What is the Structured Web?

by Mike Bergman - Wednesday, July 18, 2007

http://www.mkbergman.com/390/what-is-the-structured-web/

The ***structured Web*** is object-level data within Internet documents and databases that can be extracted, converted from available forms, represented in standard ways, shared, re-purposed, combined, viewed, analyzed and qualified without respect to originating form or provenance.

Over the past few months I have increasingly been writing about and referring to the *structured Web*. I have done so purposefully, but, so far, with little background or explication. With the inauguration of this occasional series, I hope to bring more color and depth to this topic [1].

Literally, over the past year, I have been learning and documenting on **AI3** my attempts to understand the basis, concepts and tools of the emerging semantic Web. In that process, I have come to define my own outlines of the Web past, present and future. Within this world view, I see the *structured Web* as today's current imperative and reality.

Confusing Terminology Surrounding Obvious Change

Some Web pundits have embraced a versioning terminology of Web 2.0 and Web 3.0 to describe one such world view. I don't personally agree with this silly versioning -- indeed I poked fun in a tongue-in-cheek posting about Web 98.6 more than a year ago -- but such terminology has gotten some traction and serves a purpose. I actually give my own definitions for such "versions" below if for no other reason than to close the gap with alternative world views.

We need not go back to the alternative early protocols of Usenet (and news groups), Gopher and FTP and their search engines of Veronica, WAIS, Jughead or Archie in 1991 [2] when Tim Berners-Lee first publicly announced the World Wide Web and its combination of hypertext with the Internet. More likely, the release of the Mosaic browser and CERN's decision to make access to the Web free in 1993 marked the true take-off point for the Web and the continued demise of the competing protocols.

Images and links in Web pages ("documents") plus the HTML mark-up language to enable the styling and graphical design of those pages were very much in keeping with general trends, paralleling the earlier transition of personal computers to graphical interfaces and away from terminals. Mosaic became the foundation for the Netscape browser, best links compilations became a big hit through sites like Yahoo!, and the Lycos search engine, one of the first profitable Web ventures, indexed a mere 54,000 pages when it was publicly released in 1994 [3].

This initial start to the Web -- today now referred to by some as 'Web 1.0' -- can be squarely timed to 1993-1994. By 1995, the Web was appearing on the covers of major news magazines and by 1996 the phenomenon was at full throttle. But, since these early beginnings, the Web has gone through many different "versions" and transitions, most not fitting with version numbers, as some of these examples show:

- Academic *v*. Commercial Web -- magazines like Wired, Red Herring, Business 2.0 and the mainstream press showered us with names such as e-commerce, dot-com and the gold rush for companies to establish a Web presence, B2B, etc. in the latter part of the 1990s. In fact, for some early architects of the Web, this was a period of some trauma and handwringing, since the "pure" open and academic roots of the Internet and the Web were being taken over by mainstream use, commercialization and the monied dominance of venture capital. This first major change in the Web, its first major new 'version' if you will, came back down to earth as a result of the "dot-com bust" of the bubble in 2001 [4]
- Static *v.* Dynamic Web -- all initial Web content was based on documents created by hand and posted as individual and hyperlinked Web pages. The relatively few documents of the early Web meant that hand-compiled "best of" listings such as Yahoo! worked pretty well; 'metasearchers' also emerged to overcome the limited indexing coverage of early search engines. These trends, however, were also masking another version sea-change for the Web. With growth and more content, many larger sites were moving to dynamic page generation with retrieval via search forms. This dynamic portion of the Web, called at times either the 'deep Web' or 'invisible Web,' acted like standard search engines and therefore was generally overlooked until I first popularized this change in 2000 [5]. I would argue that the shift to dynamic content, with certainly hundreds of thousands of such database-backed sites now in existence -- and content many times larger than what is indexed by standard search engines -- was also a major version shift for the Web
- Open Source and Open Data --the open source Linux and the Apache Web server have been two software foundations to the growth of the Web, and MySQL has had a leading role in supporting sites and software with database-backed designs [6]. It is beyond the scope of this piece, but I believe that the dot-com frenzy, the demise of Netscape by Internet Explorer and other tensions with commercial interests, plus the very empowering nature of the Internet itself are also leading to a version change of the Web from commercial software products to open source ones. Further, proprietary publishers and data sources have only had limited success on the Web; we are now seeing strong trends to open data as well. Additionally, the very nature of open source software

lends itself to interoperability and modularity based on naturally selected building blocks. This "open" infrastructural basis of the Web is more subtle and hard to see, but provides some powerful drivers for how more surface-oriented trends express themselves
- Social Networking Web -- the same early software that enabled dynamic Web pages and database-backed designs naturally lent themselves to early blogs, wikis and content-management systems, many backed by MySQL, which in turn led to more community-oriented designs and services such as del.icio.us for bookmarking, Flickr for photos, later YouTube for videos, and literally thousands of others. This trend, resulting from changed practices and the use of different tools and ways to harness user-generated content, and not resulting from any changes to standards *per se*, was first called 'Web 2.0' by Tim O'Reilly in about 2003
- Ajax and Widgets -- some would include Web services, APIs and 'mashups' in the Web 2.0, often as expressed through embedded Web 'widgets' and the use of Ajax or similar dynamic scripting approaches. These considerations were not part of the original Web 2.0 term, but usage today likely embraces aspects of these in many definitions of Web 2.0. In any case, there is certainly a change within the Web to more interactive, attractive, full-featured user interfaces, with interface updates no longer requiring a full Web page refresh
- Document-centric Web *v.* Data-centric Web -- however, in any event, portions of these trends and changes are more broadly combining to represent another version change in the Web from one solely focused on documents to one that is more data-centric; this topic, the basis for the term '*structured Web*,' is more fully discussed below
- Web 3.0 -- Wikipedia states, "Web 3.0 is a term that has been coined with different meanings to describe the evolution of Web usage and interaction among several separate paths. These include transforming the Web into a database, a move towards making content accessible by multiple non-browser applications, the leveraging of artificial intelligence technologies, the Semantic web, or the Geospatial Web." Of all current terms, this one is fully the silliest, since there is no consensus on what it represents nor its endpoints
- Semantic Web -- the glossary at W3C states that the semantic Web is "the Web of data with meaning in the sense that a computer program can learn enough about what the data means to process it." Elsewhere, the vision of the semantic Web is described by the Education and Outreach working group (SWEO) of the W3C "to extend principles of the Web from documents to data. This extension will allow to fulfill more of the Web's potential, in that it will allow data to be shared effectively by wider communities, and to be processed automatically by tools as well as manually." Note the importance of computer processing and autonomy in these statements, not to mention the pivotal term of 'semantics.' This is an expansive and wide-embracing vision, some challenges of which I more fully describe below, and
- Visions of the Web -- the semantic Web vision is matched with other visions, including voice activation, autonomous agents doing our bidding in the background, wireless interlinked everything, and other versions of the Web that are sometimes portrayed in science fiction. Whenever such transitions occur, they will all surely rely on all the various "versions" of the Web that have occurred in the short past 15 years of the Web's existence.

Despite these differences in viewpoint, language does matter. Though some may view language as a contest in "branding," which can legitimately apply in other venues, I think the issue here goes well beyond "branding." Language is also necessary to aid communication.

As I explain below and elaborate upon more fully throughout this series, I believe one of the correct terms

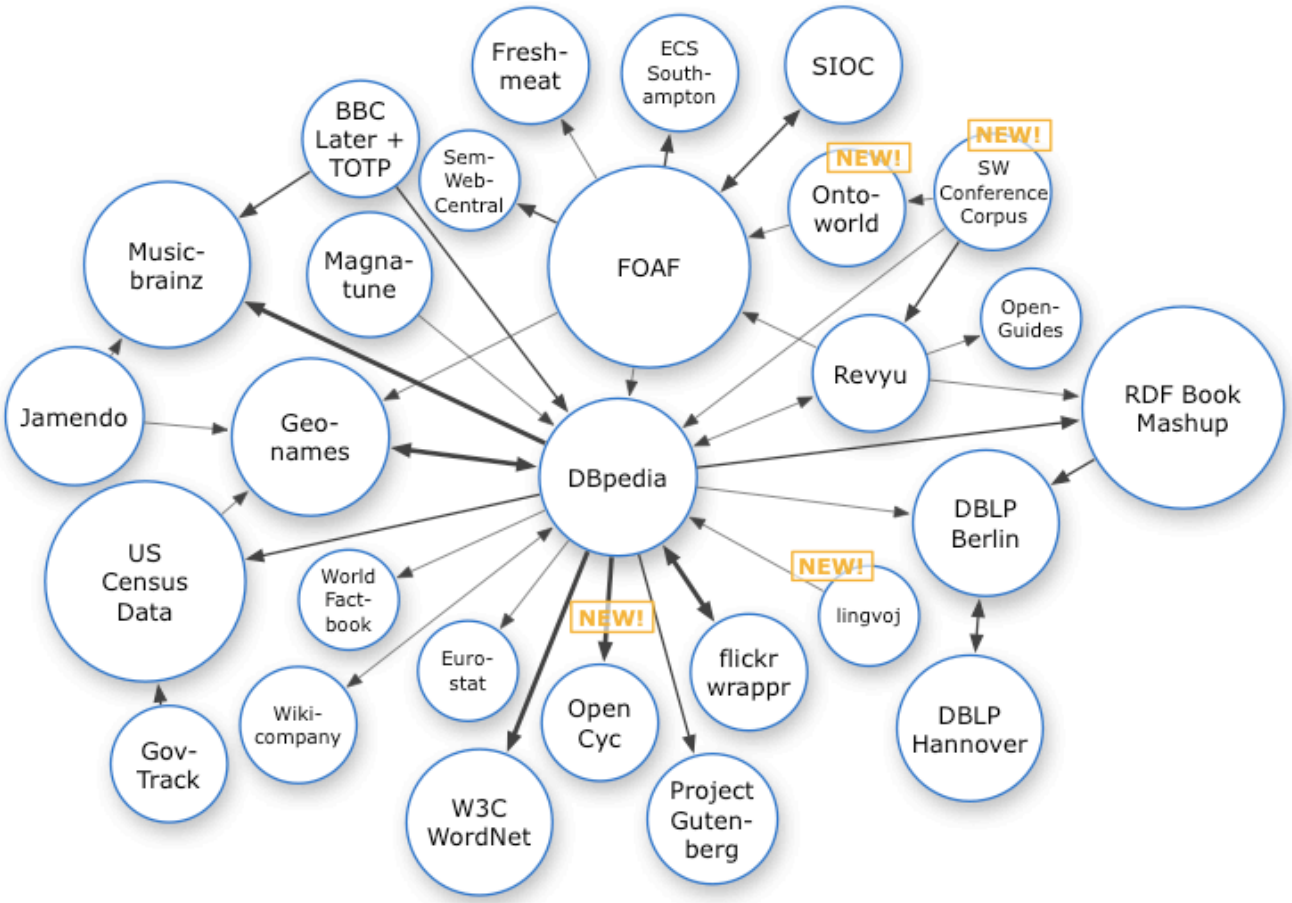
for the current evolutionary state of the Web is the '*structured Web*.'

A Clear Transition to a Data-centric Web

As noted, portions of these trends and changes are more broadly combining to represent another transitional change in the Web from one solely focused on documents to one that is more object- or data-centric. Evidence of this trend includes such factors as:

- Broad database-backed Web site designs, with content re-purposed and served up dynamically, the trend first noted as the 'deep Web'
- 'Mashups' of data from multiple sources, such as in maps, timelines, etc.
- The exposure of Web services and APIs. The programmableweb.com, for example, documents a doubling of such sources in the past nine months via its listing (as of July 2007) of about 500 APIs and more than 2,100 mashups
- Huge growth and availability of large, often public, data sources, from US government and social sources like DBpedia, an RDF data extraction from Wikipedia (and others)
- The emergence of entire data-centric sources, services and mashup platforms such as Freebase, Yahoo! Pipes, Google Base, Teqlo, QEDwiki, Ning, and OpenKapow
- The rapid -- and now almost universal -- availability of data format converters (mostly to RDF) such as the listings of the W3C's RDF Converters and MIT's 'RDFizers,' the GRDDL initiative, Triplr, and the like
- Soon, other to-be announced major data source look-up references, directories and conversion and filtering services.

One of the most popular series of presentations at this year's WWW2007 conference in Banff was from the Linked Open Data project of the SWEO interest group. The members of this LOD project -- comprised of accomplished advocates, developers and theorists -- are providing the awareness, tools and example data that are showing how this emerging version may look. In fact, the group has just announced crossing the threshold of 1 billion 'triples' with 180,000 interlinks within its online DBpedia service, via these sources:



The LOD's term for this effort is 'linked data', and a Web site has been established to promote it. Others, harking back to Tim Berners-Lee's original definition, refer to current efforts as a 'Web of data' or the 'Semantic Data Web.' Kingsley Idehen has been promoting the idea of 'data spaces' -- personal and collective -- that is also a powerful metaphor.

Frankly, I think all of these terms are correct and useful. Yet I prefer the term *structured Web* because it is both ***more*** and ***less*** than some of these other terms.

The *structured Web* is ***more*** in that it pertains to any data formalism in use on the Web and includes the notion of extracting structure from uncharacterized content, by far the largest repository of potentially useful information on the Web. Yet the *structured Web* is also ***less*** because its ambition is solely to get that data into an interoperable framework and to forgo the full objectives of the 'Semantic Web.' In that regard, my concept of the *structured Web* is perhaps closest to the idea of linked data, though with less insistence on "correct" RDF and with specific attention to structure extraction from uncharacterized content.

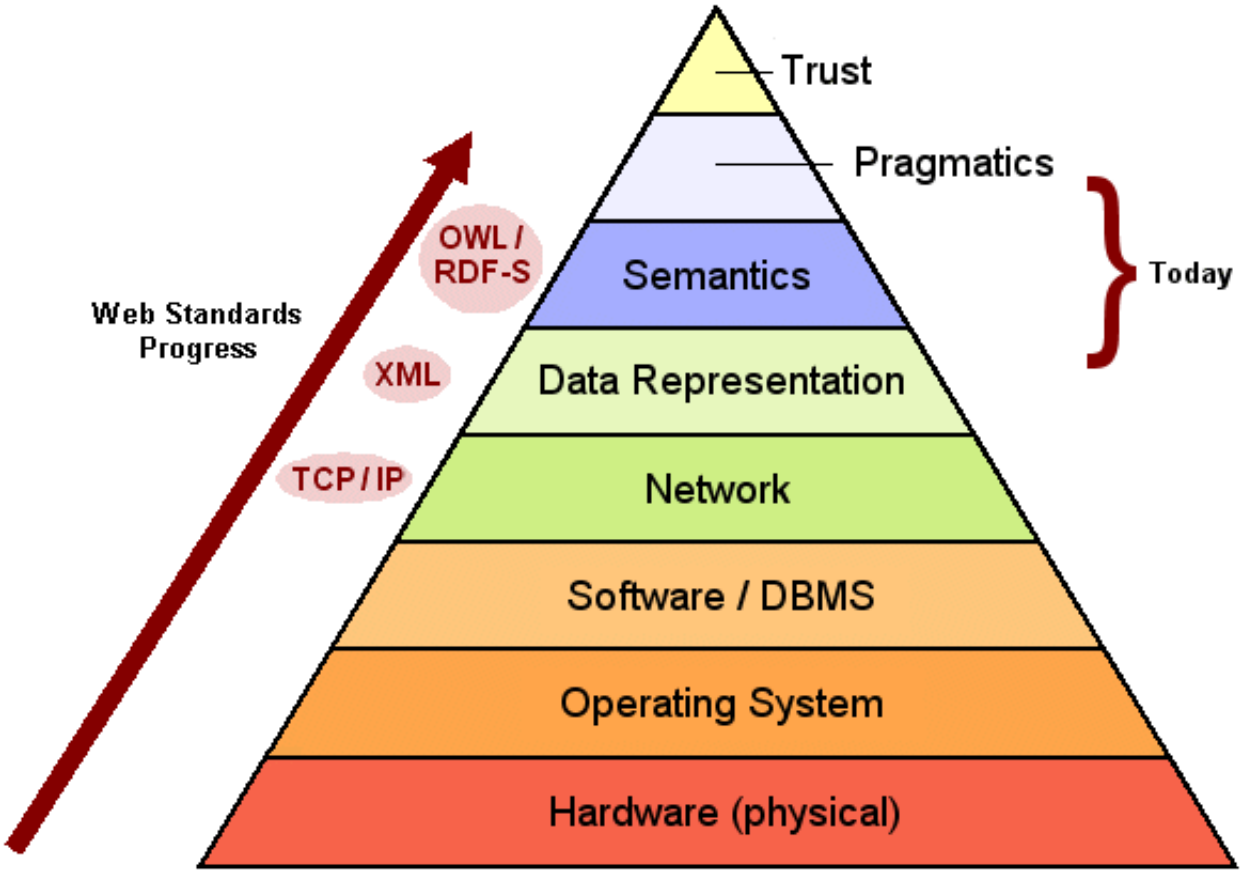
Remarkable Progress on a Still Incomplete Journey

One of today's realities is that we have accomplished much but still have a long way to go to achieve the grand vision of the 'Semantic Web' (capitalized).

More than a year ago I wrote a piece on "Climbing the Data Federation Pyramid" that noted the tremendous progress that has been made in the last twenty years in overcoming many seemingly intractable issues in data interoperability, initially of a physical and hardware nature. The Internet and Web standards have made enormous contributions to that progress.

The diagram I used in that piece is shown below [7]. Reaching the pyramid's pinnacle could be argued as having achieved the grand vision of the Semantic Web. With the adoption of the Internet and Web protocols, all layers up through data representation have largely been solved. Data representation, data models, schema for different world views, and means for reconciling and mediating those different world views are much of the focus of today's conceptual challenges.

Note, as we discuss the *structured Web* that we are largely focusing on the layer dealing with *data representation*, with some minor portions (principally in disambiguation) dealing with *semantics*. Getting data into a canonical data representation or model still leaves very crucial challenges in what does the data mean (its semantics), reasoning over that data (inference and pragmatics), and whether the data is authoritative or can be trusted. These are the daunting -- and largely remaining challenges -- of the Semantic Web.



For example, let's look solely at the layer of *semantics*, the immediate challenge after *data representation*. By *semantics*, we are referring to whether different statements from different sources indeed refer or not to the same entity or concept; in other words, have the same *meaning*. Such a determination is pivotal if we are to combine data from multiple sources.

The use of RDF, accurate name spaces and syntactically correct URIs aid this resolution, but do not completely solve it. Ultimately, semantic mediation (such as my "glad" is equivalent to your "happy") means resolving or mediating potential heterogeneities from on the order of 40 discrete categories of potential mismatches from units of measure, terminology, language, and many others. These sources may derive from structure, domain, data or language, as shown in this table [8]:

Class	Category	Sub-category
STRUCTURAL	Naming	Case Sensitivity
		Synonyms
		Acronyms
		Homonyms
	Generalization / Specialization	
	Aggregation	Intra-aggregation
		Inter-aggregation
	Internal Path Discrepancy	
	Missing Item	Content Discrepancy
		Attribute List Discrepancy
		Missing Attribute
		Missing Content
	Element Ordering	
	Constraint Mismatch	
	Type Mismatch	
DOMAIN	Schematic Discrepancy	Element-value to Element-label Mapping
		Attribute-value to Element-label Mapping
		Element-value to Attribute-label Mapping
		Attribute-value to Attribute-label Mapping
	Scale or Units	
	Precision	
	Data Representation	Primitive Data Type
		Data Format
DATA	Naming	Case Sensitivity
		Synonyms
		Acronyms
		Homonyms
	ID Mismatch or Missing ID	
	Missing Data	

	Incorrect Spelling	
LANGUAGE	Encoding	Ingest Encoding Mismatch
		Ingest Encoding Lacking
		Query Encoding Mismatch
		Query Encoding Lacking
	Languages	Script Mismatches
		Parsing / Morphological Analysis Errors (many)
		Syntactical Errors (many)
		Semantic Errors (many)

Using the same data model (say, RDF) or the same name spaces (say, Dublin Core or FOAF) helps somewhat to remove some of these sources of heterogeneity, but not all. Undoubtedly, longer term, resolving these heterogeneities will prove tractable. But they are not easily so today.

This observation does not undercut the Semantic Web vision nor negate the massive labors in support of that vision taken to date. But, hopefully, this observation may bring some perspective to the task ahead to obtain that vision.

Lowering Our Sights

If nothing else, the reality of the past 15 years shows us that the Web is a "dirty," chaotic place. If HTML coding can be screwed up, it will. If loopholes in standards and protocols exist, they will be exploited. If there is ambiguity, all interpretations become possible, with many passionately held. Innovation and unintended uses occur everywhere.

This should not be surprising, and experienced Web designers, scientists and technologists should all know this by now. There can be no disconnect between workable standards and approaches and actual use in the "wild." Nuanced arguments over the subtleties of standards and approaches are bound to fail. Robustness, simplicity and forgiveness must take precedence over elegance and theoretical completeness.

While there has been obvious growth in the sophistication of Web sites and the underlying technologies that support them, we see continued use of obsolete approaches that clearly should have been abandoned long ago (such as Web-safe colors, small displays, older browser versions, Web pages parked on some servers that have not been modified or looked at by their original authors in a decade, etc.). We also see slow uptake for clearly "better" new approaches. And we also sometimes see explosive uptake of approaches and ideas that seemingly come out of nowhere.

We also see that those approaches that enjoy the greatest success -- blogging, tagging, microformats, RSS, widgets, for example, come most recently to mind -- are those that the "citizen" user can easily and readily embrace. HTML was pretty foreign at first, but now millions of users modify their own code. Millions of users of various CMS systems and Firefox have learned how to install plug-ins and add-ins

and modify CSS themes and use administration consoles.

So, my observation and argument is not that we must always choose what is mindless and unchallenging. But my argument is that we must accept real-world diversity and seek simplicity, robustness and clarity for what is new.

After nearly a decade of standards work, the basis for beginning the transition to the semantic Web is in place. But that vision itself sometimes appears too demanding, too intimidating. The vision at times appears all too unreachable.

Of course, this perception is wrong. Measured over many years, perhaps some decades, the vision of the semantic Web ***is*** reachable. Much remains to be worked on and understood regarding this vision in terms of mediating and resolving semantic heterogeneities, capturing and expressing world views through formal ontologies, making inferences between these views, and establishing trust and authoritativeness. And those challenges do not yet address the even more-exciting prospects of intelligent and autonomous agents.

Rather, the rationale for the *structured Web* is to tone down the vision, stay with the here and now, focus on what is achievable today. And what is achievable today is very great.

Why This Series on the '*Structured Web*'?

Though version numbers for the Web are silly, with 'Web 3.0' for the semantic Web possibly being the silliest of all, such attempts do speak to the need for providing handles and language for capturing the dynamic change, diversity and complexity of the Web.

Today, right now, and all around us, a fundamental transition is taking place in the Web from a document-centric to a data-centric environment. A confluence of standards, advocacies, and previous trends are fueling this transition. Since the practical building blocks already exist, we will see this *structured Web* unfold before us at amazing speed.

The concept of the *structured Web* is thus narrower and less ambitious in scope than the 'Semantic Web.' The *structured Web* is merely a transitional step on the journey to the vision of the semantic Web, albeit one that can be fully realized today with current technologies and current understandings.

The purpose of this new series is thus to give prominence to this transition and to highlight the pragmatic, practical building blocks available to contribute to this transition. By somewhat shifting boundary definitions, the idea of the *structured Web* also aims to give more prominence to the importance of usability and structure extraction from semi-structured and unstructured content. These, too, are exciting areas with much potential.

So, as a way to provide a touchstone for continued discussion on this matter, here is one working definition of the *structured Web*:

> The ***structured Web*** is object-level data within Internet documents and databases that can be extracted, converted from available forms, represented in standard ways, shared, re-purposed, combined, viewed,

analyzed and qualified without respect to originating form or provenance.

Anticipated Topics in this Series

Some of the tentative topics that I plan to address in this series include discussion of what constitutes 'structure' in content, why structure is important, the various existing forms of structure, human *v.* machine bases for viewing and interpreting structure, the importance of finding "canonical" representation forms while also appreciating real-world diversity, the means to convert data forms and serializations, the means to extract structure from all types of content, transitioning to semantic understandings, and likely others.

Others may be added to this series over time and will be categorized under 'Structured Web' on the **AI3** blog.

This posting is the first part of a new, occasional series on the Structured Web, which also has its own new category. There are some additional prior topics in this series.

---

[1] You will note a heavy emphasis on Wikipedia definitions and histories in this piece, in keeping with the general theme of versions and transitions on the Web.

[2] News groups really did not have a good search engine until the launch of Deja News in 1995.

[3] Chris Sherman, "Happy Birthday, Lycos!," *Search Engine Watch*, August 14, 2002. See http://searchenginewatch.com/showPage.html?page=2160551.

[4] A fairly good summary of the History of the Web can be found on Wikipedia.

[5] Michael K. Bergman (Aug 2001). "The Deep Web: Surfacing Hidden Value". *The Journal of Electronic Publishing* **7** (1). An earlier version of this paper was published by BrightPlanet Corp. in July 2000.

[6] While there are variations, Linux, Apache, MySQL and the scripting languages of either Python, PHP, or Perl are often referred to as 'LAMP', one central basis for much open source software and, more broadly, interoperable open-source application packages.

[7] I would make a few changes today, notably in deprecating XML somewhat.

[8] This table builds on Pluempitiwiriyawej and Hammer's schema by adding the fourth major category of language. See Charnyote Pluempitiwiriyawej and Joachim Hammer, "A Classification Scheme for Semantic and Schematic Heterogeneities in XML Data Sources," *Technical Report TR00-004*, University of Florida, Gainesville, FL, 36 pp., September 2000. See ftp.dbcenter.cise.ufl.edu/Pub/publications/tr00-004.pdf.
