Alphabetical Accounting Machine. www.columbia.edu/cu/computinghistory/405.html
Grace's Guide (2018) *British Tabulating Machine Co.* www.grace-guide.co.uk
Fierheller, G. A. (2014) *Do not fold, spindle, or mutilate: the 'hole' store of punched cards.* Ontario, Canada: Stewart.
Greenberg, J. (2014) *Gordon Welchman: Bletchley Park's architect of ultra intelligence*. London: Frontline.
Hinsley, F. H. and Stripp, A. (1993) *Codebreakers: the inside story of Bletchley Park.* Oxford: Oxford University Press.
Lavell, C. (2018). Personal correspondence with the author.
Lewin, R. (1978) *Ultra goes to war: the first account of World War II's greatest secret based on official documents.* New York: McGraw-Hill.
Luhn, H. P. (1957) 'A statistical approach to mechanized encoding and searching of literary information', *IBM Journal* **1**(4), 309–17.
Marden, E. C. (1965) 'HAYSTAQ: A mechanized system for searching chemical information,' National Bureau of Standards, Technical Note 264. Washington DC: US Government Printing Office.
McKay, S. (2013) *The lost world of Bletchley Park: an illustrated history of the wartime codebreaking centre.* London: Aurum Press in association with Bletchley Park.
Perry, J. W. and Casey, R. S. (1952) 'Mechanized searching.' *Encyclopedia of Chemical Technology*, Vol. 8. New York: Interscience Encyclopedia.
Pugh, E. W. (2009) *Building IBM: shaping an industry and its technology.* Cambridge, Mass.: MIT Press.
*Simpson, E. (2011a) 'Solving JN-25 at Bletchley Park: 1943–1945,' pp. 127–46 in R. Erskine and M. Smith (eds), *The Bletchley Park codebreakers.* Croydon, UK: CPI Group.
*Simpson, E. (2011b). Appendix VI: Recovering by differencing. pp. 402–5 in R. Erskine and M. Smith (eds), *The Bletchley Park codebreakers.* Croydon, UK: CPI Group.
Smith, M. (1998) *Station X: the codebreakers of Bletchley Park.* London: Macmillan.
Smith, M. (2011) *The secrets of Station X: how the Bletchley Park codebreakers helped win the war.* London: Biteback.
London: Aurum.
US Army (1943) *German military abbreviations.* Washington DC: War Department.
US General Services Administration (1965) *Source Data Automation, FPMV 11.5, Federal Stockton Number 7610-782-2670.* Copy obtained from the Central Intelligence Agency.
USHMM (US Holocaust Memorial Museum) (2017) *Hollerith Machine.* www.jewishvirtuallibrary.org
UK Channel 4 (1999) *Station X: The Ultra Secret,* episode 4 of *Station X: The Codebreakers of Bletchley Park.* http://bufvc.ac.uk/dvdfind/index.php/title/20346 (accessed 1 May 2018)
Wallis, E. and Lavell, C. (2000) 'The index of Enigma messages', *The Indexer* **22**(1), 31–3.
*Welchman, G. (1982) *The Hut Six story: breaking the Enigma codes.* Kidderminster, UK: M & M Baldwin.
*Whelan, R. (1990s) *The Use of Hollerith punched card equipment in Bletchley Park.* United Kingdom National Archives, Catalog Number HW 25/22.

Eric Nelson is a former counterintelligence operator for a United States government organization, and also a former police officer. He holds a pro bono research and academic service appointment in the Department of Public Health Sciences at the University of California, Davis. Eric has earned a PhD in Criminology and Criminal Justice (University of California, Davis), an MS in Forensic Science (National University), an MA in Sociology (University of California, Davis), an MA in Family Therapy (Azusa Pacific University), as well as two undergraduate degrees. In his spare time Eric is an avid reader of Second World War history. He is quite fond of the 25 books about Bletchley Park which reside in his library.

Unfortunately the author lives 8,381 kilometers from BP. Therefore, he is seeking collaborators who can visit BP to conduct research (research experience is not required). Please contact him at elnelson@ucdavis.edu, *with a cc to* phd@fastmail.com.

Structured data for online content: how indexers can help search engines

Alexandra Bell

Search engines are constantly crawling through a vast amount of online information. Most of this process is automated, but human assistance is still needed to tease out the meaning, context, and nuances that machines can miss. Alexandra Bell suggests that indexers could help search engines provide better results for users by applying structured data to online content.

Introduction

Indexers work diligently to interpret large volumes of information so that the valuable knowledge hiding within can be discovered. There is no larger volume of information than the internet, with the amount of information published online doubling every two years (Atlas, 2018). Search engines crawl online information relentlessly, becoming increasingly efficient at interpreting what that information is about and providing an accurate response for the vague queries that are entered into the search box. Speculation that algorithms developed by major search engines to interpret content and answer users' queries effectively will soon

be able to perform their task with little human assistance is rife. However, it is clear that no matter how well web crawlers recognize keywords, the nuanced meaning, context and intent in written information will never be completely machine-readable. In an attempt to achieve this level of understanding, major search engines encourage the application of structured data to online information. For example, Google has endorsed the use of Schema.org structured data since 2011. As the use of structured data becomes more common, the benefits are becoming clearer and its uses more wide-ranging. This article suggests that the subject analysis and precision required in effective application of structured data is highly relevant to the skills of the indexer.

What is structured data and why is it important?

In the most general sense, structured data is any data that is organized in a predictable way. Predictable patterns in data enhance the potential for machine-readability, and in turn interoperability. This article focuses on structured data in the form of semantic markup that can be applied to website content in 'a standardized format for providing information about a page and classifying the page content' (Google Developers, 2018). For example, information on a recipe webpage might have markup that identifies which components of the page represent the ingredients, cooking time, temperature and calories. A more complex example is medical information with markup that identifies the names of medications, indicating their relationship to the conditions they treat. The idea of the semantic web is not new. In 2001, Tim Berners-Lee and colleagues described a web of data that is machine-readable (Berners-Lee et al, 2001). The concept has had many critics, and is still in the process of being fully realized.

In 2011 Google, Microsoft, Yahoo and Yandex launched the Schema.org initiative, which is a set of vocabularies that can be applied using several different encodings (Schema.org, 2018). Most search engines encourage the application of Schema.org vocabularies, with the premise being that application of Schema.org markup at the right level of specificity can provide crucial information about the contextual meaning of a webpage. While the inherent meaning, context and intent of a piece of information may be immediately apparent to a human reader, semantic markup makes it comprehensible to a machine also. The launch of the Schema.org initiative in 2011 and its endorsement by major search engines has been a significant step in making Berners-Lee's vision of the semantic web a reality.

Since 2011, the use of structured data has gradually gained momentum. Initial uptake was slow, and implementation was not always accurate or useful (Paulheim, 2015; Meusel et al, 2016). Accurate figures on current use are difficult to obtain. However, the growing attention by influential SEO (search engine optimized) advice websites such as Search Engine Journal, Moz, Search Engine Watch and Yoast, in addition to the increasing visibility of rich results (formerly known as rich snippets)[1] and knowledge graph items[2] in search results illustrates the traction the application of structured data is gaining. As momentum gains, there are more practical reasons to invest the time and skill in using it well.

Schema.org structured data has been described as providing a bridge to the web of linked data (Nogales et al, 2016). Linked data, another vision of Berners-Lee (2006), is about making links between sets of data, essentially creating a web. The use of URIs[3] in structured data, as illustrated in the examples below, is a key component of enabling the meaningful linking of concepts.

While the emphasis has been on search engines and the potential for search engine optimization, the importance of structured data in a broader sense is that in making the implied meaning and context of written content machine-readable, interoperability is greatly enhanced. For example, automated personal assistant applications must source information from various resources and depend heavily on interpreting the implied context of voice-based queries (Sentance, 2017). This is an area with a great deal of potential for expansion in that any tool that aggregates content and facilitates access to information could use this (Guha et al, 2015).

Methods for applying structured data to webpages

There are several ways that Schema.org structured data can be applied to webpages. The most prominent formats or encodings are detailed below where a brief set of information about The Indexer is used as an example, showing how each method interacts with the information it is describing.

Microdata

Microdata is a Web Hypertext Application Technology Working Group (WHATWG) specification used to embed metadata within content on webpages (Html.spec.whatwg.org, 2018), and at Working Draft stage with W3C (the World Wide Web Consortium). It is the format for structured data that is interpreted by most major search engines and while it has since been overtaken by other formats, it was observed to have the fastest early uptake (Paulheim, 2015). High-level values describing the main subject of a website are easy to include in online content with little risk to disrupting the site's overall code. An often-cited risk in applying microdata within the content of a webpage is the potential to disrupt the structure of the website itself. If elements of the code are not properly completed, they can have a significant effect on the site's layout by interacting with div and span elements in a site's HTML. These issues are more prominent when microdata is applied at a very specific level, wrapping code around multiple items within a page.

JSON-LD

JSON-LD (JavaScript Object Notation for Linked Data) is a W3C Recommendation. In 2015 Google announced that it would recognize JSON-LD, recommending the format as

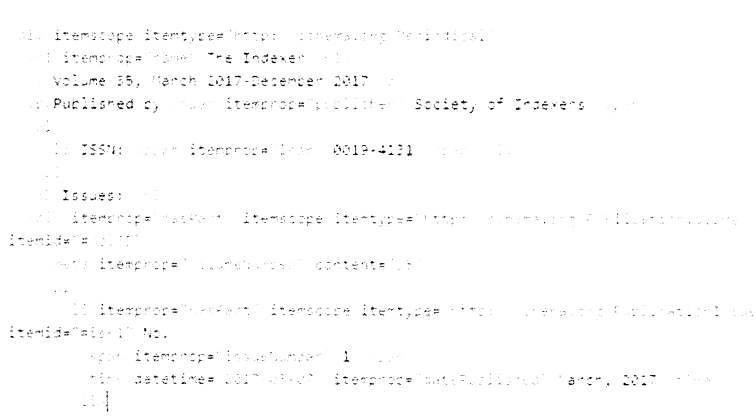

Figure 1 Microdata example from *The Indexer* articles

```
<script type="application/ld+json">
{
  "@context": "http://schema.org",
  "@type": "Periodical",
  "issn": "0019-4131",
  "hasPart": {
    "@id": "vol35",
    "@type": "PublicationVolume",
    "volumeNumber": "35",
    "hasPart": [
      {
        "@id": "issue1",
        "@type": "PublicationIssue",
        "datePublished": "2017-03-07",
        "issueNumber": "1"
      },
```

Figure 2 JSON-LD example from *The Indexer* articles

the preferred method for the inclusion of structured data on a website (Google Developers, 2018). Unlike microdata, JSON-LD is a script that can be included anywhere within a site's code without interacting with the content of the site in any way. The benefit of this is that it does not carry the risk of interrupting the structure of the website. Further, the script can describe any concept without it having to be displayed on the front end of the website.

It is expected that JSON-LD will overtake microdata as the leading format for the application of structured data due to its current endorsement by Google. However, it is still not a recommended format for other search engines, such as Bing (Bing.com, 2018).

RDFa

RDFa (Resource Description Framework in Attributes) is a W3C Recommendation that is applied in a similar way to microdata, being a set of defining attributes that are embedded within a webpage's content. As with microdata, it carries the risk of disrupting the structure of the website it is applied to by interrupting the code it is embedded within. The Open Graph Protocol encouraged by Facebook is based on RDFa. It consists of a limited set of metadata values that describe online content in a way that allows it to become a rich object in a social graph (Ogp.me, 2018). This means that it can be better recognized and interpreted by social media sites and displayed more dynamically in turn.

How is structured data being used?

A broad range of applications of structured data can be observed. Many websites are including basic structured data elements in their webpages. For example, an overall tag indicating that a page contains a recipe is a basic element that is easy to apply globally. More complex elements would identify the unique ingredients and techniques involved in each recipe, and are less prevalent. Structured data is becoming increasingly common on e-commerce sites where the benefit of a more dynamic representation in search results is significant (Caraecle, 2018) (see Figure 4). More complex areas, such as medical information have a highly specialist vocabulary, with far less examples of extensive use (Paulheim, 2015). The travel sector is working on extensions to Schema.org to add more expressive markup for describing accommodation options (Kärle et al, 2017).

Studies have observed meaningful applications of structured data in e-government and e-education websites (Navarrete and Luján-Mora, 2017). Further, some libraries are doing significant work to convert legacy metadata and catalogue records to structured data, enhancing interoperability with multiple information management systems (Jett et al, 2017; Godby and Smith-Yoshimura, 2016).

Attempts to automate the process of applying structured data are of particular interest. The authors of 'A semi-automated framework for semantically annotating web content' acknowledge that while some components can be applied programmatically, human curation is necessary to fully interpret and represent the inherent meaning of the content (Abdou et al, 2018: 94–102).

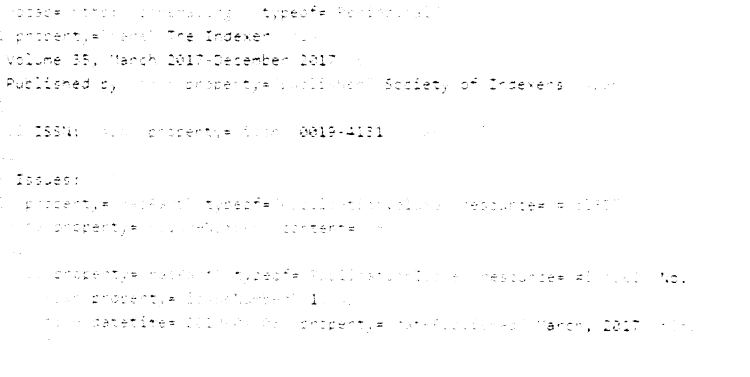

Figure 3 RDFa, *The Indexer* example

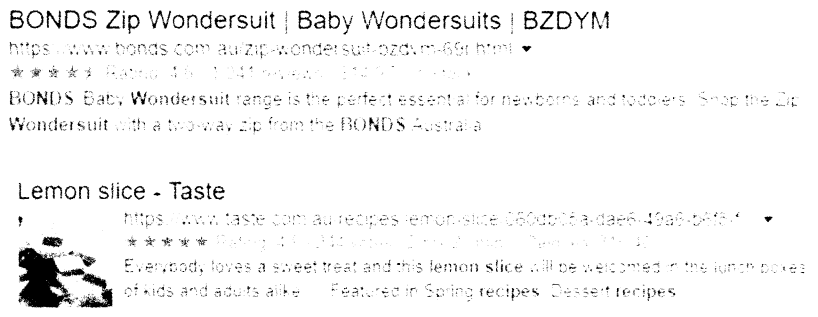

Figure 4 Examples of rich snippets (e-commerce and recipe)

How does this apply to indexers?

Indexers spend a great deal of time considering context, meaning and user intent. Professional indexers are accustomed to interpreting what content is about and then representing that in a consistent way that supports discoverability. This skill is highly relevant to the application of structured data. The effective application of structured data requires accurate subject analysis and interpretation of concepts, ensuring that relationships between concepts are not being misrepresented. The use of controlled vocabularies and adherence to formatting and data standards are also well within indexers' remit.

Going a step further into practicalities, it is helpful to consider resources that can provide guidance on the application of structured data to anyone wishing to begin using it. Given that the most immediate benefit of applying structured data to online content is the potential to display more prominently in search results, it is sensible to follow guidelines provided by Google. Google makes explicit reference to Schema.org and recommends JSON-LD as an ideal format, supporting the use microdata and RDFa also. The Schema.org website is a good place to start for guidance on vocabularies to use (`http://schema.org/`). Schema.org structured data in any of the formats accepted by Google can be tested using Google's structured data testing tool (`https://search.google.com/structured-data/testing-tool/u/0/`).

An important consideration is that with search engines gathering meaning from semantic markup there is potential for any incorrect interpretation and application of Schema.org values to be incredibly misleading. Further, there is some risk involved in a more detailed application to a large set of content, since guidelines and vocabularies can change over time. It is necessary to factor in time for recurring maintenance of any application of structured data.

Conclusion

In conclusion, structured data is a significant step towards a semantic web made up of linked data where intended meaning and context can be interpreted more effectively by search engines, formulating a rich set of results for the billions of queries that are performed every day and enhancing interoperability across platforms. Of importance is the need for human input in this process. The application of structured data to written content requires a very specific set of skills – skills that are second nature to experienced indexers.

Notes

1 Rich results are a more dynamic, information-rich version of typical search results. They can include additional information such as pricing, ratings, images and breadcrumbs (a type of secondary navigation scheme) (`https://developers.google.com/search/docs/guides/search-features`).

2 Knowledge Graph was launched by Google in 2012. Among other information, structured data is used to gain an understanding of entities such as people, places and things that are all connected. A Knowledge Graph result compiles information from multiple resources to provide a more complete set of information on a topic. The goal is to provide complete answers to search queries, rather than links (`https://searchengineland.com/library/google/google-knowledge-graph`; `https://developers.google.com/search/docs/guides/search-features`).

3 A uniform resource identifier (URI) is 'a string of characters designed for unambiguous identification of resources and extensibility via the URI scheme' (`https://en.wikipedia.org/wiki/Uniform_Resource_Identifier`).

References

Abdou, M., AbdelGaber, S. and Farhan, M. (2018) 'A semi-automated framework for semantically annotating web content.' *Future Generation Computer Systems* **81**, 94–102.

Atlas (2018) 'The dramatic rise of data creation and replication.' `www.theatlas.com/charts/E1WxoxOc` (accessed 1 June 2018).

Berners-Lee, T. (2006) 'Linked data – design issues.' `www.w3.org/DesignIssues/LinkedData.html` (accessed 1 June 2018).

Berners-Lee, T., Hendler, J. and Lassila, O. (2001) 'The semantic web.' *Scientific American* **284**(5), 34–43.

Bing.com. (2018) 'Marking up your site with structured data – Bing Webmaster Tools.' `www.bing.com/webmaster/help/marking-up-your-site-with-structured-data-3a93e731` (accessed 1 June 2018).

Caraecle, E. (2018) 'Structured data: what is it and the benefits for ecommerce websites – ecomtuning.com.' `www.ecomtuning.com/2017/07/26/understanding-structured-data-benefits-ecommerce-websites/` (accessed 1 June 2018).

Godby, C. and Smith-Yoshimura, K. (2016) 'From records to things: managing the transition from legacy library metadata to linked data.' *Bulletin of the Association for Information Science and Technology* **43**(2), 18–23.

Google Developers (2018) Introduction to Structured Data | Search | Google Developers. `https://developers.google.com/search/docs/guides/intro-structured-data` (accessed 1 June 2018).

Guha, R., Brickley, D. and MacBeth, S. (2016) 'Schema.org: evolution of structured data on the web.' *Communications of the ACM* **59**(2), 44–51.

Html.spec.whatwg.org. (2018) HTML Standard. `https://html.spec.whatwg.org/multipage/microdata.html` (accessed 1 June 2018).

Jett, J., Cole, T. W., Han, M. J. K. and Szylowicz, C. (2017) 'Linked

Digital Libraries (JCDL), Toronto, Canada, 19–23 June.
Kärle, E., Simsek, U., Akbar, Z., Hepp, M. and Fensel, D. (2017) 'Extending the Schema.org vocabulary for more expressive accommodation annotations,' pp. 31–41 in R. Schegg and B. Stangl (eds), *Information and Communication Technologies in Tourism 2017: Proceedings of the International Conference in Rome, Italy, January 24–26, 2017*. Cham, Switzerland: Springer
Meusel, R., Ritze, D. and Paulheim, H. (2016) 'Towards more accurate statistical profiling of deployed schema.org microdata.' *Journal of Data and Information Quality* **8**(1), 1–31.
Navarrete, R. and Luján-Mora, S. (2017) 'Use of embedded markup for semantic annotations in e-government and e-education websites' Unpublished paper presented at 2017 Fourth International Conference on eDemocracy & eGovernment (ICEDEG), Quito, Ecuador, 19–21 April.
Nogales, A., Sicilia, M., Sánchez-Alonso, S. and Garcia-Barriocanal, E. (2016) 'Linking from Schema.org microdata to the web of linked data: an empirical assessment.' *Computer Standards and Interfaces* 45, 90–99.
Ogp.me. (2018) Open Graph protocol. http://ogp.me/ (accessed
Paulheim, H. (2015) 'What the adoption of schema.org tells about linked open data - Semantic Scholar.' www.semanticscholar.org/paper/What-the-Adoption-of-schema.org-Tells-About-Linked-Paulheim/90cb85afee8b793b3356457b4fa740d33783769a (accessed 1 June 2018).
Schema.org. (2018) Home: schema.org. http://schema.org/ (accessed 1 June 2018).
Sentance, R. (2017) 'The State of Schema.org: what are the biggest challenges surrounding Schema markup?' Search Engine Watch https://searchenginewatch.com/2017/04/18/the-state-of-schema-org-what-are-the-biggest-challenges-surrounding-schema-markup/ (accessed 6 June 2018).

Alexandra Bell is a metadata enthusiast who has been working with online content for over 7 years, developing a genuine interest in how we describe the things we are looking for. Alexandra holds a Master of Information Studies (Information Architecture) and Master of Applied Linguistics. Email: alexandra.jane.bell@gmail.com

Indexing databases for our users, not ourselves

Valerie Nesset

Beauty lies in the eye of the beholder – and the usefulness of an index is decided by the user. Valerie Nesset argues that indexers of online databases will produce a more acceptable product if they focus on the user rather than on traditional indexing methods.

Introduction

Since the latter part of the 20th century the art and science of database indexing has been evolving constantly, but never so rapidly and dramatically as after the advent of the Internet or World Wide Web. For material previously only available in printed format, within a few decades through advances in information technology – specifically digitization – full-text electronic documents came to the fore, resulting in a proliferation of information to be organized. To best address this information explosion and facilitate retrieval of documents, the development of information retrieval systems and automated indexing became a necessity, not a choice, resulting in the development of indexing tools such as online thesauri to facilitate retrieval. But what if the end-user does not make use of these tools? Studies have shown that while thesaurus use can greatly improve retrieval, most if not all of the participants did not know either that one existed or how to use it (Sunny and Angadi, 2018). Furthermore, the terms used in the various thesauri, chosen by experts, may not be understood by the inexpert end-user who may often use other terms to express their search query (Hert et al, 2012; Spiteri, 2007; White, 2013).

So what is a good indexer to do? Index for your user, not for yourself. While it perhaps sounds simple and straightforward, indexing for our users is becoming increasingly difficult to do. Professional indexers who are skilled in various indexing and retrieval techniques may not be able to put themselves back into the mindset of the novice searcher. A simple example – the Library of Congress Subject Heading (LCSH) preferred term for trains is 'railways', yet how many young searchers would even know this term? They would most likely enter the word, 'train' and hope for the best. Yes, they would retrieve documents, but how many more would they miss? As indexers, we have to think about what makes sense to our users, meaning that sometimes we may have to bend the rules. After all, one of the greatest benefits of human indexing is our intellectual input. Unlike computers, which as of yet can only perform exact matches to terms entered in a search query, indexers can use techniques such as redundancy – including as many synonyms as possible for a particular term – to facilitate retrieval. Indexers could also develop their abstracting knowledge and skills and offer to bring them into play since with the proliferation of electronic resources, a well-done abstract can be crucial to improving information retrieval.

The importance of information retrieval brings me to another issue concerning indexing that is often not considered – that of social justice. Think about it: if we are not indexing materials in a way that the potential users of those