Aplicaciones de descripción documental y recuperación de la información en el entorno del fin de siglo y en el marco de las bibliotecas públicas

Alejandro Delgado Gómez
Ayuntamiento de Cartagena

0.1 Resumen

El Ayuntamiento de Cartagena está llevando a cabo un programa de investigación en nuevas tecnologías que pretende fomentar la interoperabilidad entre diferentes centros de gestión de la información, así como entre diferentes tecnologías de la información, con la finalidad de facilitar las tareas de los profesionales de la información y el acceso amigable del ciudadano a diversos ítems documentales e informativos, a nivel local y remoto. El programa diseña un sistema de descripción local, mediante el uso de formatos MARC, y de etiquetado remoto, mediante lenguajes de metadatos. También permite la recuperación mediante el uso tanto de OPACs locales como del protocolo Z39.50 y medios de difusión genérica de información.

Palabras clave: Sistemas de información. Archivos. Museos. Bibliotecas. Descripción. Recuperación de información. Formatos de descripción. Lenguajes de metadatos. Algebra de Boole. Redes neurales. Norma ANSI/NISO Z39.50.

0.2 Abstract

The Cartagena City Council is developing a research project on new information technologies, with the aim of promoting interoperability between different information management centres, as well as between different information technologies, in order to make easier tasks of information professionals and an user-friendly access to information items, at a local and remote level. The project is carrying out an in-house description system, using the MARC formats, as well as a tagging system, using metadata languages. It also allows retrieval of information through the use of OPAC's and Z39.50 protocol, and tools to disseminate information in a more generic way.

Keywords: Information System. Archives. Museums. Libraries. Description. Information Retrieval. Description Formats. Metadata Languages. Boolean Algebra. Neural Nets. ANSI/NISO Z39.50 Standard.

1. Introducción

En primer lugar, debe señalarse que el trabajo que a continuación se expone es fruto del esfuerzo de un amplio equipo del que el comunicante es, en este caso, únicamente portavoz. De igual modo, cabe precisar que, respondiendo al título de la comunicación, se hablará de manera casi exclusiva de bibliotecas públicas, aunque el programa de investigación del que depende el presente trabajo parte del supuesto de que la integración de distintos servicios de información es deseable, posible y necesaria. Sirvan de apoyo los más de 30.000 documentos que pueden hallarse en las bases de datos de la Unión Europea, si se realiza una búsqueda no selectiva. El hecho de priorizar las bibliotecas sobre otros servicios tiene un fundamento puramente práctico: históricamente, las bibliotecas han avanzado más que los archivos, los museos u otros servicios de información en lo referente a tareas de descripción y recuperación. Pero cuanto se diga en el presente texto de aquéllas valdrá también para éstos.

Justifíquese también de manera previa la relevancia que el presente texto concede a la descripción. Ésta, entendida como descripción local, al modo tradicional, pierde progresivamente importancia. Aun así, creemos que la estructura que subyace a la descripción debe conocerse, no sólo para que el profesional éste al corriente del esqueleto de su trabajo, sino sobre todo porque la catalogación bibliotecaria y el formato MARC han alcanzado en el curso de la historia un grado de solidez tal que permite su utilización como base para la integración de la descripción procedente de diversos ámbitos, tal y como pretende el programa de nuevas tecnologías. Por decirlo con sencillez, el programa precisa una estructura sólida y simple que permita hacer muchas cosas, y esta estructura la proporcionan la norma ISO 2709 y los formatos MARC.

El trabajo que se lleva a cabo en el marco del programa de nuevas tecnologías del Servicio de Archivos, Bibliotecas y Documentación del Ayuntamiento de Cartagena parte de un conjunto de evidencias. La primera de ellas se ha repetido hasta la saciedad, de manera que su enunciado resulta suficiente: existe, en las sociedades occidentales tardo-modernas, una inflación de recursos informativos de distinto tipo, sobre diferentes soportes, ubicados en diversos lugares y recuperables mediante múltiples procedimientos. Si no se establece un orden en tales recursos informativos, éstos devienen inasimilables y, por tanto, inutilizables por el ciudadano.

La segunda evidencia es: la administración local se encuentra más próxima al ciudadano que otras administraciones y, en consecuencia, está en mejores condiciones potenciales de proporcionar a éste la información que precisa.

La tercera evidencia, en cierto modo derivada de o paralela a la segunda, sería: la propia administración local genera una enorme cantidad de información distribuida en diferentes servicios o unidades que en numerosas ocasiones duplican su trabajo.

En cuarto lugar, existe la evidencia de que no todos los ciudadanos potenciales beneficiarios de la recepción de información disponen de las mismas destrezas para ello ni cuentan con los medios sociales, culturales, económicos o intelectuales encaminados a tal fin. A ello debe sumarse el hecho de que no es deseable que toda la información se recupere de la misma manera: en los procesos de recuperación intervienen factores legales, administrativos, técnicos o estéticos. A modo de ejemplo, el programa de investigación ha concluido que un PDF es el mejor procedimiento para recuperar información administrativa, técnica o la generada en un proceso de gestión electrónica de la documentación. Pero cuando se quiso aplicar el mismo sistema a los resultados de la descripción museológica, se advirtió que resultaba inadecuado, incómodo e incluso visualmente desagradable para el usuario.

Por último, a pesar de que tradicionalmente se han hecho esfuerzos para diferenciar tipos, soportes o servicios de información, éstos presentan más similitudes que diferencias. Al menos, presentan la similitud de que todos ellos almacenan y gestionan documentación y la información en ella contenida, con el fin de suministrarla a quienquiera que se muestre interesado en ella. No resulta, por tanto, descabellado plantear la posibilidad de integración de las tareas de descripción y recuperación de información procedente de diversos ámbitos.

El programa de investigación en nuevas tecnologías pretende, como consecuencia de las anteriores evidencias, crear o promover un procedimiento de descripción y recuperación de información que facilite, por una parte, la integración de los diferentes servicios dedicados a ello y, por otra, la accesibilidad del ciudadano a diferentes niveles.

En lo que sigue se exponen, en filigrana, los pasos proyectados por el programa de nuevas tecnologías del Servicio de Archivos, Bibliotecas y Documentación del Ayuntamiento de Cartagena, y se explican las realizaciones obtenidas hasta el momento.

2. El Programa de Nuevas Tecnologías del Servicio de Archivos, Bibliotecas y Documentación del Ayuntamiento de Cartagena

Tal y como ha sido definido para el período 1997-2000, el plan de trabajo sobre nuevas tecnologías es el que a continuación se expone, no sin antes hacer la salvedad de que la siguiente exposición es puramente analítica y diacrónica, es decir, muchas de las tareas enunciadas son simultáneas y no pueden darse unas sin otras.

2.1. Fase de entrada de datos

2.1.1. Diseño de un programa de descripción para diversos soportes informativos, basado en los cinco formatos USMARC y en las recomendaciones del ICA y del ICOM.

El hecho de haber seleccionado los formatos USMARC se debe a la vocación integradora de éstos, que permite homogeneizar la descripción de distintos tipos de documentos, así como al énfasis puesto en el uso de enlaces, autoridades, referencias, etc., de manera tal que con la mayor economía de medios posible pueda obtenerse la mayor cantidad posible de información relevante. No son ajenas a la elección de USMARC circunstancias tales como la posibilidad de describir documentos electrónicos, mediante el campo 856; la posibilidad de describir información no soportada necesariamente por un determinado material, haciendo uso del formato para descripción comunitaria; o la posibilidad de contextualizar la información descrita, mediante el formato de holdings. De los formatos USMARC, sin embargo, se han eliminado, sin alterar la estructura de la norma ISO 2709, los campos y subcampos que sólo resultan pertinentes en el contexto norteamericano. Por otra parte, y dado que a pesar de su riqueza los formatos USMARC no satisfacen plenamente las necesidades de archiveros y museólogos, al menos en el contexto europeo, ha parecido procedente elaborar dos formatos adicionales para documentación de archivo y de museo. Estos formatos respetan la estructura de la norma ISO 2709 y, con el fin de garantizar la homogeneidad, se ajustan, hasta donde es posible al formato general USMARC para la descripción bibliográfica, suprimiendo sin embargo los campos y subcampos irrelevantes e incorporando los campos que el ICA, en ISAD(G), considera pertinentes, en el caso de los archivos, y, en el caso de los museos, los campos qu el ICOM, a través del CIDOC y de otros organismos o programas asociados —*International Guidelines for Museum Object Information: The CIDOC Information Categories, The AFRICOM Handbook of Standards, CIDOC Core Data Standard for Archaeological Sites and Monuments, y Visual Resource Association Core Categories for Visual Resources*— considera pertinentes. El cuerpo de formatos de descripción se compone, pues, de los siguientes: descripción bibliográfica, descripción archivística, descripción museológica, *holdings*,

información comunitaria, autoridades y datos de clasificación. Todos los formatos son homogéneos y, partiendo del supuesto de que los dos últimos se utilizan adecuadamente, debe ser posible para el ciudadano recuperar, desde un solo punto, información tan diferente como, por ejemplo, monografías y publicaciones periódicas sobre pintura, los fondos de un determinado pintor que posee un Ayuntamiento, el expediente por el que estos fondos fueron legados o la dirección y horario de apertura de una asociación de amantes de la pintura. De los siete formatos enunciados, en este momento se encuentran en funcionamiento los de descripción bibliográfica, archivística, museográfica y de autoridades. Los tres restantes se han finalizado pero aún no se encuentran operativos, salvo, parcialmente, el de *holdings*. El formato de datos de clasificación no es utilizable; sin embargo, a los programas de descripción en uso, se ha incorporado un fichero de autoridades de clasificación.

2.1.2. Diseño de un lenguaje de metadatos que permita la incorporación de las anteriores descripciones a una red.

Excluídos los lenguajes genéricos de metadatos, y dado que no siempre resultará necesaria una recuperación exhaustiva de información, el programa de nuevas tecnologías propone tres opciones, de uso alternativo, y cuya viabilidad se encuentra pendiente de contrastación, habida cuenta del creciente número de facilidades de conversión de lenguajes:

2.1.2.1. *Dublin Core*: La opción simple está basada en Dublin Core, que, como se sabe, es un lenguaje de metadatos sencillo y extremadamente pertinente. A niveles primarios de recuperación de información y en contextos donde prime la integración sobre la especialización, tanto su sencillez como su pertinencia lo convierten en una herramienta útil; sin embargo, a niveles más complejos de recuperación y en contextos donde prime la especialización sobre la integración, Dublin Core resulta insuficiente o debe sufrir modificaciones *ad hoc* que no se compadecen con lo que se entiende por un lenguaje bien definido de metadatos. En lo fundamental, y siempre a nuestro juicio, el principal problema de Dublin Core es que resulta demasiado simple, mientras que la información a tratar y recuperar no siempre lo es. No obstante, o quizá por ello, ha alcanzado una notable popularidad y un alto nivel de difusión.

2.1.2.2. *MARC-DTD*: La opción compleja está basada no en uno, sino en cuatro lenguajes de metadatos: los dos lenguajes MARC-DTD, para descripción bibliográfica, información comunitaria y *holdings*, por una parte; y para autoridades y clasificación, por otra. EAD, para las descripciones archivísticas, y CHIO, para las descripciones museológicas. EAD, según declaración de su principal responsable, Daniel Pitti, persigue la incorporación de algunos de los logros más relevantes de MARC, por ejemplo, independencia del software y del hard-

ware, etiquetado de diferente nivel, estructuración de la información, información descriptiva y no procedimental o exportabilidad a otros sistemas. CHIO respeta la filosofía genérica de SGML y adopta inicialmente la estructura de TEI, lenguaje destinado al etiquetado de documentos procedentes del área de humanidades, dejando la puerta abierta a futuras modificaciones. Los MARC-DTD, elaborados en Perl, permiten la conversión de registros MARC a SGML y a la inversa, de manera extremadamente intuitiva, al simplificar el uso de atributos como, por ejemplo "Nombre" o "Tipo"; o al hacer uso de una cómoda sintaxis apoyada en SGML. Tanto EAD, desarrollado para etiquetar instrumentos de descripción archivística, como CHIO, desarrollado para etiquetar mediante SGML información museológica y que ha incorporado además un perfil Z39.50, resultan bastante más complejos. Además, CHIO, si bien susceptible de recibir otros usos, fue concebido para etiquetar instrumentos de difusión, más que instrumentos de descripción en sentido estricto.

En realidad este paso bien pudiera haberse reducido al uso de un solo lenguaje de metadatos, MARC-DTD, que, fundado, aunque no de manera explícita, en los formatos USMARC, resulta completo e intuitivo. Pero presenta la grave carencia de no reconocer contextos locales, lo que significa, entre otras cosas, que los formatos de descripción archivística y museológica hubieran quedado fuera de contexto, a menos que sobre ellos se llevaran a cabo profundas modificaciones. Así las cosas, el uso de EAD y de CHIO, a pesar de su extrema complejidad y de su ubicación en cierto modo artificiosa dentro del proyecto, parece inevitable.

2.1.2.3. HTML, XML/RDF: La tercera opción no es sino un estudio de la aplicabilidad de los lenguajes genéricos de metadatos -HTML, XML/RDF- a contextos específicos.

A estas opciones han de sumarse dos problemas transversales: el de la conversión de series de caracteres no latinos y el del etiquetado de imágenes.

Este paso ha sido finalizado, pero aún no es de uso público, por un doble motivo. En primer lugar, la realización del siguiente paso obligaría a cuestionar la validez del presente; en segundo lugar, y como ya se adelantó, la solución propuesta en cuanto al uso de lenguajes de metadatos diríase excesivamente complicada y merece un análisis detallado de las posibilidades de economizar medios.

2.1.3. Diseño de un cliente/servidor Z39.50.

Como se sabe, y la propia norma explicita, Z39.50 es un protocolo que especifica formatos y procedimientos que gobiernan el intercambio de mensajes entre cliente y servidor, posibilitando que el cliente pida al servidor una búsqueda en una base de datos y la identificación de los registros que satisfagan ciertos criterios, así como la recuperación de algunos o todos los registros identificados.

Dado que Z39.50 es una norma de aplicación genérica, se han desarrollado también perfiles determinados de diferente tipo. Este paso se encuentra en período de discusión, por dos motivos. El primero de ellos, su implantación obligaría a cuestionarse, como se ha dicho algo más arriba, la validez de los lenguajes de metadatos: desde un punto de vista documental, tal y como se está desarrollando la discusión en el programa de nuevas tecnologías, los lenguajes de metadatos son útiles, puesto que permiten el etiquetado de información para usos no locales, así como la estructuración de aquélla con fines de control, navegación y uso; desde un punto de vista informático, la norma ANSI/NISO Z39.50 los convierte en innecesarios, en la medida en que la accesibilidad queda garantizada a través de ella. En este momento expreso una opinión estrictamente personal, y de este modo pido que se entienda. A mi juicio, los lenguajes de metadatos y la norma ANSI/NISO Z39.50 son herramientas diferentes y útiles cada una de ellas en su contexto; pero la metaprogramación es un proceso terriblemente complicado y anti-económico, de manera que me parece legítimo cuestionar hasta qué punto es rentable la realización de un esfuerzo doble. Con todo, no debe olvidarse que Z39.50 es un protocolo de comunicaciones, es decir, un conjunto de requisitos a cumplir para que se establezca contacto entre cliente y servidor a través de una *gateway*. Un lenguaje de metadatos es, simplemente, una herramienta para estructurar datos atómicos en la web.

El segundo motivo de discusión acerca de la implantación de Z39.50 reside en la necesidad de decidir acerca de las múltiples posibilidades de la norma. Como se sabe, existen, al menos, dos versiones oficiales de la misma, cada una con distintas capacidades, no siempre necesarias; al tiempo que se han generado derivados que pretenden mejorar la norma, no siempre con igual fortuna. También aquí expreso un punto de vista personal: si la vocación del programa de nuevas tecnologías tiende a la integración y a la economía de medios, entonces creo que un diseño genérico cliente/servidor ha de ser suficiente.

Por otra parte, aunque han sido muchas las modificaciones que la norma ha sufrido desde su origen, éste es esencialmente bibliotecario, como sugeriría, por ejemplo, la serie de atributos de uso Bib, lo que justificaría en principio su integración en el proyecto.

2.2. Fase de recuperación de la información

2.2.1. Diseño de un buscador avanzado para uso local.

Diseño de un buscador avanzado de información, al que podríamos llamar de primera generación, puesto que se basa, por una parte, en el álgebra convencional de Boole, herramienta lógica con una semántica y una sintaxis muy simples pero de gran potencia; y, por otra, en un proceso de indización humana. Aunque el programa de nuevas tecnologías incorpora numerosas utilidades para ayudar al usuario, éstas no son otra cosa que derivaciones del álgebra de Boole, de manera que, en lo sustancial, seguimos dentro de los buscadores de primera generación. Además, este buscador ha sido ideado, con independencia de futuras aplicaciones, para uso local, no para su empleo en Internet. Este paso ya ha finalizado y se encuentra parcialmente operativo y asociado a los procesos de descripción mencionados en la fase 1. Algo más adelante intentaremos explicar su funcionamiento.

2.2.2. Optimizador basado en el algoritmo de Som

A efectos estrictamente teóricos, y con el objeto de incrementar la eficacia en la recuperación de información mediante el uso de Internet, investigación acerca de un algoritmo basado en redes neurales y algoritmos genéticos y, de manera específica, en el algoritmo de Som. Como se sabe, este concepto no es novedoso; de hecho, se viene utilizando desde hace cierto tiempo en contextos no documentales. Pero la creciente presencia de instituciones documentales en la web, así como los fallos de pertinencia y exhaustividad de los buscadores tradicionales, han forzado la recuperación del concepto. En filigrana, el algoritmo de Som es un procedimiento para distribuir o mapear unidades o neuronas sobre una rejilla bidimensional, de tal manera que las unidades semejantes se encuentren cerca unas de otras y las desemejantes más alejadas. Teóricamente, ésto debe permitir que el usuario que lleva a cabo búsquedas difusas o ambiguas recupere información relevante sin necesidad de atenerse a una sintaxis estricta. En este momento se está discutiendo la mejor manera de llevar a cabo este paso, aunque la investigación reconoce su dependencia de los trabajos de Timo Honkela. En cualquier caso, en este mismo congreso el profesor Alonso ha explicado con mayor profundidad de la que podemos hacer uso ahora el concepto de redes neurales.

2.2.3. Diseño de un OPAC amigable

En relación con la recuperación local de información citada en el paso 1 de la fase 2, diseño de un OPAC, basado en las recomendaciones de la IFLA, capaz de hacer amigable la recuperación de información por parte del usuario. En el último borrador de la IFLA, el diseño de OPAC's se aleja cada vez más del modelo de asiento bibliográfico y, dependiendo de la finalidad perseguida, se aproxi-

ma a la idea de los buscadores convencionales en Internet, por una parte, y, por otra, a la disposición convencional de términos en un tesauro. Este paso ya ha finalizado y se encuentra operativo, aunque está siendo sometido a revisión para mejorar la accesibilidad. En lo fundamental, se pretende precisamente que en un OPAC de uso en una red de área local se pueda utilizar el álgebra de Boole de manera similar a como se hace en un buscador convencional, con las adiciones que el programa presenta, en principio escasamente intuitivas para un cliente general, aunque susceptibles de optimización con el fin de optimizar paralelamente las búsquedas.

2.3. Fase de protección de información

Aunque la protección de la información no se relaciona inmediatamente con las tareas de descripción y recuperación, una de las premisas fundamentales del programa de nuevas tecnologías es que la información se encuentra contextualizada, como se intentará explicar más adelante, de manera que no parece inoportuno mencionar, siquiera en filigrana, esta fase del programa.

2.3.1. Digitalización de la documentación

La digitalización de la documentación que corre peligro de desaparición se realizará con un doble fin: conservar los originales que merecen ser conservados, o conservar la información, si los originales no son recuperables. Este paso se está verificando en áreas pequeñas (en concreto el centro de documentación de literatura juvenil), para avanzar posteriormente a áreas más amplias. Se hace uso para ello de scanners de sobremesa convencionales, aunque la empresa asociada al proyecto utiliza cámaras fotográficas en otras unidades cuya documentación tiene más valor, en cuanto al soporte físico se refiere. El señor Mascordá, en su ponencia sobre edición electrónica, distinguía fotografía de transcripción. Nuestra propuesta se basa en el escaneado como base para la trasncripción, en el caso, claro está, de obras literarias sin valor físico de patrimonio histórico.

2.3.2 Edición electrónica de los documentos digitalizados.

El procedimiento que ha parecido más adecuado es la creación de un fichero padre en texto plano, para convertirlo después al formato pertinente (HTML, Adobe, Word, compiladores de ayuda, etc.). En lo que dice, por ejemplo, bien que modesto, al centro de documentación de literatura juvenil citado, un sólo fichero ha servido recientemente para crear una exposición virtual de obras de literatura fantástica haciendo uso de Hcrtf y Shed, los compiladores de ayuda de Visual FoxPro 5; para editar un catálogo y un boletín impresos mediante Adobe Pagemaker y el paralelo PDF; y para preparar una página, escrita evidentemente mediante HTML, con vistas a su futura inclusión en la web.

2.4. Fase de difusión de datos

Los pasos llevados a cabo anteriormente deben permitir:

a) Que el ciudadano pueda consultar desde un solo OPAC información de diverso tipo de manera amigable.

b) Que la información introducida, tanto en la fase de descripción como en la de edición electrónica, sea recuperable también mediante Internet o una intranet.

c) Que al ciudadano le sea posible diseñar de manera cómoda páginas de información, comodidad que se entiende en un doble sentido: general, orientado al turismo, el rastreo, etc.; y especial, orientado a la formación de grupos especiales de clientes (jóvenes, mujeres, educación de adultos, etc.)

Esta fase aún no ha sido desarrollada, aunque se está experimentando en el segmento de jóvenes.

2.5. Fase de utilidades

2.5.1. Herramientas

Los pasos anteriores deben completarse con herramientas de introducción de datos: recomendaciones, diseño de páginas de presentación, thesauri. Este paso se está llevando a cabo en la actualidad por etapas, habiendo terminado, por ejemplo, el diseño de presentación de descripciones en pantalla, las recomendaciones para introducir documentos de archivo o un thesaurus de música popular para jóvenes. En este punto, cabría decir, en cierto sentido, que lo urgente se sobrepone a lo importante.

2.5.2. Microtesauros

Precisamente por lo que hace a la elaboración de thesauri o algún otro tipo de autoridades, ha parecido aconsejable la fragmentación o limitación temática por el siguiente motivo: a pesar de la vocación integradora del programa de nuevas tecnologías, no se puede negar que persisten notables diferencias en cuanto a soportes documentales, servicios de información o incluso tipos de clientes. Si en un proceso de descripción física estas diferencias pueden minimizarse, se acentúan sin embargo en un proceso de indización. Resultaba, pues, necesario, optar entre un thesaurus o fichero de autoridades faraónico y con voluntad de abarcarlo todo, o un conjunto de micro-thesauri o ficheros de autoridades independientes pero elaborados de manera coherente, a partir de la norma ISO 2788:1986, con el fin de dejar la puerta abierta a una futura integración. Esta segunda solución es la que nos ha parecido de sentido común.

2.5.3. Programa de estadiística, evaluación y gestión de la calidad

Diseño de un programa de estadística, evaluación e indicadores de rendimiento, basado en ISO 11620:1998. Este paso se encuentra en fase de desarrollo, y ya se encuentran operativas las opciones estadísticas cuyos datos se pueden obtener inmediatamente de la gestión diaria de una biblioteca, es decir, aquellos relativos a la circulación, los clientes o la adquisición en sus diversas formas. De igual modo, se encuentran operativos algunos procedimientos o conceptos estadísticos destinados a la elaboración de informes. Sin embargo, aún no es posible medir la aspectos de extrema relevancia para el programa, como, por ejemplo, la calidad o la velocidad de la descripción. Asociado a la fase de indicadores de rendimiento, se ha desarrollado también un procedimiento de encuesta, el llamado SERVQUAL, basado en los trabajos de Zeithaml, Parasuraman y Berry -citados ya por la profesora Pinto en su ponencia sobre gestión de calidad- para la empresa privada, que, si bien algo fuera de fecha, dado que procede del año 1990, posee la virtud de medir no sólo la opinión del cliente, sino el grado de divergencia entre ésta y la opinión de directivos y empleados, es decir, en definitiva, SERVQUAL mide los errores de percepción. Este procedimiento ha sido finalizado, pero aún no se encuentra operativo.