Jason Stewart | UMID: 2415-3556 | SI 541 Professor Edwards

Standardization and Gateway Technologies

The course readings throughout this semester have explored, among many other things, the countless factors that influence the development, adoption, and evolution of infrastructures. This "think piece" will reflect on some of those course readings and focus on the perceived role of standardization as it relates to gateway technologies in the adoption and evolution of electronic information infrastructures.

In our "information society" we, as information consumers, place high value on the availability of usefulness information. The degree of information usefulness in a particular context is greatly influenced by the portability of data between the various independent infrastructures that support the relevant data management and storage systems. For example, as train passenger train time data is only useful to us at the train station if it makes its way from the "master schedule" to the big flipping board above the terminal. The transactions that take place in order to display data from the master schedule to the big flipping board require the interactions of a number of infrastructures. These interactions cannot occur naturally[1] without a facilitating agent, commonly referred to as a gateway technology.

Given the high value information consumers place on the availability of contextually useful information, and the variety of data management and storage

---

[1] By "naturally", I mean to say that the components of the infrastructure itself can not facilitate the interaction

systems in existence, we can start to see the integral part gateway technologies play in getting information to consumers. In fact, with the advent of the contemporary Internet, coupled with the usefulness and popularity of “mashups” as services and sources of information, and the degree to which an increasing number of infrastructures need to be interconnected to produce such information services, it seems that the viability of future electronic information infrastructures is correlated with the ability of gateway technologies to make its payload transferable to other infrastructures.

Many infrastructures, such as the Berners Lee’s era Internet, were constructed with limited foresight and for an immediate specialized application, and thus were not explicitly designed with infrastructure integration in mind. Berners Lee, for example, did not envision the now varied use of the “his” creation during its development, but instead designed it to address the issues of data redundancy. At that point, there was no need to think about how this infrastructure would interact with infrastructures that supported airline reservations systems, or electrical grids, or telephone networks etc; all it needed to do was facilitate the transfer of valuable data from one point to another.

Gateway technologies with respect to standardization, as discussed by Tineke M. Egyedi, can manifest in three major varieties: Dedicated; Generic; and Meta-Generic. Dedicated gateways, due to lack of standardization, only link a limited and specific number of subsystems while Generic Gateways, due to standardization, have a wider scope and can link many more of subsystems. Lastly, Meta-Generic

Gateways have an even wider scope than Generic Gateways as they specify a framework or protocol for the creation of specific generic standards, without specifying those standards directly[2].

Continuing with the example of the Internet, we can explore how various gateway technologies were developed and used by interested parties and given the available options. Google arose as the need to index, categorize, and search the Internet became apparent. At the time, there was not a widely adopted standardized method of collecting relevant information about web sites being indexed by Google. One of infrastructures that facilitated the display of content on the web, Hyper Text Markup Language (HTML), could not directly communicate relevant data to Google's spiders; there was not a gateway technology in place. Nevertheless, there was a market need for a searchable index on the web, so instead hopelessly pushing for some sort of immediate standardization among web sites, Google developed a complex content analysis system that parses HTML and makes a reasonable guess as to the content of that page. Of course, the ideal interaction between the Google spider and web sites would be something like:

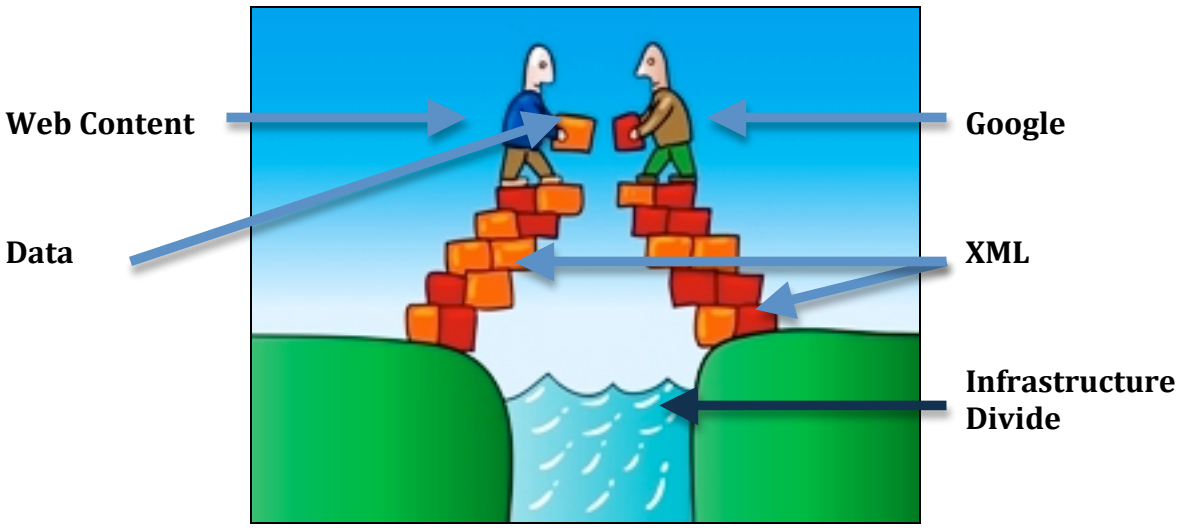
> **Google Spider:** "Hi, I'm Google. What's this page about?"
>
> **Web Site being indexed:** "This page is about…"

---

[2] "Understanding Infrastructure: Dynamics, Tensions, and Design" Paul Edwards, et al. Page 15. NSF Cyber Infrastructure Workshop, 2007. http://www.si.umich.edu/InfrastructureWorkshop/documents/UnderstandingInfrastructure2007.pdf

The infrastructures in place were not specifically designed for aforementioned type of interaction, so Google resorted to a "dirtier" method of indexing which was to do some sophisticated content analysis on parsed HTML.

Now, as Google is prominent figure in the information economy, standardized methods of presenting web content are more widely adopted by web developers in order to make their content more accessible. According to the Google Web Master Tools website[3] web site developers can provide sitemaps in XML format to ensure the proper content is indexed when the site is crawled. In this case, XML can be considered a Gateway technology as it bridges part of the gap between web content infrastructures (HTML, MySQL, etc.) and Google's content database. Other gateways technologies of this type have also been widely adopted to meet search engine providers demand for standardized content retrieval such Really Simple Syndication (RSS) and the Atom syndication format. Both of these "syndication" formats, known as feeds, use XML as a standard for formatting data and since web content providers can produce them and web content crawlers can read them, RSS and ATOM can be seen as gateway technologies.



[3] Web Master Tools Tour. Google Inc. Accessed 2/15/2009. http://www.google.com/webmasters/tour/tour4.html

This image represents XML as a gateway technology between search engine and web content infrastructures.

XML can be described as Meta-Generic Gateway in this case because rather than specifying exact standards for information transfer it provides a framework for each use of XML to specify its own standards. Such can be seen when comparing RSS and ATOM feed standards: both employ XML but produce different code and have different standards for the same content.

It is important to note the trade-offs between the two gateway technologies used to bridge the infrastructure divide. The first, Google's "crawl and scrape" technology, relies on sophisticated algorithms to infer meaning from parsed content. The interaction might be perceived as something like:

> **Website Content:** "My mother was a cook on TV. I remember her first show… 'Today on the show, we're going to make blueberry pancakes'"
>
> **Google Analysis of Content:** "Cooking website; blueberry pancakes; …"

Where as with XML feeds (Sitemaps, RSS, ATOM, etc.) the interaction might be perceived as something like:

> **XML Feed:** <content> My mother was a cook on TV. I remember her first show… "Today on the show, we're going to make blueberry pancakes" </content><content_description>Memoirs of my mother's cooking</content_description>
>
> **Google Analysis of Content:** "Memoirs of my mother's cooking"

For the Google search engine user, it would be much more useful to have the second example rather than be lead to believe they were being sent to a cooking website (the probable outcome of the first example).

Bringing this back to David and Bunn's question of "What forces us to choose a gateway or infrastructure technology", we can see how market pressures (in this case: the need to index and search the web) caused interested parties, such as Google, to forgo standardization considerations when developing their gateway technology. Not enough web content developers adopted the XML standard of information presentation, which would have elicited more precise search results as

demonstrated in the above examples, so Google developed an alternative gateway technology. Now that Google has more influence on the web content providers due to the public's adoption of their service, more web content providers are adopting more precise methods of presenting content on their site in order to appear higher in search results or at least in relevant search results (cooking website versus memoirs of a mother).

The role of flexible standardization within gateway technologies in this case served to increase the value of data being passed through it by making it more relevant. However, when standardized gateways are not readily available, the market will adapt if there is a demand and sometimes use less precise methods within their gateway technologies to bridge the gap.