios la información de las series y los catálogos la de los documentos. Además mediante los índices podemos hacer búsquedas cronológicas, topográficas... La normalización y la irrupción de la informática, han generado un doble cambio en este terreno. Por una parte, ha evolucionado el modo de preparación, actualización, y consulta de estas herramientas tradicionales, proporcionando una mayor versatilidad, facilidad de modificación y rapidez de localización de los documentos. Por otro lado, es posible acceder al contenido de los documentos sin necesidad de utilizar instrumentos de descripción como puerta de entrada, haciendo el proceso más complejo y más fácil a la vez.

2.1. La Recuperación Automatizada de la Información.

Por Recuperación de Información (en adelante R.I.) entendemos el proceso por el cual, una vez analizado el documento e identificada la necesidad de información, se produce una comparación entre ambos para obtener resultados satisfactorios para el usuario[2] .

Según esta definición el documento ha de estar procesado para la recuperación. Las técnicas de preparación, en las que no vamos a entrar, podrán ser más o menos complejas y las podremos denominar catalogación, descripción archivística, o simplemente tratamiento documental. Lo único importante en este caso es señalar que esta preparación condicionará el proceso de recuperación y por lo tanto el resultado de la búsqueda.

Desde los años 50 en que comenzaron a desarrollarse las técnicas automatizadas de R.I., son muchos los autores que han investigado este tema. En un principio, más que de R.I. se debería hablar de recuperación de documentos, ya que según la definición de Mooers[3] el resultado del proceso de búsqueda era una *"lista de citas de documentos almacenados"*, más que la propia información en sí, concepto que también se ajusta a la recuperación en archivos. En este sentido es importante el matiz que introduce Rijsbergen[4] quien dice que *"el sistema de recuperación de la información no informa, no cambia el estado del conocimiento del usuario en la materia que está preguntando, sólo informa de la existencia o no existencia y del paradero de los documentos relativos a una pregunta"*. Estos conceptos hoy en día han quedado un tanto desfasados, y si lo aplicamos a los soportes electrónicos, lo que espera el usuario es que en la recuperación se incluya el enlace al propio documento o a una versión digital del mismo, con lo que la distancia entre recuperar una lista de referencias y los documentos o la información que contienen es menor.

3. LA R.I. EN LA ESPECIFICACIÓN MOREQ.

La especificación Moreq describe un modelo para la gestión de documentos electrónicos de archivo, incidiendo especialmente en los requisitos funcionales de los sistemas de gestión de documentos electrónicos de archivo (SGDEA). A pesar de este título, también incluye directrices para la gestión de documentos en soporte tradicional, haciendo extensibles a éstos la mayor parte de sus recomendaciones.

El modo de acceso al documento está íntimamente relacionado con su soporte, así, en el caso de los documentos en soporte tradicional, no podemos acceder directamente, sino que necesitamos un paso intermedio en el que la pregunta se compare con la representación del mismo, lo que en la especificación MoReq se identifica con metadatos. En el caso de los documentos electrónicos este paso intermedio se puede obviar y podremos buscar directamente en el documento en la mayoría de los casos.

---

[1] *Modelo de requisitos para la gestión de documentos electrónicos de archivo. Especificación MoReq.* Bruselas: CEE, 2001. Consultado 30-08-2004 http://www.cultura.medc.es/archivos/oa/pdf.moreq.pdf

[2] Gómez Díaz, Raquel. *Estudio de la incidencia del conocimiento lingüístico en los sistemas de recuperación de la información para el español* [CD ROM] Salamanca : Universidad de Salamanca, 2002 ISBN 84-7800-831-4

[3] Spink, A.; Losse, R. M. "Feedback in Information Retrieval". En"*Annual Review of Information Sciencie and Tecnology.* 1996, v. 13 pp. 31-81

[4] Rijbergen, K. V. *Information Retrieval.* 2nd print. London : Butterworth, 1979

Dentro del campo informático, el término metadatos fue acuñado en los años 60 en referencia a los datos que proporcionan información sobre otros datos gestionados dentro de una aplicación o contexto determinado, pudiendo incluir información descriptiva sobre el *contexto, calidad, condiciones y características de los datos*[5]. En el glosario de la especificación MoReq, se define metadato como la *"información estructurada o semiestructurada que permite la creación, gestión y utilización de documentos de archivo a lo largo del tiempo, tanto dentro de los ámbitos en que se crearon como entre ellos"*[6]. Esta definición difumina la barrera existente entre soporte informático y otros soportes, unificando la terminología[7].

El capítulo octavo de la especificación está íntegramente dedicado a los requisitos de la búsqueda y la recuperación de expedientes y documentos de archivo y su posterior presentación, sea en pantalla o mediante una copia impresa, desarrollando en la primera parte el conjunto de características que debería incluir un SGDEA dotado de "buenos instrumentos de recuperación".

Como archiveros puede resultarnos muy útil el conocimiento de ciertos conceptos y métodos de recuperación. Estas nociones nos ayudarán a comprender por qué en unos sistemas se recupera más rápido o con mayor éxito que en otros, facilitándonos la evaluación, el uso y la realización de propuestas de mejora de nuestros propios sistemas.

3.1. Las técnicas de recuperación.

Apenas hay estudios de recuperación asociados a los entornos archivísticos, ya que la mayor parte de los trabajos de este tema o bien son teóricos y explican los modelos puros que rara vez se aplican en la práctica, o se centran en otras áreas documentales.

Aunque con una visión excesivamente simplista, entre los requisitos MoReq encontramos referencia a algunas de las técnicas que se pueden emplear en recuperación de información y documentos. Vamos a tratar contextualizarlas archivísticamente y de explicarlas, aunque sea de forma breve, indicando el punto concreto de la especificación en que se mencionan, puesto que su desarrollo no es objeto de la especificación.

En el capítulo octavo se incide en las funcionalidades de búsqueda que debe presentar el SGDEA, haciendo referencia a la inclusión de una gama flexible de opciones [8.1.1.] que permitan la recuperación a través de parámetros definidos por el usuario. Aunque la recuperación no haya sido siempre considerada una de las funciones clásicas de la gestión de documentos, esta cuestión está incluida en la especificación que indica cómo se tiene que relacionar el usuario con el sistema, es decir, cómo hace la pregunta y como el sistema proporciona la respuesta. Partiendo de la afirmación previa en relación a los soportes y el modo de acceso a la información, las búsquedas [8.1.1; 8.1.4.; 8.1.5; 8.1.6.][8] las podremos hacer a través del propio documento o de la representación del mismo, sea ésta generada de manera automática o voluntariamente

En realidad, un modelo de recuperación de información es un concepto mucho más amplio e incluye el modo en el que el SGDEA realiza la gestión de una consulta, es decir cómo lleva a cabo las operaciones de búsqueda. En estas cuestiones de implementación práctica la especificación se mantiene al margen ya que no forman parte de su objetivo, pero es un aspecto que no podemos olvidar y de ello depende, en gran medida, la calidad de los resultados obtenidos.

Uno de los primeros aspectos abordados en la especificación es el lenguaje de acceso a los documentos, que puede ser natural o controlado. No podemos considerarlo como una técnica de recuperación en sí misma, sino que es la forma en la que está consignado el documento, su representación (metadatos) o la pregunta, sirviendo para establecer la comparación entre ellos.

La especificación MoReq habla de búsquedas en texto libre [8.1.8.; 8.1.9], expresión que desde el punto de vista de la recuperación, podemos entender en dos sentidos. Por una parte, la búsqueda en texto libre es aquella que se hace en información no estructurada en campos; en este caso, sólo podremos hacer búsquedas automatizadas sobre documentos o metadatos en soporte electrónico. También podemos entender esta expresión como sinónimo de lenguaje natural, es decir el lenguaje que no requiere tratamiento documental, y que se caracteriza por ser más flexible y tener un vocabulario más rico, en contraposición al lenguaje controlado[9]. La principal ventaja que tiene la búsqueda en este tipo de lenguaje, es que se reduce considerablemente el tiempo de

[5] Howe, Denis *FOLDOC Free Online Dictionary* [Documento HTML] Foldoc, 1993. Citado por Méndez Rodríguez, Eva. Metadatos y recuperación de información: estándares, problemas y aplicabilidad en bibliotecas digitales. Gijón: TREA, 2002

[6] Glosario de términos de la especificación MoReq

[7] Por tanto, según la especificación, se considera metadato cualquier representación del documento, por ejemplo, una descripción elaborada con ISAD(G), una descripción tradicional o la información generada automáticamente al crear un documento informático.

[8] Entre corchetes se indica la referencia a los requisitos de la especificación.

[9] Pinto Molina, María. *Análisis documental. Fundamentos y procedimientos.* 2ª ed. Madrid: EUDEMA, 1993

preparación de los documentos para la recuperación al no necesitar ser convertido a un lenguaje controlado, pero el gran inconveniente es que necesitamos que la consulta contenga exactamente los mismos términos que aparecen en el documento o en la representación del mismo.

Además del lenguaje libre, en la especificación también se habla de los lenguajes controlados, en concreto de los tesauros [8.1.10]. La ventaja que aporta el tesauro a la recuperación es la precisión en la búsqueda, pero el principal inconveniente radica en el excesivo tiempo de preparación que implica la elaboración del mismo. Este objetivo puede lograrse también, aunque no con unos resultados tan afinados, mediante otros lenguajes controlados como las listas de materias, autoridades... Hay algunos estudios, aún en fase experimental, que tratan de aplicar tesauros a la recuperación en archivos[10] [11] y que nos darán un punto de valoración para elegir una herramienta u otra en función de su rentabilidad.

Una técnica muy extendida de recuperación es la aplicación del álgebra de Boole [8.1.8.] para la combinación de los elementos en la pregunta. Belkin[12] en su clasificación de técnicas de recuperación la denomina "*modelo de coincidencia exacta*". Esta técnica combina los términos mediante los operadores "y", "o", "no", que además pueden combinarse con paréntesis para darle una mayor flexibilidad.

No deben obviarse los problemas que plantean los operadores booleanos. En primer lugar, es necesario que los términos de la pregunta sean exactamente iguales a los que aparecen en el documento y ésta siempre se plantea en valores absolutos (presente / ausente), sin reflejar la importancia de los términos en el contexto exigiendo un alto grado de precisión en los términos utilizados. Además requieren claridad en la composición de las expresiones a buscar, por eso habitualmente lo que se hace es combinar esta técnica con otras como los truncamientos, los operadores de proximidad o de comparación para darle una mayor flexibilidad a las búsquedas, tal y como se recomienda la especificación MoReq. Por último hay que tener en cuenta que este tipo de búsquedas no nos ofrece una respuesta ponderada, es decir, la salida no presenta ningún orden en relación a la relevancia de los términos.

Como técnica que permite reflejar en la pregunta la necesidad de que varios términos no sólo aparezcan en el documento -resultado que como hemos visto podemos conseguir con los operadores booleanos- sino que además estén muy próximos o juntos (coocurrencia de términos) disponemos de los operadores de proximidad ("*cerca*", "*seguido*") [8.1.12]. Su utilización nos posibilita afinar más las búsquedas obteniendo así una mayor precisión.

Los operadores booleanos y de proximidad se complementan a su vez con los operadores de comparación y aunque éstos explícitamente no aparecen en la especificación, consideramos conveniente explicarlos aquí ya que muchos sistemas SGDEA los tienen implementados.

Los operadores de comparación nos permiten especificar el rango de búsqueda fijando unos límites para la misma. Pudiéndose expresar de manera alfabética ("mayor que", "menor que", "igual que", o "distinto de") o con el correspondiente signo ("<", ">", "=" y "<>"), se utilizan principalmente en documentos que contienen datos numéricos, y son especialmente prácticos cuando queremos buscar por fechas o intervalos temporales.

Finalmente, mencionar los truncamientos como herramienta que podemos utilizar independientemente o como complemento de las técnicas anteriores. Se aplican generalmente sobre el lenguaje natural y el procedimiento consiste en sustituir un carácter por otro"– que se denomina comodín— o un conjunto de ellos por un solo carácter, generalmente un asterisco "*". La ventaja de esta técnica es que ahorra tiempo, con una sola búsqueda podremos agrupar un conjunto de ellas, de forma que si introducimos en el sistema "corregi*", éste nos buscará "corregidor", "corregidores", "corregimiento", "corregimientos". El principal inconveniente es que introduce un alto factor de ruido al recuperar también aquellos términos que los que coincidan los caracteres ("corregimos", "corregir") a pesar de que no exista una relación semántica. En la especificación se habla de comodines [8.1.11], pero en realidad es una referencia a los truncamientos tal y como son mencionados en la mayor parte de la literatura.

En este capítulo deberían incluirse al menos algu-

[10] Rubio Liniers, María Cruz "El análisis de contenido y el control del vocabulario en los archivos municipales. Principales problemas y posibles soluciones". En"*XV Jornadas de Archivos Municipales: la descripción multinivel en los archivos municipales: la norma ISAD (G).* Madrid: Consejería de Cultura y Deportes de la Comunidad de Madrid. Dirección General de Archivos, Museos y Bibliotecas, 2004

[11] Grupo de Archiveros Municipales de Madrid. Materiales para un Tesauro de Archivos Municipales. 2ª ed. Madrid: Consejería de Cultura de la Comunidad de Madrid, 2000.

[12] Belkin, N.J. ; Bruce, C.W "Retrieval Techniques""*Annual of Information Science Technology.* 1987. v. 22 pp. 109-145

nos comentarios relativos a la evaluación como etapa final de las fases de la recuperación[13]. Teniendo en cuenta que la especificación ni tan siquiera menciona los distintos aspectos susceptibles de valoración y, por tanto, tampoco ninguna de las maneras posibles de hacerlo, hemos preferido obviar el tratamiento de este tema a pesar de considerarlo un punto de especial relevancia.

4. CONSIDERACIONES FINALES.

El trabajo del archivero cobra sentido pleno en el momento de la recuperación de información y documentos, una de las diversas facetas de la archivística que han dado un paso de gigante bajo el auspicio de las nuevas tecnologías. La especificación MoReq respalda esta realidad al incluir este aspecto como una funcionalidad imprescindible en un SGDEA, dedicando un capítulo completo a los requisitos necesarios para disponer de un buen sistema de búsquedas.

Consideramos que el archivero debe conocer las posibilidades de implementación de estas herramientas, conocimientos que deben completarse con un estudio más profundo de cómo el programa procesa la información contenida en el sistema, la relaciona con la pregunta planteada y devuelve una salida en forma de respuesta que debe cumplir de la mejor manera posible con las expectativas del usuario

BIBLIOGRAFIA.

❑ BelkKin, N.J. ; Bruce, C. W. "Retrieval Techniques". *Annual of Information Science Tecnology*. 1987. v. 22 pp 109-145

❑ Gómez Díaz, Raquel. *Estudio de la incidencia del conocimiento lingüístico en los sistemas de recuperación de la información para el español* [CD ROM] Salamanca : Universidad de Salamanca, 2002 ISBN 84-7800-831-4

❑ Grupo de Archiveros Municipales de Madrid. Materiales para un Tesauro de Archivos Municipales. 2ª ed. Madrid: Consejería de Cultura de la Comunidad de Madrid, 2000.

❑ Méndez Rodríguez, Eva. Metadatos y recuperación de información: estándares, problemas y aplicabilidad en bibliotecas digitales. Gijón: TREA, 2002

❑ Pinto Molina, María. *Análisis documental. Fundamentos y procedimientos*. 2ª ed. Madrid: EUDEMA, 1993

❑ Rijsbergen, K. V. *Information Retrieval*. 2nd print. London : Butterworth, 1979

❑ Spink and R. M. Losse. Feedback in Information Retrieval. *Annual Review of Information Sciencie and Tecnology*, 1996 vol 13 p. 31-81

❑ Tramullas, Jesús *Introducción a la Documática* Zaragoza: Kronos, 1997.

❑ *XV Jornadas de Archivos Municipales: la descripción multinivel en los archivos municipales: la norma ISAD (G).* Madrid: Consejería de Cultura y Deportes de la Comunidad de Madrid. Dirección General de Archivos, Museos y Bibliotecas, 2004

[13] Gómez Díaz, Raquel. op. cit.