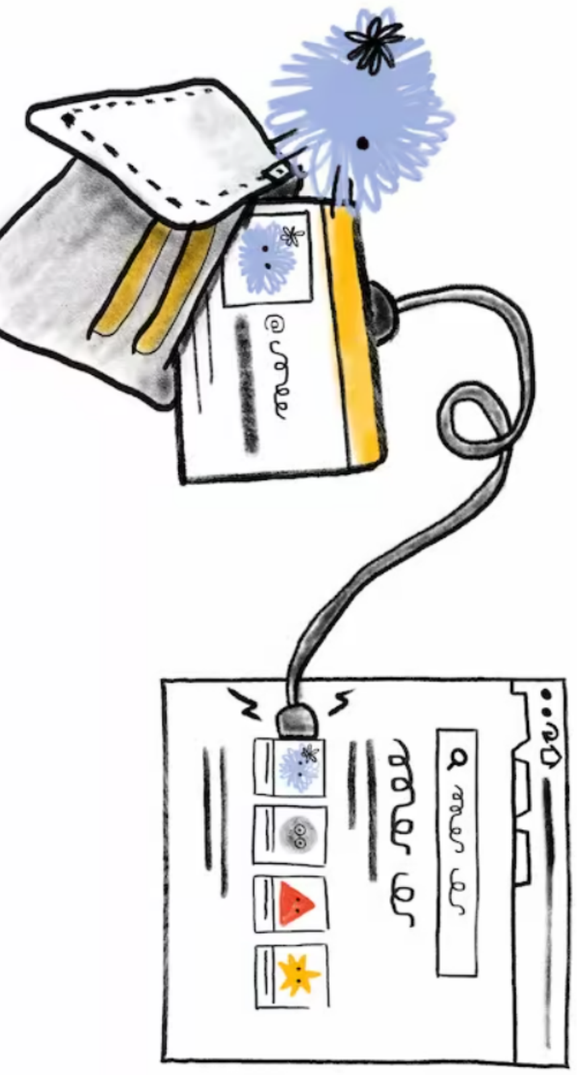
PARTE I - CÁPITULO 4

DATOS ESTRUCTURADOS

Por Alejandro Vázquez Sánchez

Puntos Clave

1. Introducción
2. Conceptos clave
3. Tipos de cobertura y datos estructurados
4. Uso por tipo
5. Cobertura por tipo de sintaxis
6. Conclusión final

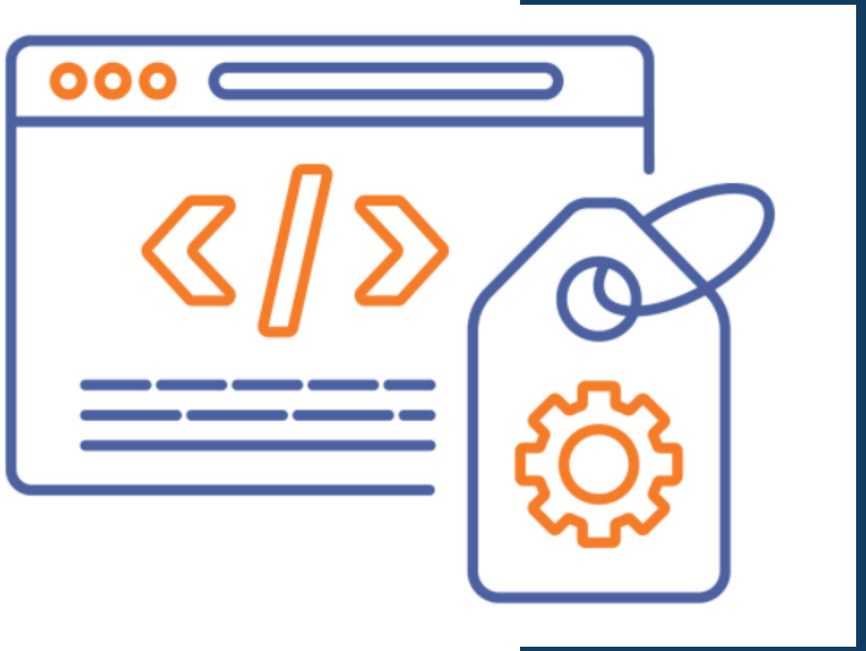


Introducción

Al leer páginas web, consumimos contenido no estructurado. Los humanos somos muy buenos identificando temas, puntos de datos, entidades, relaciones. Aunque esto es difícil de replicar para el software ya que es difícil hacerlo de una manera confiable con un alto nivel de confiabilidad.

Las limitaciones pueden restringir el tipo de cosas que podemos construir y crear de manera efectiva, mediante la estructura de la información, podemos hacer que el software entienda el contenido, ya sea agregando etiquetas y metadatos que identifican conceptos y entidades claves, asi como propiedades y relaciones.

Cuando la maquina pueda extraer datos de manera fiable a escala, habilitamos tipos de software, sistemas, servicios y negocios nuevos mas inteligentes



Conceptos claves

Los datos estructurados son un paisaje complejo y por naturaleza, abstracto y "meta". Para comprender la importancia y el impacto potencia de los datos estructurados hablaremos ahora de "La web semántica" y "Los motores de búsqueda y más"

Conceptos clave

LA WEB SEMÁNTICA

Cuando agregamos datos estructurados a páginas web públicas, definimos las entidades que esas páginas contienen, creamos una forma de datos vinculados.

Hacemos declaraciones sobre las cosas en nuestro contenido en forma de triples ya sea por ejemplo "Este video habla sobre los leones". Al describir nuestro contenido de esta manera, las maquinas pueden tratar las páginas web y los sitios web como bases de datos. A escala, crea una web semántica que es una gigantesca base de datos mundial de información. Creando a su vez, una gran cantidad de posibilidades para negocios, la tecnología y la sociedad

MOTORES DE BÚSQUEDA Y MÁS

A día de hoy, algunos de los consumidores mas amplios de datos estructurados son los motores de búsqueda y las plataformas de redes sociales.
En los motores de búsqueda, los propietarios de sitios web pueden ser elegibles para diversas formas de resultados enriquecidos mediante la implementación de varios tipos de datos estructurados en sus sitios web, teniendo un papel muy importante en la adopción general de datos estructurados en la web, además la influencia de los motores de búsqueda también ha popularizado schema.org
Las redes sociales se basan en datos estructurados para influir en la forma en que leen y muestran el contenido cuando se comparte en sus plataformas.

LA WEB SEMÁNTICA

La Web Semántica es el nombre de un proyecto a largo plazo iniciado por el W3C con el propósito declarado de hacer realidad la idea de tener datos en la Web definidos y vinculados de manera que puedan ser utilizados por máquinas no solo con fines de visualización, sino tambien con fines de visualización, automatización, integración y reutilización de datos en diversas aplicaciones

GREG ROSS

Tipos de cobertura y datos estructurados

Los datos estructurados vienen en muchos formatos, estándares y sintaxis. Hemos recopilado datos sobre los más comunes como pueden ser:

- Schema.org
- Núcleo de dublín.
- Metaetiquetas utilizadas por las redes sociales
    - Abrir gráficos
    - Gorjeo
    - Facebook
- Microformatos
- RDFa, microdatos y JSON-LD

Tipos de cobertura y datos estructurados Advertencias sobre los datos

INFLUENCIA DE LOS SISTEMAS DE GESTION DE CONTENIDO

Provienen de sitios web que utilizan un sistema de gestión de contenido (CMS) como WordPress o Drupal. Estos sistemas/módulos/temas/complementos mejoran su funcionalidad responsables de generar el marcado HTML

SUPERPOSICIÓN DE DATOS

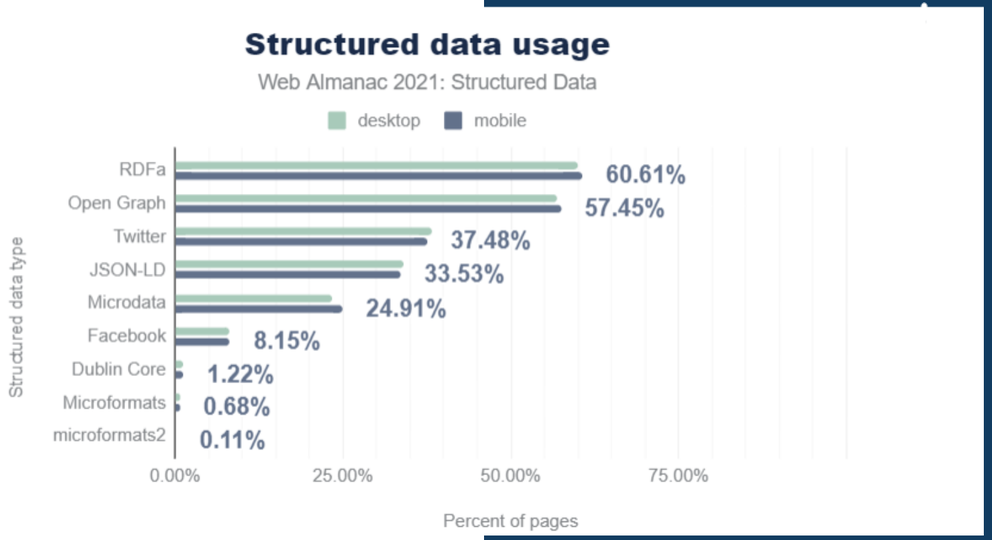
La naturaleza de algunos formatos de datos estructurados dificulta realizar este tipo de análisis de forma limpia a escala. se implementan varios formatos, las lineas entre sintaxis y los vocabularios se vuelven borrosas.
Por ejemplo, los metadatos de Facebook y Open Graph son técnicamente un subconjunto de RDFa.

LIMITACIONES DE LOS DATOS DE LA PAGINA PRINCIPAL

La naturaleza y la escala de nuestros métodos de recopilación de datos limitan nuestro análisis a las páginas de inicio únicamente. Limitando significativamente la cantidad de datos que podemos recopilar y analizar sesgando los tipos de datos que hemos recopilado.
Mayoria de paginas de inicio actuan como portales a páginas mas especificas , por el contrario, indexamos en exceso la información que normalmente se encuentra en las páginas de inicio y la información de todo el sitio que está presente en todas las paginas, como información sobre páginas web, sitios web y organizaciones

METRICAS MÓVILES

En todo nuestro conjunto de datos, la adopción y presencia de datos estructurados varía solo muy ligeramente entre nuestros conjuntos de datos de escritorio y móviles. Centrandose en el conjunto de datos móviles



Uso por tipo

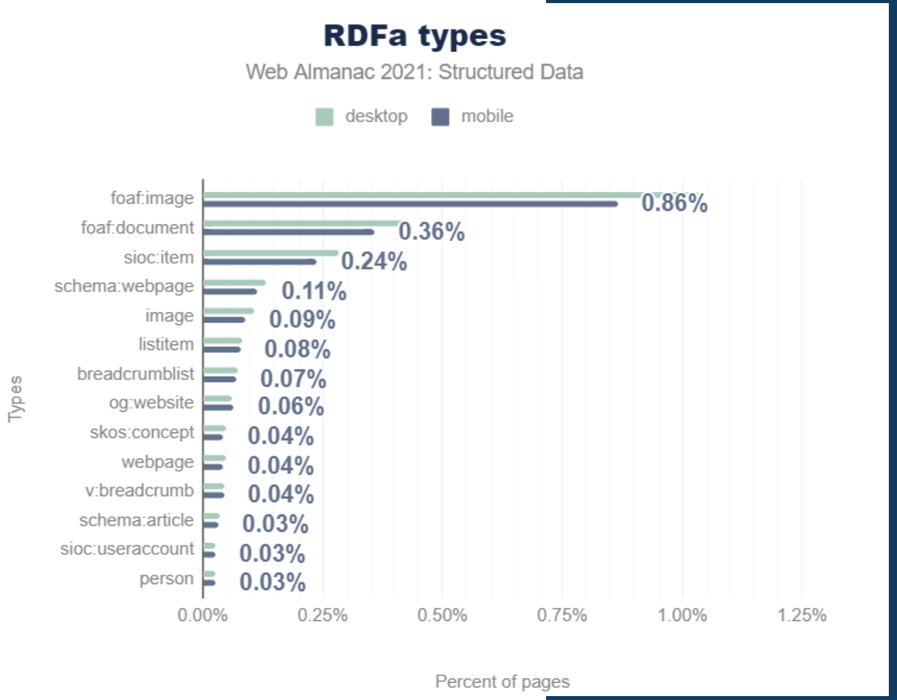
Podemos ver que hay una amplia gama de diferentes tipos de datos estructurados en muchas páginas de nuestro conjunto. También podemos ver que las etiquetas RDFa y Open Graph en particular son extremadamente frecuentes, apareciendo en el 60,61% y el 57,45% de las páginas, respectivamente.

En el otro extremo de la escala, lso formatos heredados, como los microformatos y los microformatos2, aparecen en menos del 1% de las páginas

Cobertura por tipo de sintaxis

Además de identificar cuándo está presente un cierto tipo de datos estructurados, recopilamos información sobre los tipos de datos que describe. Podemos desglosarlo en los que veremos a continuación.

RDFa



Resource Description Framework in Attributes (RDFa) es una tecnología para el marcado de datos vinculados, que fue introducida por W3C en 2015. Permite a los usuarios aumentar y traducir información visual en una página web agregando atributos adicionales al marcado.

Por ejemplo, el propietario de un sitio web puede agregar un "rel="license"" atributo a un hipervínculo para describirlo explícitamente como un vínculo a una página de información de licencia.

Cuando evaluamos los tipos de RDFa, podemos ver que la "foaf:image" sintaxis está presente en muchas más páginas que cualquier otro tipo.

Más allá de este valor, el uso del RDFa disminuye y se fragmenta considerablemente.

Dentro de este punto podemos ver los subpuntos de "En FOAF" y "Sobre otros hallazgos notables de RDFa"

En FOAF

FOAF es un diccionario de datos vinculados de términos relacionados con las personas, creados a principios de la década de 2000. Se puede utiliza para describir personas, grupos y documentos.
FOAF usa la sintaxis RDF de W3C y en su introducción original se explico de la siguiente manera:

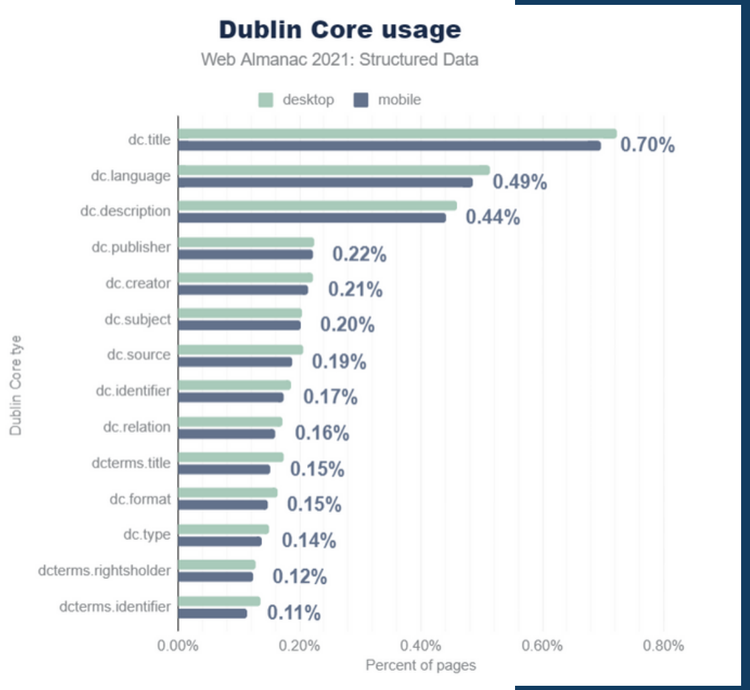
"Considere una red de páginas de inicio interrelacionadas, cada una de las cuales describe cosas de interés para un grupo de amigos. Cada nueva página de inicio que aparece en la Web le dice al mundo algo nuevo, proporcionando hechos y chismes que hacen de la Web una mina de fragmentos de información desconectados. FOAF proporciona una forma de darle sentido a todo esto."

Como anécdota podemos atribuir una prominencia de "foaf" marcado en nuestros resultados a los sitios que se ejecutan en las versiones anteriores del CMS Drupal, que históricamente ha añadido "typeof="foaf:image" y "foaf:document" de marcas a su HTML de forma predeterminada

Sobre otros hallazgos notables de RDFa

Además de las propiedades FOAF, en nuestra lista aparecen otros estándares y sintaxis.
Podemos ver varias "sioc" propiedades, como "sioc:item" y "sioc:useraccount". SIOC es un estándar diseñado para describir datos estructurados relacionados con comunidades en línea, como tableros de mensajes, foros, wikis y blogs
Tambien podemos ver una propiedad SKOS (Sistema de organización de conocimiento simple), SKOS es otro estándar, cuyo objetivo es proporcionar una forma de describir taxonomías y clasificaciones.

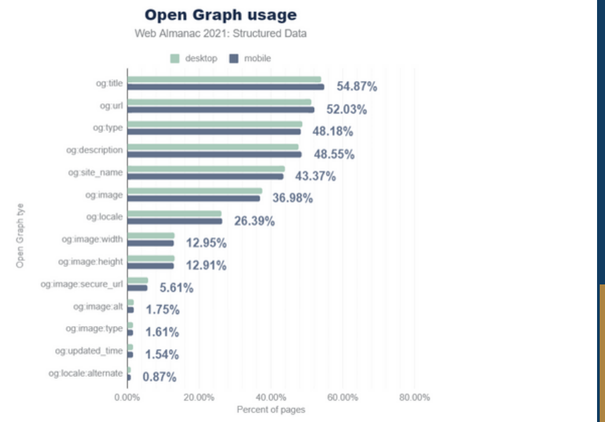
Dublin Core

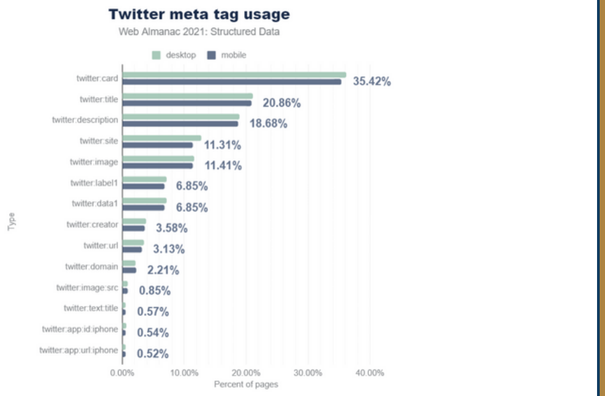


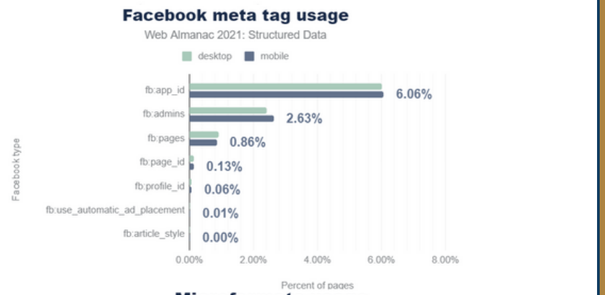
Dublin Core es un vocabulario interoperable con estándares de datos vinculados que se concibió originalmente en Dublin, Ohio, en 1995 en un taller de OCLC (Online Computer Library Center) y NCSA (National Center for Supercomputing Applications).

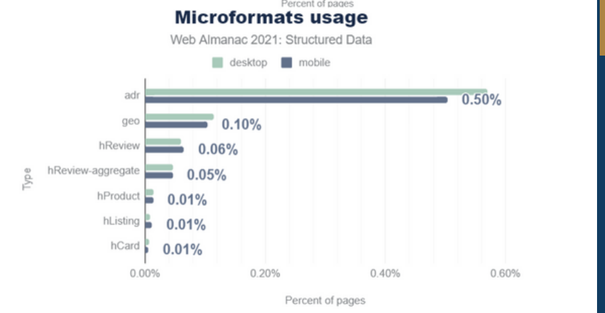
Fue diseñado para describir una amplia gama de recursos y se puede utilizar en varios escenarios comerciales. A partir de 2000 se hizo extremadamente popular entre los vocabularios basados en RDF y recibió la adopción del W3C.

Desde 2008 es administrado por Dublin Core Metadata Initiative (DCMI) y sigue siendo altamente interoperable con otros vocabularios de datos vinculados. Por lo general, se implementa como una colección de metaetiquetas en un documento HTML
Si bien a muchos les puede parecer que Schema.org es predominante en el contexto del SEO, el papel de DC sigue siendo fundamental debido a su amplia interpretación de conceptos y sus profundas raíces en el movimiento de datos abiertos vinculados.









Metadatos sociales

Las redes y plataformas sociales son algunos de los mayores editores y consumidores de datos estructurados.

- Abrir gráficos

El protocolo Open Graph es un estándar de código abierto, creado originalmente por Facebook. Es un tipo de datos estructurados especificos para el contexto de compartir contenido, basado libremente en Dublin Core, Microformats y estándares similares.

- Gorjeo

Aunque Twitter usa etiquetas Open Graph como alternativas y predeterminadas, la plataforma admite su propio tipo de datos estructurados. "Twitter:" se puede utilizar un conjunto de metaetiquetas específicas (todas con el prefijo) para definir cómo se deben presentar las paginas cuando comparten las URL de Twitter

- Facebook

Además de las etiquetas Open Graph, Facebook admite metadatos adicionales para relacionar páginas web con marcas, propiedades y personas específicas en su plataforma

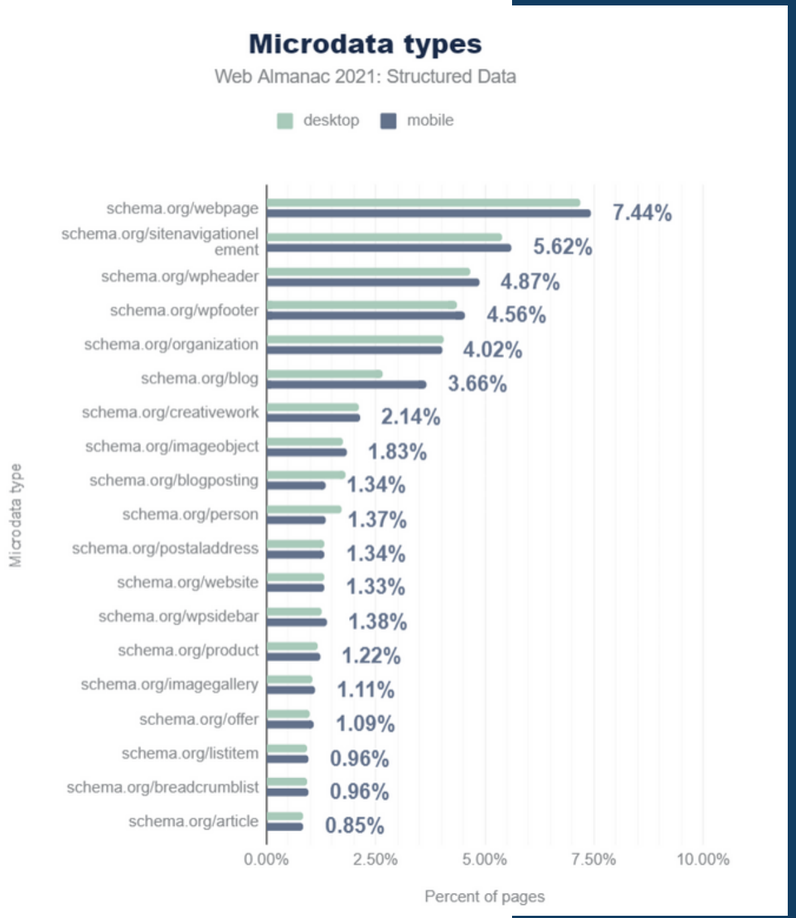
- Microformatos y microformatos2

Los microformatos son un estándar de datos abiertos para que los metadatos incrusten semántica y datos estructurados en HTML.

Estan compuestos por un conjunto de clases definidas que describen los significados detras de los elementos HTML normales, como encabezados y párrafos.

Existen dos versiones, Microformats v1 y Microformats v2. Este último, introducido en marzo de 2014 y reemplaza a v1 y aprovecha algunas lecciones importantes aprendidas de las sintaxis de microdatos y RDFa

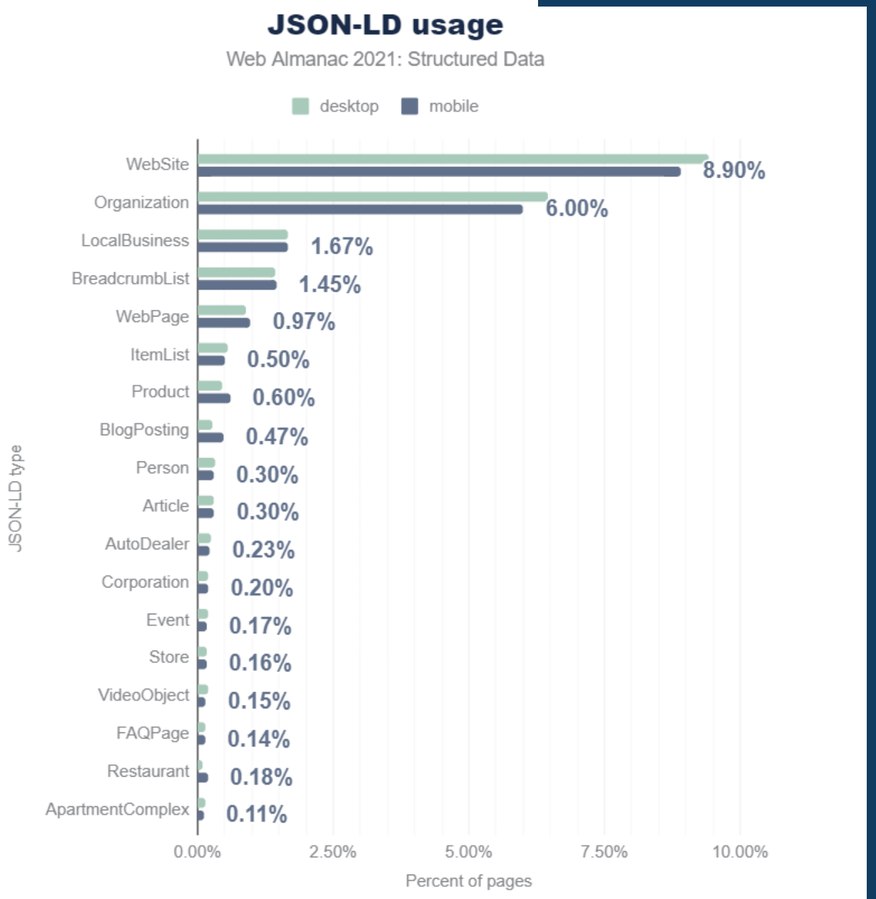
Microdatos



Los microdatos se basan en agregar atributos a los elementos HTML. A diferencia de los microformatos, pero al igual que RDFa, no está vinculado a un conjunto de significados definidos. El estándar es extensible y permite a los autores declarar que vocabularios de datos están describiendo, mas comúnmente schema.org.

Una de las limitaciones de los microdatos es que puede resultar difícil describir relaciones abstractas o complejas entre entidades, cuando esas relaciones no se reflejan explícitamente en la estructura HTML de la página

JSON-LD



A diferencia de los microdatos y microformatos, JSON-LD no se implementa agregando propiedades o clases al marcado HTML. En su lugar, el código legible por máquina se agrega a la página como uno o más blobs independientes de la notación de objetos JavaScript. Este código contiene descripciones de las entidades en la página y sus relaciones.

Debido a la implementación no está vinculada directamente a la estructura HTML de la página, puede ser mucho más fácil describir relaciones complejas o abstractas, así como representar información que no está disponible en el contenido legible por humanos de la página.

Como era de esperar, nuestros hallazgos son similares a los hallazgos de la evaluación del uso de microdatos. Ambos enfoques están fuertemente sesgados hacia el uso de schema.org como estándar predominante.

Debido a que el formato JSON-LD permite que los propietarios de sitios describan su contenido independientemente del marcado HTML, puede ser más fácil representar relaciones complejas más abstractas, que no están vicnuladas tan estrictamente al contenido de la página, como por ejemplo "BreadcrumbList" y "ItemList"

JSON-LD

ESTRUCTURAS Y RELACIONES JSON-LD

Una ventaja clave de JSON-LD es que podemos describir más fácilmente las relaciones entre entidades que en otros formatos.
Sin embargo estas relaciones a menudo pueden volverse profundas, complejas y entrelazadas.
Entonces, para los propósitos de este analisis, solo estamos viendo los tipos mas comunes de relaciones entre entidades; no evaluar árboles enteros y estructuras de relaciones.

PROFUNDIDAD DE LA RELACIÓN

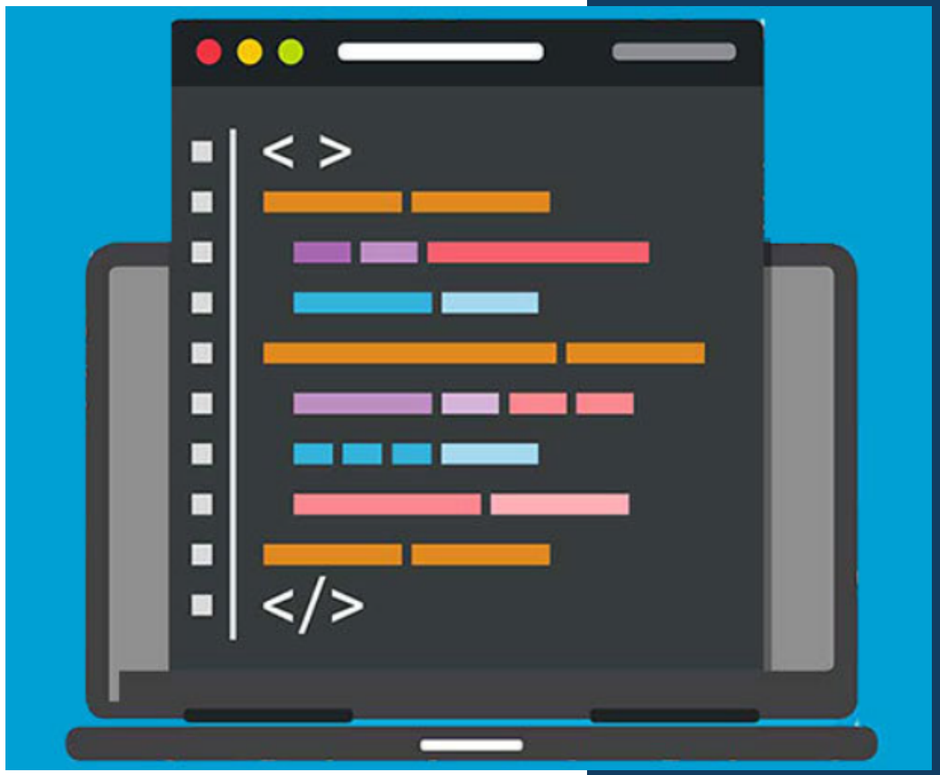
Por curiosidad, también calculamos las relaciones más profundas y complejas entre entidades, tanto en nuestros conjuntos de datos móviles como de escritorio.
Las relaciones más profundas tienden a equivaler a descripciones más ricas y completas de las entidades.

Las entidades más profundas son:
- En el escritorio, una profundidad de 18 conexiones anidadas
- En el móvil, una profundidad de 12 conexiones anidadas.

Vale la pena considerar que estos niveles de profundidad pueden indicar una generación programática de resultados, en lugar de un marcado hecho a mano, ya que estas estructuras se vuelven difíciles de describir y mantener a escala

USO DE "SAMEAS"

Uno de los casos de uso más poderosos para que los datos estructurados declaren cuándo una entidad es la "sameAs" otra entidad. Desarrollar una compresión integral de una cosa a menudo requiere consumir información que existe en múltiples ubicaciones y formatos. Tener una forma en la que cada una de esas instancias pueda hacer referencias cruzadas con las otras hace que sea mucho más fácil "conectar los puntos" y construir una comprensión más rica de esa entidad.



Conclusión

Los datos estructurados se utilizan de forma amplia y diversa en la web. Si bien parte de esto es obsoleto, también existe una fuerte adopción de estándares nuevos y emergentes.

Una web hecha de datos estructurados profundamente conectados que impulsa un mundo más integrado ha sido durante mucho tiempo un sueño de ciencia ficción. Pero quizás, no por mucho más tiempo. A medida que estos estándares continúan evolucionando y su adopción continúa creciendo, allanamos el camino hacia un futuro emocionante.

CÁPITULO 4

FIN

DATOS ESTRUCTURADOS

ALEJANDRO VÁZQUEZ SÁNCHEZ