XIII

La documentación del grupo Vocento

Mauricio Martín

Jefe de documentación del Grupo Vocento
Vocento Taldeko Dokumentazio Burua

Laburpena: Hurrengo ponentzian, Vocento taldeko hedabideek erabiltzen duten dokumentazio sistemaren funtzionamenduaren ikuspegi orokorra ematen du egileak. Dokumentazio zerbitzu nagusiak hedabideetako base-datuak mantentzen ditu. Hedabide bakoitzak, bere aldetik, bere dokumentazio saila dauka, tokiko materiala lantzeko.

Resumen: En la siguiente ponencia, se ofrece una visión general del funcionamiento del sistema de documentación utilizado en las cabeceras del grupo Vocento. El servicio central de documentación alimenta las bases de datos comunes a todas ellas, mientras que cada cabecera tiene su departamento de documentación propio que trabaja con el material local.

1. INTRODUCCIÓN

En esta presentación me voy a limitar a mostraros el contexto global de las cabeceras o las multimedias regionales que trabajan aquí en el País Vasco, que es básicamente lo que corresponde al enunciado de estas jornadas.

En cuanto a la historia, refiriéndome a la parte vasca de Vocento, cuando no existían ni Vocento, ni Grupo Correo, me hice cargo del centro de documentación de *El Correo* en 1988 (en aquella época Jesús Oleaga estaba en *El diario vasco*). En ese año arrancamos con la base de datos puramente referencial, es decir, en aquella época lo que hacíamos era catalogar, hacer una ficha de referencia que referenciaba fotos. Poníamos un número de referencia que decía en qué sobre estaba guardada esa foto y, cuando se necesitaba, se recogía manualmente y se pasaba a fotomecánica para publicarse en el periódico.

En el 92, *El diario vasco* empezó a montar un archivo digital (yo creo que fue uno de los primeros periódicos de España en hacerlo, y quizás no sólo de España), tenía las fotos alojadas físicamente en formato digital en discos ópticos. En 1994 hicimos algo parecido en *El Correo* y, a partir de ahí, en 1996 se montó el Grupo Correo con varios periódicos, se habían comprado unos cuantos: *El diario vasco*, *El ideal de Granada*… hasta los ocho que había en esa época; y se decidió optar por un sistema distinto de documentación. Se

montó un departamento de documentación central para todo el grupo, que pasé a llevarlo yo, y se añadió a los centros de documentación preexistentes en cada una de las cabeceras de los periódicos. Jesús Oleaga pasó a llevar la parte de *El Correo*. Como no había ninguna herramienta que solucionara bien nuestras necesidades lo que hicimos fue crearla nosotros. Compramos la licencia del motor de búsqueda Verity Topic, que era como se llamaba entonces, en 1996, y montamos una aplicación por encima para que permitiera trabajar con los documentos con los que trabajamos nosotros y que ahora mencionaré.

En principio el esquema fue ese y lo mantenemos ahora, después de más de diez años. Hay un departamento de documentación central que trabaja para todas las cabeceras del Grupo… iba decir en España, ha habido en algún momento cabeceras fuera de España: hay participación en *Sud ouest* y hubo una época en la que había periódicos de Argentina (se han vendido hace poco). El planteamiento general sería: documentación central trabaja, alimenta, mantiene las bases de datos comunes, mientras que cada cabecera tiene su departamento de documentación propio que trabaja con el material local. Básicamente, en términos generales, documentación central trabaja mucho en el caso de fotos con agencias; mientras que documentación local trabaja mucho en el tema de fotos con material propio, fotógrafos propios, agencias regionales y cosas por el estilo. Documentación local trabaja con sus periódicos, mientras que documentación central no tiene una publicación en papel o similar que corresponda en igualdad a todos.

A partir de ahí, en el 2000, montamos el primer metabuscador con el cual permitimos que cualquier periódico entrara a buscar en la base de datos de los demás periódicos, o en todas a la vez. Si, por ejemplo, *El diario vasco* en su base de datos no tenía algo y tampoco lo tenía la base central, podría entrar a buscarlo en Murcia, en Málaga, en Santander, en Gijón, en cualquier base de datos montada y vinculada a la multimedia correspondiente en alguno de los casos. Este metabuscador del 2000 probablemente será de lo primero que tengamos que renovar porque ya se ha quedado un poco viejillo, aunque sigue dando mucho juego.

En 2006, hicimos otra herramienta, pero bajo los mismos principios filosóficos, y desde hace unos años estamos trabajando básicamente con vídeo. El vídeo es problemático, la gestión de contenidos de vídeo es complicada por varios aspectos. Uno: tamaños monstruosos de ficheros, con lo cual hasta hace bien poco los almacenamientos eran muy complejos. Ahora por suerte ha bajado muchísimo el disco duro e incluso ya están empezando a entrar cosas como el *Blu ray,* por fin, los almacenamientos holográficos ya están a la vuelta de la esquina, ya hay desde hace nada un producto comercial. Y, por otro lado, las transmisiones de ficheros de las dimensiones que tienen los ficheros de vídeo (si hay alguien de ETB por aquí supongo que también lo estará sufriendo). Imaginaos, si una foto comprimida en un formato JPEG te puede ocupar 500 kbs o 1 mega, en un ficherito de vídeo «de tres al cuarto» y de poca duración puedes estar hablando de muchos, muchos gigas. Cuando hablas de almace-

namiento de vídeo ya te vas a los teras, mientras que en el almacenamiento de artículos o fotos hablas de megas o gigas como mucho.

Digamos que los últimos desafíos que tenemos ahora son avanzar con la parte de vídeo, que tenemos ya instalada en tres televisiones, y seguir básicamente con lo que serían las búsquedas federadas agrupadas en muchas bases de datos.

En la actualidad lo que compone el Grupo Vocento sería, trabajando sobre el mismo sistema de búsquedas, 14 cabeceras que corresponden a bastantes periódicos, más que 14 periódicos, revistas, agencias como Colpisa, gratuitos etc.; lo último que hemos montado ha sido la base de datos de *Qué*, que tiene unas 18 delegaciones. Luego veremos los números que da esto, pero tiene unas dimensiones complejas y el mantenimiento es también bastante complejo.

La filosofía sería: cada cabecera, cada usuario del sistema, accede de forma natural a la búsqueda en todo el material alimentado en las bases de datos propias, más al material que es gestionado en la base de datos central, y que se replica todas las noches a todas las cabeceras donde se trabaja. Es decir, en una cabecera como podría ser *Las provincias*, en Valencia, cuando un usuario, redactor o fotógrafo… va a buscar algo de material en su base de datos, automáticamente lo que está haciendo (sin que se dé cuenta muy bien de qué está mezclado y de cómo está separado) es entrar a la base del datos central, que se está alimentando y manteniendo aquí en Bilbao, y, al mismo tiempo, a la base de datos local, que se está alimentando directamente en *Las provincias*. La búsqueda, la descarga de material, se hace sin que prácticamente te des cuenta de dónde viene eso, es decir, la base de datos a la que accede cada usuario de Vocento está compuesta de esos dos elementos.

2. FONDOS Y TRATAMIENTO DOCUMENTAL

¿Qué documentos controlamos? Uno: fotos. Digamos que es la perla de la corona porque es donde más dinero se está ahorrando en archivo frente a la compra de material. Las fotos son bastante caras, el vídeo es aún más caro pero de momento vídeo se compra bastante poco por parte de las televisiones locales, que tienen un nivel de contenido relativamente limitado y se trabaja sobre todo en cadena. Trabajamos también con gráficos vectoriales tratables con *Free Hand* o *Illustrator*.

Trabajamos también con artículos, evidentemente, y con páginas. Elegimos en su momento, en 1996, el formato PDF cuando todavía no era estándar y que gracias a Dios, por suerte, se ha convertido en estándar absoluto. Trabajamos también con vídeos de alta resolución y distintos formatos GIF, AVI y similares. Tenemos vaciados de revistas de cine, desde 1995, y alguna cosilla más.

Los archivos de papel de foto que todavía no están digitalizados en todos los periódicos y que en algunos casos alcanzan dimensiones vistosas, como puede ser el archivo de fotos en papel de ABC… En la última estimación se

iba por encima de los 6 millones de fotos, es decir, te puedes ir a ver la guerra de Marruecos, y tienes un cuerpo de armario completamente lleno de fotos de la guerra de Marruecos (desde el desastre de Annual…), pasando por un cuerpo entero dedicado a Franco, un cuerpo dedicado a la República y cosas por el estilo; eso todavía no está digitalizado y se va pasando poco a poco. Algunos periódicos sí que lo están haciendo tranquilamente.

Lo que sí que se está digitalizando es toda la producción de papel, es decir, todas las páginas de los periódicos están entrando ya directamente en los archivos de forma que supongo que en algún momento se empezarán a comercializar, pero no sé cuando. Hubo un proyecto en su momento que se quedó parado porque parecía que no iba a dar suficiente dinero. Aquí no estamos hablando de empresas públicas sino de empresas privadas y hay que justificar que cada euro que se gasta va a revertir en el euro y algo más… aquí no podemos contar con mucho margen.

3. EL GRUPO VOCENTO

En cuanto a números, estamos hablando de 14 multimedias regionales. Para que os hagáis una idea de lo que significa un multimedia regional en el esquema de Vocento, en el caso de *El Correo* empresas o cabeceras usuarias o redacciones usuarias del sistema de documentación serían, por supuesto, la redacción central de *El Correo* (sita en la calle Pintor Losada en Bolueta), más las diez delegaciones, de las cuales hay algunas, como la de Vitoria, que en realidad construyen un periódico nuevo porque cambian prácticamente todas las páginas. Aparte de eso, tendríamos la edición digital de *El Correo*, las televisiones locales vinculadas como *Bilbovisión*, *Álava 7*, un gratuito que es *El Nervión*, otro gratuito que sería *Qué* en la edición de Bilbao, tendríamos también el Master de la UPV con *El Correo* que también tiene sus usuarios, las comercializadoras CM NORTE, las redacciones paralelas como SRB (Servicio de Redactores Bilbaínos) que hacen suplementos comerciales y partes del periódico, tendríamos las tiendas, ahora también tenemos un producto con el que si te quieres sacar la página del día que naciste puedes ir a Ikatz y te lo proporcionan y te lo enmarcan... subproductos que se sacan a partir de documentación.

Pues esto, multiplicado por las 14 cabeceras, más luego los postes de radio, aunque estos no dan normalmente mucho trabajo. Es decir, cada una de las cabeceras tiene un grupo de satélites alrededor de forma que el número de usuarios potenciales que puede tener en este momento Vocento sumadas las cabeceras es de miles. La verdad es que nunca hemos podido calcularlo con precisión, 1.000 o 2.000 usuarios profesionales, redactores, fotógrafos, infógrafos, gente en las áreas de publicidad y diagramación… Documentación, por ejemplo, se usa para emitir justificantes de que la publicidad se ha publicado; si alguien mete un anuncio la forma de justificar que se ha publicado es enviarle un PDF que justifica que ha salido tal día, al devolver la factura por correo electrónico al cliente. Eso evidentemente se saca a partir de las herramientas de documentación.

4. LA DOCUMENTACIÓN EN CIFRAS

En cuanto a tamaños (decíamos que hay un número de usuarios entre 1.000 o 2.000 profesionales), en documentación central hay tres documentalistas, en *El Correo* hay otros tres y en total entre todos los centros documentación vinculados al grupo en este esquema, andaríamos entre 20 o 30 repartidos por todas las cabeceras, más luego evidentemente la vinculación de becarios, vacaciones, suplencias etc. Hay un grado de autonomía importante, es decir, no hay una relación jerárquica estricta entre lo que sería por ejemplo documentación de Vocento y documentación de alguna de las cabeceras. Hay una relación de supervisión a veces, pero no hay una relación jerárquica como se podía esperar en otros esquemas de tipo multinacional.

En cuanto a datos, ¿qué tenemos en las bases de datos? Sumadas las cabeceras, las bases de datos ahora mismo digitales, con todos los documentos digitalizados ya andaríamos en torno a los 7 millones de fotos digitalizadas y referenciadas con sus correspondientes vías para ser localizadas; artículos, en torno a los 11 millones y medio; gráficos, algo más de los 20.000; y páginas digitalizadas, he renunciado a calcularlo porque ya nos vamos bastante más arriba. Entre *El Correo,* que cumple 100 años el 4 de mayo de 2010, *ABC* que cumplió 100 años hace uno o dos años, y *El Norte de Castilla,* que cumple 150 este año, los tres sumados, sin considerar las otras 12 o 14 publicaciones que puede haber, nos vamos por encima de los 10 millones de archivos PDF, ya digitalizados y en vías de indexación (parte está indexado y parte no), se le pasa un OCR al PDF y eso es buscable. Eso todavía no ha salido al exterior y son herramientas que se usan de momento sólo internamente. Sumado todo lo demás, teniendo en cuenta que todos los periódicos están intentando llevar todo su papel a PDF, se va a bastantes millones, unas cuantas decenas de millones de documentos. A esto habría que sumar la parte de papel todavía no digitalizada, que sólo en el caso de *ABC* estaba comentando que se iba a los 6 millones.

5. CRITERIOS DE SELECCIÓN

Criterios de selección, en artículos entra todo, en páginas entra todo; se trata muy poco, en artículos apenas se referencian cositas como si se trata de una entrevista, de una biografía, de una cronología y cosas por el estilo, si lleva gráficos, si lleva a datos…

En el caso de fotos, el proceso de selección es más complejo, para que os hagáis una idea en el caso de documentación central de Vocento vienen a entrar diariamente unas 6.000 fotos, de esas 6.000 fotos de agencias, de flujo interno, etc. aproximadamente entrarían en la base de datos en torno al 5% de lo que llega a nuestros receptores. Es decir, se vendría a catalogar, se vendría a dedicar más o menos tiempo a meter términos que permitan sacar las fotos para un

monto diario de algo menos de 100 fotos, en el caso de central. Hay otras 200 y pico fotografías diarias que se dejan pasar con poco tratamiento documental.

Por suerte, en el caso de fotografías tenemos una norma internacional que es la del consorcio internacional IPTC que ha promovido desde hace muchos años que las agencias metan información vinculada a los fichero de fotos cuando van a formato JPEG y, gracias a esa norma, a la que bendigo junto a la del PDF, conseguimos que mucho material venga ya al menos precatalogado, bien sea en inglés o en otros idiomas, pero tenemos herramientas que permiten tirar del castellano a otras cosas y básicamente encontrar fotos con pocos problemas. Por suerte, los topónimos los tenemos bien relacionados entre bastantes idiomas y tenemos la suerte de que las transliteraciones se han ido unificando a lo largo del tiempo. Os recuerdo que, por ejemplo, antes Mao Zedong no era Mao Zedong, Mao Zedong se escribe desde que el país adoptó la transliteración pinyin, en 1977, antes era Mao tse Tung. Ese tipo de cosas pueden provocar que en las búsquedas no te salga un documento; eso lo solucionamos con herramientas internas como diccionarios de sinónimos, derivaciones y cosas por el estilo.

En el caso de local, evidentemente la selección es menos estricta, porque hay mucho más material propio sobre el que se tienen derechos y que en un futuro podría ser revendible, reutilizable, sin tener que pagar a terceros y cosas por el estilo, el rango de selección podría moverse en un entorno del 40% aproximadamente. En un caso como el de *El Correo* entrarían en torno a 600 fotos diarias aproximadamente, de esas se catalogarían en torno a 100, es decir, se dedicaría tiempo a 100 fotos, a meter términos, a hacer títulos razonables y este tipo de cosas, y se meterían en torno a 150 a la situación de temporalidad.

En estado temporal aprovechamos la información que tiene previamente, o hacemos catalogaciones masivas, o agrupaciones por bloques en dosieres temáticos. Por ejemplo, para fotos de fútbol, fotos de baloncesto o cosas similares durante la temporada en curso se coge todo, pero no se cataloga todo. Se agrupan paquetes, de forma que habrá un dossier «Athletic», por ejemplo, en el cual todas las fotos que no se van a catalogar se agrupan de forma que están disponibles durante toda la temporada. Una vez que ha pasado la temporada, han cambiado los patrocinadores (aunque en el caso del Athletic la cosa es un poco distinta, pero pensad en el Barcelona), han cambiado las camisetas, los uniformes, los fichajes, el equipo del año pasado ya no vale, entonces se supone que tiene que quedar solamente lo que merece la pena que quede, es decir, el partido que jugaron contra «tal», «la final que jugaron no sé dónde», fotos buenas de los jugadores para siluetear y cosas de ese tipo, pero no todas las fotos. De un partido Barcelona-Real Madrid por agencias para central pueden entrar perfectamente 200 o 300 fotos sin ningún problema.

Ahora con el asunto de la antorcha olímpica están entrando en torno a las 500 o 1.000 fotos diarias más o menos; cuando empiecen las Olimpiadas de Pekín, probablemente nos vayamos solamente en fotos dedicadas a las Olimpiadas en torno a las 2.000 o 3.000 diarias mínimo, todos los días. Es impo-

sible, necesitaríamos un ejército de catalogadores para tratar mínimamente eso. Lo que hemos hecho es desarrollar herramientas con las cuales podemos trabajar muy bien sobre pocos documentos y muy poco sobre muchos; de forma que están visibles para los redactores, son buscables, pero no nos suponen demasiado trabajo en documentación.

6. HERRAMIENTAS DE APOYO A LAS BÚSQUEDAS

En cuanto a búsquedas, en el apoyo de búsquedas ya os he comentado el diccionario de sinónimos que hemos montado nosotros, las reglas de derivación que también hemos montado nosotros, vaciando las que nos ofrecía el producto comercial Verity. Recuerdo que cuando compramos Verity metía como sinónimos un diccionario de Hispavox de 1991 en el que hacía equivalencias de tipo pollera/falda, a mí pollera/falda no me vale de nada, a mí me vale que San Sebastián sea Donosti, que la ONU sea UN, sea Naciones Unidas, sea United Nations… cosas de ese tipo, equivalencias entre idiomas, siglas, errores frecuentes, equivalencias entre euskera, valenciano, catalán, gallego, castellano… los lugares más importantes en inglés, los códigos que utilizan las agencias para denominar a los países, ese tipo de cosas es lo que hemos ido metiendo nosotros con el tiempo y que no nos ofrecían en su momento los productos que teníamos.

Reglas de derivación para evitar trabajar con asteriscos, aunque sí mantenemos operadores a la antigua usanza… pero estamos pensando si movernos a la norma Google. Nosotros seguimos manteniendo todavía los operadores «y», «o», «no», pero estamos pensando si debemos plegarnos a lo que es hoy la norma estándar que sería la búsqueda de Google, que es más imprecisa respecto a la nuestra pero que al final es la que acaba controlando todo el mundo. Vamos, que cuando pongas «patata frita» te busque «patata» aquí, «frita» aquí y que te lo saque y punto, no como hacemos nosotros ahora que hace la frase «patata frita»… pero eso serían cuestiones más o menos menores.

Este es básicamente el contexto en el que se movería lo que en nuestro lenguaje interno sería un «local», un centro de documentación local como sería *El Correo*. Las herramientas que ha mostrado Jesús Oleaga son similares para prácticamente en todas las cabeceras de nuestro grupo, bien sea en Granada, Murcia, Santander, Donosti, Gijón; con esto mismo que estáis viendo aquí vais a *El Sur* de Málaga y las herramientas serían exactamente iguales.