Del *mobile-first* al *data-first*: *schema.org*, búsquedas *zero-click* y la incertidumbre sobre los asistentes de voz

From mobile-first to data-first: *Schema.org*, zero-click searches and the uncertainty about voice assistants

Tomás Saorín; **Juan-Antonio Pastor-Sánchez**

Saorín, Tomás; **Pastor-Sánchez, Juan-Antonio** (2020). "Del *mobile-first* al *data-first*: *schema.org*, búsquedas *zero-click* y la incertidumbre sobre los asistentes de voz". *Anuario ThinkEPI*, v. 14, e14f04.

https://doi.org/10.3145/thinkepi.2020.e14f04

Publicado en *IweTel* el 17 de abril de 2020

Tomás Saorín

https://orcid.org/0000-0001-9448-0866

Universidad de Murcia
Departamento de Información y Documentación
Campus de Espinardo, 30071 Murcia, España
tsp@um.es

Juan-Antonio Pastor-Sánchez

https://orcid.org/0000-0002-1677-1059

Universidad de Murcia
Departamento de Información y Documentación
Campus de Espinardo, 30071 Murcia, España
pastor@um.es

Resumen: Se realiza una reflexión sobre el marcado semántico desde el punto de vista de la edición y la publicación digital. Se acuña el término *data-first*, entendido como el uso de datos estructurados en los contenidos digitales para mejorar cómo los buscadores entienden con precisión la información que contienen. Se revisa la evolución y adopción del vocabulario *Schema.org* como estándar *de facto* en la Web, dado su impacto en el posicionamiento web y en las características mejoradas de las páginas de resultados de búsqueda. Finalmente se apuntan las consideraciones sobre el efecto de los datos estructurados en las respuestas directas, el discutido fenómeno de las búsquedas *zero-click* y en la interacción a través de asistentes de voz.

Palabras clave: Datos estructurados; *Schema.org*; Búsqueda web; Optimización para buscadores; Edición digital; Marcado semántico; Web de datos; Asistentes de voz; *Zero click*.

Abstract: This work reflects about semantic markup from the point of view of digital publishing and editing. It's coined the term "data-first" with the meaning of the use of structured data in digital content to improve how search engines understand precisely the information these contents embodied. Evolution and adoption os *Schema.org* vocabulary is outlined, as a de facto web standard, due to its impact in web engine optimization and featured snippets in SERPs. Finally, contented phenomena as direct answers in search results and zero-click searches are presented as results of the spread of interaction with conversational assistants.

Keywords: Structured data; *Schema.org*; Web search; Search engine optimization; Digital edition; Semantic markup; Web of data; Voice assistants; Zero click.

1. Rompiendo el hielo: fondo y forma

La Web está recorriendo un camino donde la semántica incorporada a los contenidos forma parte del proceso de publicación digital. **García-Marco** (2013) apuntó la importancia de *Schema.org*, entonces incipiente, y la enfocó como una "catalogación revisitada". En la Web actual, los datos estructurados son una parte valiosa del ciclo de vida de los contenidos y pueden marcar la diferencia entre "ser y estar", entre aparecer o no entre los resultados y recomendaciones que nos proporcionan los buscadores. Por este motivo, resulta oportuno hablar de *Schema.org* como una "edición digital revisitada". Se trata de un modelo de edición en el que los metadatos y los contenidos se editan y generan de forma simultánea para expandir el paradigma bibliotecario de búsqueda y archivo de contenidos, hacia otro que podemos llamar editorial, orientado al consumo y difusión de datos y contenidos. Un paradigma donde los contenidos digitales funcionan también como datos.

¿Qué podemos entender por edición digital? El mercado actual denomina a esto simplemente "diseño web" o, en el mejor de los casos, "gestión de contenidos":

- por diseño suele entenderse la organización visual de páginas HTML-CSS para crear una capa visualmente atractiva, legible y usable para el usuario. Aquí priman los enfoques del "diseño de experiencia de usuario" y del "diseño responsivo" donde la interacción y la visualización de los contenidos se articulan con claridad, orientándose generalmente al uso de contenidos y servicios mediante dispositivos móviles;
- con gestión de contenidos se hace referencia a la construcción de un sitio web, estructurando y presentando un conjunto de información en crecimiento, navegable, y prestando especial atención a la herramienta que permite a un equipo crear y administrar esos contenidos ágilmente con un claro control de permisos de publicación.

El contenido, en sí mismo, puede verse como la información que se transmite, los significados vehiculados mediante textos o elementos audiovisuales. En los contenidos web fluyen tanto el lenguaje natural como el visual para una lectura humana fluida y una interacción eficiente. Esta situación puede enfocarse como el clásico dilema entre fondo y forma, entre lo que se transmite y cómo se transmite. Sin embargo, para comprender esta situación en el contexto de la Red debe considerarse que el lector que tiene que acceder al fondo (el contenido) a través de la forma (la página) tiene una naturaleza dual: personas y máquinas. El lector máquina (el buscador, el algoritmo) es el intermediario que en primera instancia otorga al lector humano el acceso a los contenidos.

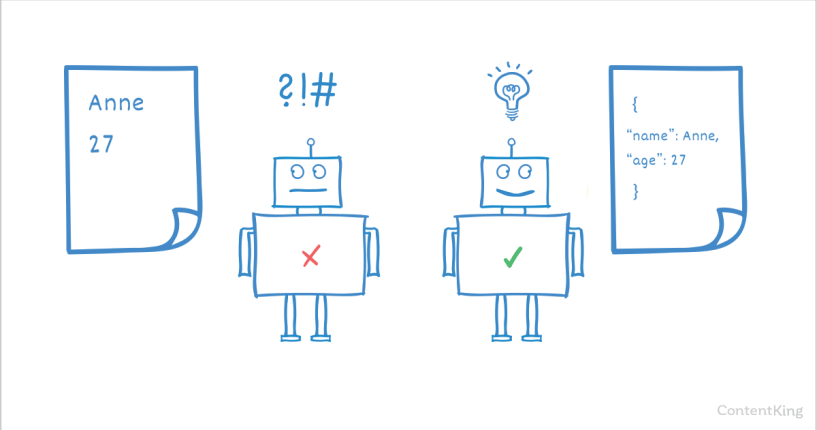

Figura 1. *https://www.contentking.es/academia/schema-org-introduccion*

Las estrategias de edición para que un lector humano entienda el contenido se asocian con la organización del discurso, la ilustración y claridad expositiva. Las estrategias de edición para el lector máquina están vinculadas con la inclusión de datos estructurados en los contenidos mediante técnicas de marcado semántico. Así que podríamos acuñar, variando el aceptado *mobile-first*, el término *data-first*: el móvil es la principal vía de acceso, pero los datos estructurados procesados por máquina actúan sobre el principal filtro de entrada.

2. Filtrado y recomendación

El lector que accede en primera instancia a los contenidos no es humano, es el *crawler* de *Google* que los indexa. Dentro de lo que se conoce como SEO técnico, que asegura la legibilidad por los *bots* de indexación, veremos que el marcado con datos estructurados cada vez requiere mayor atención. Pero *Google* no solo procesa lo que se publica, sino que al mismo tiempo nos "lee" a nosotros, los usuarios, a través de las palabras que usamos para buscar. De este modo conecta lo que preocupa o necesita el usuario con lo que hay en la Web, y posee un mapa colectivo, continuo y único en la historia tanto del conocimiento disponible como del deseado (**Galloway**, 2018).

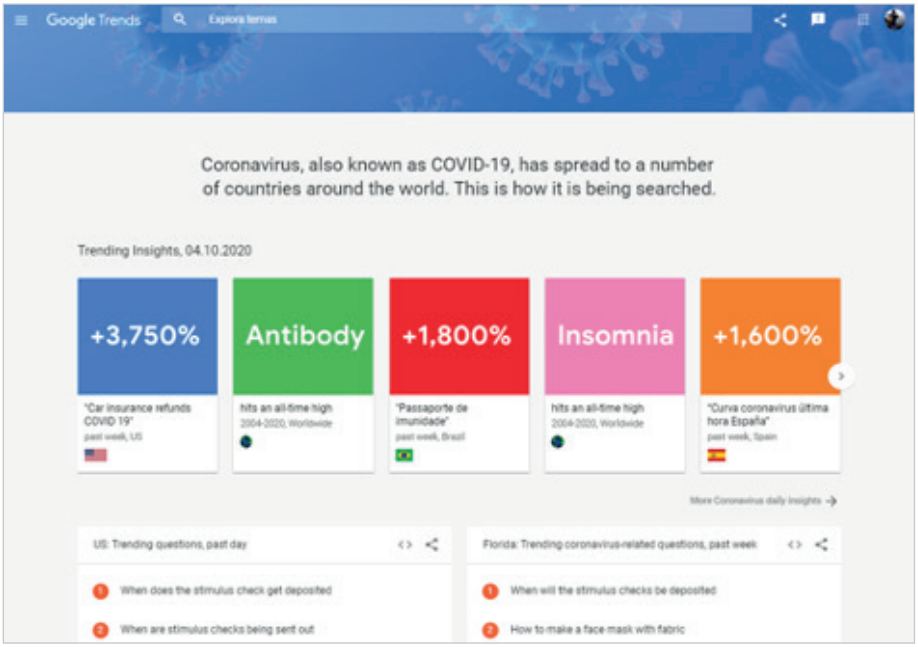

Figura 2. Tendencias de búsqueda sobre Coronavirus en *Google Trends*. *https://trends.google.com/trends/story/US_cu_4Rjdh3ABAABMHM_en*

Tomemos como ejemplo la información que en el momento de escribir esta nota muestra *Google Trends* en relación con búsquedas sobre el Covid-19. Se trata de una simple muestra del inimaginable potencial de observación del comportamiento colectivo que posee el buscador.

Google ha construido un espacio de confianza y eficiencia: siempre responde y trata de mostrar un resultado natural (lo que denomina búsqueda orgánica) donde cada entrada de la página de resultados de búsqueda responde a criterios más allá de *PageRank*: usabilidad, rapidez, fiabilidad, comodidad, accesibilidad, adaptabilidad, etc. (**Vaidhyanathan**, 2012). *Google* no solo pretende ser un simple buscador, sino también domar la Web salvaje, moldearla a su visión y sus propósitos.

Para ese objetivo tan ambicioso HTML se queda corto y CSS es irrelevante. Entonces ¿cuál es el lenguaje de la máquina? *Google* quiere trascender los aspectos estructurales y visuales. Es decir, quiere llegar al significado, a la semántica. En un principio *Google* no precisó un contenido web previamente catalogado, descrito y estructurado, sino que creó herramientas para procesar el lenguaje natural en bruto, derivando información de la misma estructura de enlaces y de la dinámica de los usuarios que usan y alimentan la Web. Pero el siguiente paso, en el que ya nos encontramos inmersos, precisa más herramienta para representar explícitamente el significado de los contenidos a través de datos estructurados, es decir, se requiere un lenguaje de metadatos para la Web real: *Schema.org*. Precisamente son los buscadores los que están interesados en que haya más claridad en la Web, más datos semánticos estructurados integrados en los contenidos, para poder ofrecer respuestas cada vez más afinadas a nuestras cada vez más precisas preguntas (**Sulé-Duesa**, 2015). Enriquecer los contenidos con datos adecuados para la Web permite ofrecer búsquedas y servicios más eficientes derivados de su procesamiento y filtrado: la idea central es *Enrich then filter* (**Alemu**; **Stevens**, 2015), que el contenido pueda ser enriquecido con diversos sistemas de metadatos a lo largo de su ciclo de vida y difusión, para incrementar la posibilidad de filtrarlo de formas sofisticadas y contextuales.

"Los buscadores están interesados en que haya más claridad en la Web, más datos semánticos estructurados integrados en los contenidos, para poder ofrecer respuestas cada vez más afinadas a nuestras cada vez más precisas preguntas"

En este escenario, el contenido digital, independientemente de su valor intrínseco, tiene que fluir de forma conveniente y competitiva, al tiempo que los sistemas automatizados intermedian entre usuarios y las publicaciones. Lo que habitualmente conocemos como "buscadores" también pueden plantearse como "mecanismos de filtrado y recomendación". Todo esto, en el negocio digital, se condensa en SEO, la optimización para buscadores (**Serrano-Cobos**, 2015), donde cada vez es más relevante entender qué es y cómo afecta *Schema.org* a la vida digital de los contenidos web.

3. *Schema.org*, el vocabulario de datos para la Web

Desde su puesta en marcha en 2011, *Schema.org* ha adquirido la inercia que se podía entrever para un proyecto nacido de los intereses comunes de los principales actores en el negocio de la búsqueda (*Google*, *Bing*, *Yandex*, *Yahoo*). Como otras iniciativas de la Red, posee cierto grado de flexibilidad, ensayo-error, despliegue parcial y, sobre todo, capacidad para demostrar su propuesta de valor con resultados tangibles. *Schema.org* se ha convertido en esencial en el sector de servicios para la optimización para buscadores. No hay consultora SEO que no incorpore esta tecnología de marcado semántico en su línea de servicios: *Moz*, *SemRush*, *Search Engine Land*, *Search Engine Watch*, etc. Para comprender su impacto en la Web, cabe diferenciar los dos extremos de la cadena:

- Publicación: incorporar datos estructurados en origen durante el proceso de publicación de los contenidos para la comprensión del contenido por los buscadores durante la indexación.
- Búsqueda: enriquecer los resultados de búsqueda, añadiendo características avanzadas en las páginas de resultados (SERP) tales como *rich snippets* o *featured snippets* gracias a la disponibilidad de datos estructurados.

En marzo de 2014 *Searchmetrics* condujo un estudio para *Google* (*Searchmetrics*, 2014), detectando que apenas el 0,3% de la Web usaba *Schema.org*. Todo ello a pesar de que casi un tercio de las SERP (*Search engine results page*) incluían características enriquecidas a partir de datos estructurados y que las páginas con marcado semántico solían estar mejor posicionadas o tener una mayor tasa de CTR (*Click through rate*).

El informe de la consultora *Forrester* en 2016, sobre las técnicas usadas por las agencias de marketing de contenidos, situaba el uso de *Schema.org* como la menos preferida, con presencia solo en un 17% de los casos (**Sentance**, 2017).

En 2017 la firma *SchemaApp*, especializada en integración de tecnologías semánticas de analítica y publicación de contenidos, realizó un estudio sobre cómo y por qué se realizaba el trabajo de marcado semántico (**Van-Berkel**, 2017) que detectó que existía un incremento de la relevancia de los datos estructurados como un factor de posicionamiento, pero que era poco habitual su integración en procedimientos robustos de gestión de contenidos. Los datos de *Web Data Commons* (*WDC*) en el proyecto *Common Crawl*, sobre el uso de propiedades y clases de *Schema.org* en noviembre de 2019 iluminan sobre la "semantización" de diversos tipos de información. Por ejemplo, para *CreativeWork* se recogen 361.338.579 *quads* o tripletas, 5.870.968 urls y 160.335 *hosts*. La propia web de *Schema.org* cuantifica el despliegue de su uso en 10 millones de sitios web. Esto supone alrededor de menos del 1% de la Web, pero que es usado en cerca del 30% de los dominios (**Bizer**; **Primpeli**; **Peeters**, 2019).

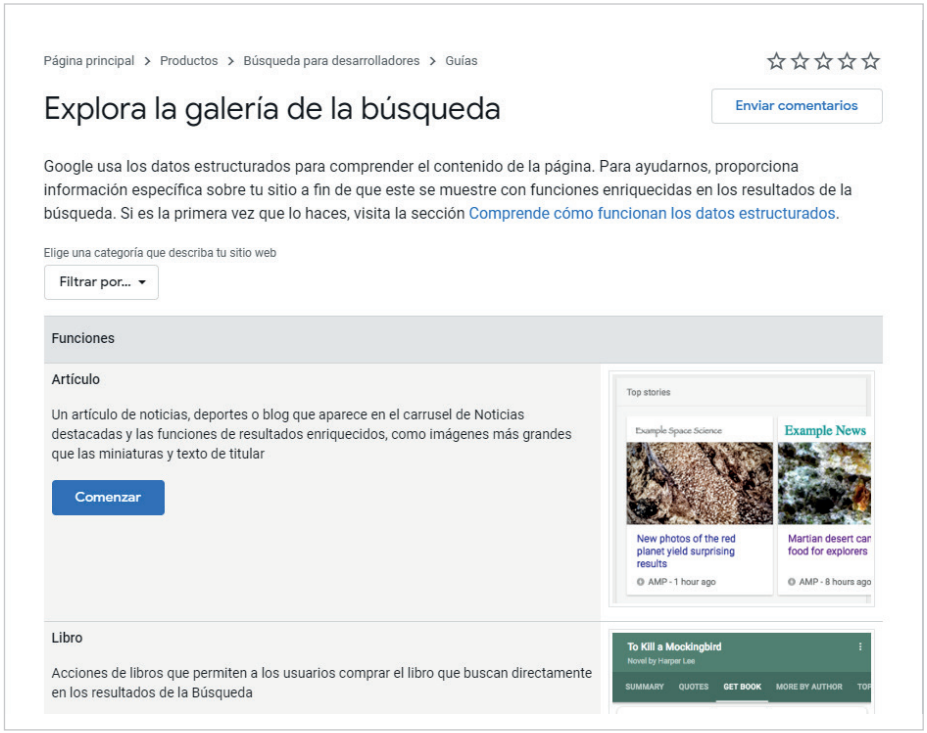

Figura 3: Guía de la galería de búsqueda de *Google*.
https://developers.google.com/search/docs/guides/search-gallery?hl=es-419

Todos estos estudios reflejan la naturaleza desestructurada de la Web, diferencias de capacidad, inversión y prioridades en la edición de contenidos, así como una larga cola de editores descuidados y sitios poco o nada optimizados. El uso de *Schema.org* aportaría ciertas ventajas competitivas por lo que cabe preguntarse sobre las razones para que no se aprovechen. ¿Barreras tecnológicas, insuficientes expectativas de recompensa? O simplemente que todavía no ha llegado el "semantic divide", el momento en el que aquellos contenidos peor construidos semánticamente vayan siendo relegados en la recomendación de los asistentes conversacionales y descubriendo que se han vuelto irrelevantes.

No obstante, la evolución de *Schema.org* sugiere que el mercado valora los conocimientos sobre vocabularios, esquemas, ontologías y organización de información. *Schema.org* comparte ciertas características con *Linked Open data*: un modelo de datos en grafo con polijerarquía de clases, entidades tipadas y propiedades (**Guha**; **Brickley**; **Macbeth**, 2015). Es un modelo *cross-domain* flexible y organizado jerárquicamente (**Fensel**; **Simsek**; **Angele**, 2020). En nueve años de recorrido ya se han lanzado siete versiones. En la actualidad cuenta con 625 tipos de contenidos (eventos, libros, reseñas, lugares, servicios, personas, etc.) más del doble de los 297 que tenía en sus comienzos. Las propiedades descriptivas han pasado de las 187 iniciales a 903. A diferencia de *Linked Open Data*, no obliga al uso de identificadores URI. Esta característica lo hace menos preciso, pero más fácil de implementar en entornos abiertos: reduce las barreras de entrada en lo que se refiere a identificación y control referencial. Se anima a los editores a incorporar el máximo de información extra sobre cada elemento descrito, para la consiguiente explotación a posteriori por parte de los agentes.

Schema.org también reduce la ambigüedad del discurso en la Red asignando tipos, nombres y propiedades a las cosas. Esto permite, por ejemplo, identificar cuando la palabra "Santiago" hace referencia a "Santiago Ramón y Cajal" (persona) o "Santiago de Cuba" (ciudad). Las plataformas sociales resuelven este problema extendiendo la gramática de escritura de los usuarios mediante punteros precisos (*hashtags* de temas y menciones de cuentas) obteniendo un proceso para "alimentar a la máquina" (**Saorín**, 2020). Pero en una Web abierta de contenidos heterogéneos y discursos complejos y extensos (un vídeo, un documento, un podcast), los buscadores tienen que analizar y comprender estos elementos semánticos, su significado y su contexto para poder ofrecer respuestas relevantes. Enriquecer los contenidos con datos estructurados incrementa la capacidad de servicio del buscador e intermediador y beneficia la legibilidad global de la Red a favor de sus usuarios. Y *Schema.org* alimenta el *Google Knowledge Graph*, que está detrás de su increíble capacidad para "entender las cosas", para construir contexto y sentido.

Lo que más atrae la atención de *Schema.org* no es su condición de ontología, sino la forma en que se visualiza en la SERP (página de resultados de búsqueda). Un recorrido por su evolución en *Google* revela que se han ido incorporando funciones complementarias en forma de *snippets* que acercan de una forma más directa al contenido. En la experiencia de búsqueda estos *featured snippets* desempeñan un papel cada vez más relevante frente al listado de páginas encontradas (**Soulo**, 2020). En contraposición al mero listado de páginas (búsqueda orgánica) los *featured snippets* muestran respuestas a partir de datos concretos, sugieren contenidos como publicaciones, productos o servicios. Estos resultados que permiten responder de forma precisa a preguntas concretas proceden del contenido marcado con datos estructurados. Es revelador que las reglas de catalogación con *Schema.org* se denominen "Galería de búsqueda". Allí puede encontrarse cómo catalogar en origen un artículo, receta, evento, producto o podcast para que los contenidos compitan entre sí por la atención de los usuarios en la SERP.

"Se necesita disponer de datos precisos, inmediatos y geolocalizados. El usuario quiere respuestas, no páginas"

Los podcasts son un ejemplo claro de un tipo de contenido con unas características peculiares de publicación y consumo sobre el que *Google* despliega su atención desde 2018 para incorporarlos mejor a los resultados de búsqueda. El resultado es que ahora puede localizarse fácilmente su serie, obtener una guía de su contenido e incluso buscar a texto completo en sus transcripciones automáticas. Google explota el marcado semántico aplicado en los podcasts para convertirse en una pseudoplataforma de podcasts que, al ser publicados con datos estructurados, se convierten en "ciudadanos de primera clase" para el buscador, que sabe cómo incluirlos en la búsqueda para optimizar su consumo, especialmente desde el móvil. *Google* no aloja, intermedia porque los entiende y organiza.
https://podcasts.google.com

La reciente crisis mundial del Covid-19 ha provocado una proliferación de información de forma sostenida y acelerada en la Web. Esta información se publica de forma distribuida y sin instrumentos de control centralizados. Además, necesita codificaciones con significados precisos para hacerlos disponibles de forma efectiva a los usuarios. El 21 de enero de 2020 se publicó la versión 6 de *Schema.org*

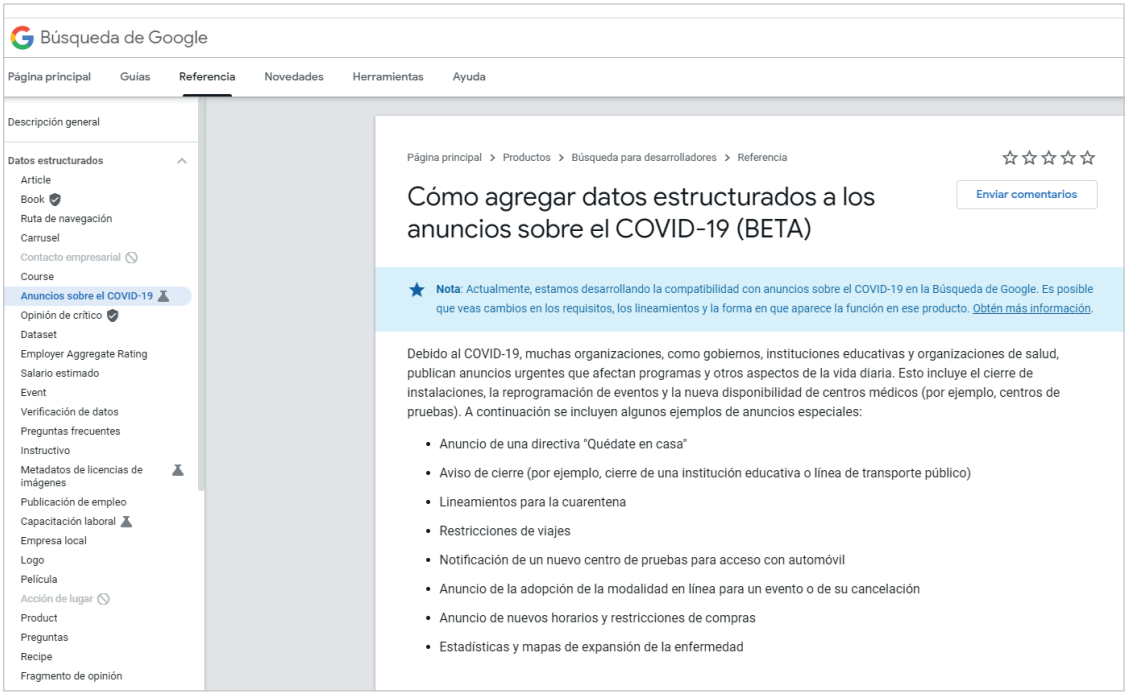

Figura 4. Guía de *Google* para agregar datos estructurados a los anuncios del Covid-19.
https://developers.google.com/search/docs/data-types/special-announcements?hl=es-419

y, pocas semanas después, ya en plena emergencia sanitaria en Europa, se publicaba la versión 7 que esencialmente reaccionaba a necesidades derivadas del confinamiento:

- incorporó el tipo *SpecialAnnouncement* para etiquetar la información de avisos especiales: cierres, ajustes de servicios… necesarios en un contexto de comunicación de crisis y conciencia del valor de los buscadores como servicio público (**Brickley**; **Guha**; **Marsh**, 2020; **Harvey**; **Sullivan**, 2020);
- definió *CovidTestingFacility* para informar de entidades que prestan el servicio del test del Covid-19;
- para los eventos que se desarrollan online se creó el tipo *VirtualLocation*, y la propiedad *eventAttendanceMode* permite señalar las formas de participación, indicando que un evento había cambiado a modo online mediante *EventMovedOnline*.

Estos cambios persiguen que los buscadores puedan determinar sin ambigüedad, y de forma inmediata, los cambios introducidos en diversos servicios de la comunidad, así como disponer de datos precisos para responder a millones de preguntas de valor crítico (lo que se conoce en jerga como YLYM "*your life, your money*"). Frente a la pregunta de un ciudadano "dónde puedo hacerme una prueba del coronavirus" no basta con obtener un listado de páginas que contengan los términos de búsqueda u otros relacionados, sino que se necesita disponer de datos precisos, inmediatos y geolocalizados. El usuario quiere respuestas, no páginas.

La adopción masiva en mercados de consumo y no en entornos especializados (cultura, ciencia, investigación) convierte a *Schema.org* en los metadatos *mainstream*, en el estándar *de facto* de la Web semántica en la Web real (**Fensel**; **Simsek**; **Angele**, 2020). Mientras que *Dublin Core* era una apuesta por un conjunto mínimo de elementos descriptivos, flexible y para cualquier ámbito, surgidos del sector tradicional de las bibliotecas, la academia y la cultura, *Schema.org* nace desde los intereses de los buscadores y el mercado de masas, y se expande a través de la fuerza del consumo de productos, servicios y contenidos orientados al consumidor digital que busca entretenimiento, información, consejo, o guía. ¿Puede un estándar de metadatos de consumo servir para el contexto cultural y científico? Su modelo de datos posee, desde luego, más especificidad que *Dublin Core* y una gran vitalidad al ser un factor de ventaja competitiva. Podría servir para sacar a los metadatos de su nicho aislado y hacerlos omnipresentes: estructurar y sistematizar toda la información digital, desde las entradas de un concierto a una pieza de museo. Posee un gran potencial transformador del acceso y organización global del conocimiento, pero también riesgos que ya se adivinan.

4. La voz y las respuestas: el reverso tenebroso del marcado semántico

Una vez que los datos (accesibles, procesables, comprensibles y precisos) están disponibles de forma generalizada, se manifiesta una situación conflictiva: el contenido original ya no es funcionalmente necesario.

Cuando, por ejemplo, *Google* es capaz de capturar la fecha, hora y lugar del concierto de un artista, puede responder directamente a nuestra búsqueda desde su propia página de resultados sin necesidad de abrir la página original que contenía dichos datos. Esto se conoce como *zero-click searches* y cada vez preocupa más a los publicadores de la Web, que ven como *Google* parasita su inversión en contenidos y servicios sin obtener a cambio tráfico hacia sus sitios web. Los cada vez más abundantes *featured snippets* en las búsquedas parece que pueden suponer hasta más de la mitad de las búsquedas servidas por el gigante de la Web (**Alvy**, 2019) y están cambiando las reglas del juego SEO.

Este fenómeno es observable en las SERP, pero en los asistentes de voz es consustancial: *Alexa*, *Siri*, *Google Nest Hub*, dan respuestas y realizan acciones, pero no dirigen a páginas. Su funcionamiento se apoya en la extracción de datos de la Web, poniéndolos al servicio del usuario en modo conversacional. Aquí también entra en juego el marcado con *Schema.org* puesto que

> "las consultas de voz dependen en gran medida del contexto implícito, y el marcado de *Schema* puede ayudar a dar ese contexto a una página de texto ambigua" (**Sentance**, 2017).

Este campo se conoce como AEO (*Answer Engine Optimization*) y conlleva que ya no vale estar en la primera página de Google, hay que estar en la primera respuesta.

La búsqueda por voz es diferente de la búsqueda por texto: implica una mayor capacidad de procesar conocimiento preciso y formas diferentes de ofrecerlo al usuario. En este marco ya surgen las alertas sobre "la respuesta por defecto": la primera respuesta hablada es la única, aquí el concepto de aparecer en la primera página carece de sentido y la información se convierte en un resultado cuya fuente se vuelve invisible (**Belsky**, 2018). Uno de los efectos disruptivos de estos nuevos interfaces son la reducción de los procesos de selección, dejando de lado la comparación y la elección. Esto implicaría además una ventaja brutal para el producto que se ofrece como primera respuesta al consumidor final. Comprender este nuevo escenario es esencial para los actores de la Web y la consultoría en optimización para buscadores. La búsqueda por voz es

> "una lucha vía UX por la interfaz de usuario y vía resultados por la inteligencia artificial" (**Muñoz**, 2017).

Google es cada vez más conversacional, volviéndose más etéreo e influyente en nuestra manera de resolver tareas y tomar decisiones. El contenido se desmaterializa en datos y precisamente el modo en el que se embeben esos datos en el contenido hace que este fluya a un modelo u otro de reutilización digital.

5. La picadora de contenido: el futuro es multimodal

Sirva como anécdota el caso de Alex Hinojo, durante un tiempo el omnipresente activista digital de *Amical Wikimedia*. Era imposible hablar de la febril actividad de *Wikipedia* en catalán sin que apareciera su nombre. Sin embargo, ahora es conocido como "el de la tostadora", porque usa la idea de poder hablarle en catalán a sus electrodomésticos para explicar la visión de las nuevas estrategias para la eficacia social de una lengua y una cultura (**Dedéu**, 2019). La tostadora es el futuro de la Web, porque ha captado que el futuro es de la interacción hombre-máquina a través de los asistentes, en una nueva era para la oralidad (**Rodríguez-de-las-Heras**, 2019) que es realidad una era de la conversación. El contenido está mediado y filtrado, y es crítico posicionarse en este nuevo orden de las cosas.

¿Qué pistas pueden hallarse en estos momentos de las características constitutivas que deberá tener el contenido digital que necesita la Web que viene? En este sentido deben considerarse determinados aspectos para que el contenido pueda fluir con un enfoque de gestión de contenidos *COPE* (*Create, Once, Publish, Everywhere*) para la publicación omnicanal. El contenido está armado sobre ciertas estructuras que determinan su plasticidad para incorporarse a flujos de republicación y transformación sofisticados, y su materialidad, su estructura interna es determinante. Esta estructura interna implica que el contenido ha de estar construido sobre datos articulados, ha de contener metadatos no solo para su descubrimiento, sino para su uso dentro de sistemas complejos de procesamiento de información. El futuro es multimodal (**Goebel**, 2020) y el contenido debe adoptar diferentes aspectos según las necesidades situacionales. Debe poder ser descompuesto en unidades significativas que puedan manifestarse a través de la voz, el texto impreso o visualizarse en pantallas, poder ser conciso y extenso al mismo tiempo, que sus significados sean independientes del idioma y presentarse contextualizado. El contenido ha de estar desligado del formato y del discurso, y para ello el contenido ha de incorporar datos

sobre sí mismo, ha de tener una estructura interna de datos estructurados, invisible para el consumidor humano, como lo son los átomos en la materia física, pero accesible para el lector máquina, que actúa como mediador, filtro y pantalla.

6. Referencias

Alemu, Getaneh; Stevens, Brett (2015). *An emergent theory of digital library metadata. Enrich then Filter*. Chandos Publishing. ISBN: 978 0081003855

Alvy (2019). "Hemos llegado al punto en que más de la mitad de las búsquedas en *Google* ya no producen clics hacia fuera de *Google*". *Microsiervos*, 14 agosto.
https://www.microsiervos.com/archivo/internet/busquedas-google-clics-fuera-zero-click-search.html

Belsky, Scott (2018). "Disruptive interfaces & the emerging battle to be the default". *Positive slope*, 9 septiembre.
https://medium.com/positiveslope/disruptive-interfaces-the-emerging-battle-to-be-the-default-23a6485a6f29

Bizer, Christian; **Primpeli, Anna**; **Peeters, Ralph** (2019). "Using the semantic web as a source of training data". *Datenbank spektrum*, v. 19, pp. 127–135.
https://doi.org/10.1007/s13222-019-00313-y

Brickley, Dan; **Guha, Ramanathan V.**; **Marsh, Tom** (2020). "Schema for Coronavirus special announcements, Covid-19 testing facilities and more". *Schema blog*, 16 marzo.
http://blog.schema.org/2020/03/schema-for-coronavirus-special.html

Dedéu, Bernat (2019). "La tostadora de Àlex Hinojo". *El nacional.cat*, 3 julio.
https://www.elnacional.cat/es/opinion/el-tostador-de-alex-hinojo_400261_102.html

Fensel, Dieter; **Simsek, Umutcan**; **Angele, Kevin** (2020). *Knowledge graphs: Methodology, tools and selected use cases.* Springer. ISBN: 978 3030374389

Galloway, Scott (2018). *Four: El ADN secreto de Amazon, Apple, Facebook y Google*. Conecta. ISBN: 978 8416883271

García-Marco, Francisco-Javier (2013). "*Schema.org*: la catalogación revisitada". *Anuario ThinkEPI*, v. 7, p. 169–172,
https://recyt.fecyt.es/index.php/ThinkEPI/article/view/30355

Goebel, Tobías (2020). "The future is multimodal: Why voice alone will never be the answer". *CMSWire*, 11 febrero.
https://www.cmswire.com/digital-experience/the-future-is-multimodal-why-voice-alone-will-never-be-the-answer

Guha, Ramanathan V.; **Brickley, Dan**; **Macbeth, Steve** (2015). "*Schema.org*: Evolution of structured data on the Web". *ACMQueue*, v. 13, n. 9.
https://queue.acm.org/detail.cfm?id=2857276

Harvey, Lizzi; **Sullivan, Danny** (2020). "Introducing a new way for sites to highlight COVID-19 announcements on Google Search". *Google webmaster central blog*, 03 abril.
https://webmasters.googleblog.com/2020/04/highlight-covid-19-announcements-search.html

Muñoz, Fernando (2017). "El ranking zero, los *featured snippets* y el futuro del SEO". *Señor Muñoz*, 21 junio.
http://www.senormunoz.es/SEO-MARBELLA/ranking-zero-featured-snippets

Rodríguez-de-las-Heras, Antonio (2019). "Un mundo para interrogar y escuchar". *Telos*, n. 111.
https://telos.fundaciontelefonica.com/telos-111-cuaderno-la-voz-antonio-rodriguez-de-las-heras-un-mundo-para-interrogar-y-escuchar-voz

Serrano-Cobos, Jorge (2015). *SEO, introducción a la disciplina del posicionamiento en buscadores*. Barcelona: Editorial UOC. ISBN: 978 84 9064 956 5.

Saorín, Tomás (2020). "Así alimentamos los algoritmos de *Google* (sin pretenderlo)". *The conversation*, 29 marzo.
https://theconversation.com/asi-alimentamos-los-algoritmos-de-google-sin-pretenderlo-132290

Searchmetrics (2014). "Over a third of *Google search* results incorporate rich snippets supported by *Schema*". *Searchmetrics*, 22 abril.
https://www.searchmetrics.com/news-and-events/schema-org-in-google-search-results

Sentance, Rebecca (2017). "The state of *Schema.org*: What are the biggest challenges surrounding Schema markup?". *SchemaApp*, 18 abril.
https://www.searchenginewatch.com/2017/04/18/the-state-of-schema-org-what-are-the-biggest-challenges-surrounding-schema-markup/

Soulo, Tim (2020). "Ahrefs' study Of 2 million featured snippets: 10 important takeaways". *Ahrefs blog*, 03 abril.
https://ahrefs.com/blog/featured-snippets-study

Sulé-Duesa, Andreu (2015). "*Schema.org*, la mejora de la visualización de los resultados en los buscadores y mucho más". *BID*, n. 34.
http://bid.ub.edu/es/34/sule.htm

Vaidhyanathan, Siva (2012). *The googlization of everything (and why we should worry)*. University of California Press. ISBN: 978 0520258822

Van-Berkel, Martha (2017). "The state of Schema Markup". *SchemaApp*, 1 abril. *https://www.schemaapp.com/research/state-schema-markup*

Tomás Saorín
tsp@um.es

Juan-Antonio Pastor-Sánchez
pastor@um.es

* * *

Nuevos retos

Natalia Arroyo-Vázquez

Enhorabuena por esta nota ThinkEPI, Tomás y Juan Antonio, que nos plantea retos para nuestros sitios web.

Al pensar en la aplicación práctica de todo esto, me surgen dos cuestiones:

1. Cómo se integra *Schema.org* en los principales gestores de contenidos (CMS).
2. Qué tipos de *Schema.org* pueden ser especialmente útiles para el sitio web de una biblioteca. Echando un vistazo rápido a la jerarquía completa, *https://schema.org/docs/full.html*

se encuentran dos tipos que son especialmente aplicables en este caso:

- *LibrarySystem*, que describe a un sistema de bibliotecas; *https://schema.org/LibrarySystem*
- *Library*, aplicable a bibliotecas (entendidas como organizaciones o como negocios locales). *https://schema.org/Library*

¿Qué otros tipos recomendáis en este caso concreto?

Natalia Arroyo-Vázquez
Universidad de Navarra. Servicio de Bibliotecas
natalia.arroyo@gmail.com

* * *

Schema.org en bibliotecas

Tomás Saorín y Juan-Antonio Pastor-Sánchez

Respondemos a las preguntas de Natalia:

1. Cómo se integra *Schema.org* en los principales gestores de contenidos (CMS)

Este es un punto relevante. Si el marcado semántico no está integrado en el propio gestor de contenidos con el que publicamos, la capacidad real de usarlo se reduce porque se hace costoso, irregular e infragestionado. Hace años que nos hemos acostumbrado a los editores visuales al crear contenido web (edición *Wysiwyg*: *what you see is what you get*) y el HTML de la página se genera solo: nosotros solo editamos negritas, enlaces, listas, etc. Pero en el caso que nos preocupa aquí tendríamos que hablar de edición *Wysiwym* (*what you see is what you mean*) en la que lo que se busca es anclar los significados a las piezas de un contenido, para que el precio se entienda como precio, las coordenadas como coordenadas, *Wilco* como un grupo de música y *Abbey Road* como el título de un disco de música. ¿Cómo hace esto un CMS? Cuando hablamos de gestión de contenidos web conviene empezar por el estándar *de facto* de la Web, *WordPress*, porque más o menos sirve para entender el punto de partida. Entendemos por contenido web el contenido no-estructurado, o en el mejor de los casos semi-estructurado: se trata de un contenido con un título, unas etiquetas y un texto, accesible desde una dirección web. En este caso, el marcado suele realizarse mediante *plugins* que permiten especificar al editar el tipo de contenido y sus propiedades, conforme a los tipos disponibles en *Schema.org* y sin elementos de control de vocabulario ni de referencia. Supone una duplicación del esfuerzo del editor, se escribe el contenido

en una parte del editor y en otra se describe. Sin embargo, otros *plugins* definen tipos de contenidos personalizados, de forma que el sitio web puede tener contenidos que son concebidos como noticias, otros como entrevistas, otros como podcasts, otros como reseñas, etc. En este caso el marcado del tipo puede venir de serie, en la propia plantilla de generación dinámica del contenido. Pero como su contenido sigue siendo no-estructurado, la situación es la misma que la descrita anteriormente. Un modo más avanzado se da en los *plugins* que definen nuevos tipos de contenidos con campos, como los de tienda online o recetas. En estos existen campos específicos para información como tiempo de cocción, precio, ingredientes, etc. Estos *plugins*, cuando están bien realizados, generan el marcado semántico de forma automática.

"Cuando los CMS manejan contenido más estructurado, pueden y deben generar el marcado con *Schema.org*, y además deben permitir el uso de vocabularios controlados y referencias claras"

En resumen, cuando los CMS manejan contenido más estructurado, pueden y deben generar el marcado con *Schema.org*, y además deben permitir el uso de vocabularios controlados y referencias claras. El CMS de un medio de comunicación integra estas funciones de gestión de contenidos para la publicación digital y la organización interna de su archivo. Esos metadatos integrados deben mapearse con el esquema de *Schema.org* y generarse en la salida web de las páginas. Si un medio al crear una noticia la etiqueta como "Deuda pública" e indica que es una "Entrevista", esa información se especificará en el código HTML de forma automática, para que sea indexada con calidad por *Google*. A cada modelo propio de gestión de contenidos le conviene poder producir descripciones en *Schema.org*: un repositorio, una revista electrónica, una plataforma de podcasts, una comunidad virtual… Por eso, los gestores de contenidos que mejor pueden optimizarse para *Schema.org* son aquellos que posean mayores prestaciones de estructuración de la información, como podría ser *Drupal* con su estructura nativa de tipos de contenido, campos y vocabularios.

2. ¿Qué tipos de *Schema.org* pueden ser especialmente útiles para el sitio web de una biblioteca?

Para responder a tu pregunta conviene diferenciar varios niveles de descripción: un sitio web completo; un contenido individual y un listado.

Las dos propiedades que mencionas pertenecen a la rama de "Organization", y sirven para describir el sitio web de una biblioteca y de una red de bibliotecas. Sirven para controlar la imagen de la marca, para si alguien te busca como organización, *Google* sepa dirigir hacia ti directamente, e incluso pueda aparecer tu logo, teléfono y dirección en la parte derecha de la página de resultados de búsqueda. Es interesante también indicar las cuentas de redes sociales oficiales vinculadas con el sitio. Toda esta información es la que irá al *Knowledge panel*, pero más que por el marcado *Schema*, *Google* prefiere recogerla de la información de *Local Business*, para tener mayor certeza y estar menos expuesto a *metadata spam*. En la identificación de la organización como fuente institucional el marcado con *Schema* es redundante (pero la redundancia es también un mecanismo de validación para *Google*).

Pero no es aquí donde se libra la batalla por el contenido, porque muy tonto tiene que ser *Google* para no identificar claramente nuestra biblioteca, red u organización. Eso es fácil, es un objetivo bien visible. Donde se libra la competición por la atención es en los contenidos individuales: si una biblioteca ha creado una página con una guía de lectura de biografías de mujeres artistas, o tiene un blog de reseñas sobre series de televisión, o si ha programado un concierto para el día del libro. Estos contenidos son los que están al final de la cadena de la búsqueda, a donde el usuario puede querer llegar después de hacerle una pregunta a *Alexa*, o los que *Google* tiene que determinar que son interesantes en un contexto determinado.

Es decir, no es llegar al sitio web de *PcComponentes*, sino al portátil en oferta con pantalla táctil o a una página con valoraciones y reseñas sobre un producto o servicio.

Disculpa la extensión, es que es un tema apasionante, el de dónde está y cómo se comparte el significado preciso de un texto: de qué entidades se habla, qué se dice de ellas… y *Schema.org* está orientado más hacia contenidos identificables, una página que habla de un libro. Es más complicado cuando una página repasa varios libros de varios autores… ahí el significado es más difícil de capturar.

Tomás Saorín
tsp@um.es

Juan-Antonio Pastor-Sánchez
pastor@um.es