

## Chapter 2

### **Digital Libraries and the Web: Origins, Impact, and Evolution**

In the previous chapter I claimed that the choice of the term “digital library” influenced the information model adopted by the digital library research community and shaped the technical components of digital library applications that realize that model. I further claimed that this shaping has had an impact on the coexistence of digital libraries with the remainder of the web information space.

This chapter explores those issues in greater detail. It begins by describing the motivations underlying the reuse of the term “library” to describe the new digital information research area. As will be described, those motivations reflected the expedient interests of each participant community and their respective definitions of a library. It next describes the influence of that term on the architectures and applications that were produced by the digital library research program over the subsequent years. The chapter continues with an explanation of how the architecture of those digital library applications impacts the technical coexistence of digital libraries with the mainstream web. Finally, the chapter describes how the evolution of the web from Web 1.0 to Web 2.0 has not only increased the technical incompatibilities between the two information environments, but has led to a fundamental conceptual difference in their information models.

#### **Origins of the “digital library”**

The decision in the early 1990’s to extend the notion of the library forward into the emerging digital information context was the result of the collective and distinctive

assumptions of three stakeholder communities: the funders, the technology-focused researchers, and the practitioner library community [82]. Each community responded to the opportunities offered by emerging networked computing technologies in a unique, opportunistic (and sometimes myopic) fashion. While they all agreed that “digital libraries” was an appropriate term for the new endeavor, they each had a different idea of the meaning of the term and different allegiances to the components of what they considered a library.

The following sections demonstrate this “interpretive flexibility” [468] by describing the different meanings attributed to digital libraries by the three major communities involved in the research effort.

#### *Perspective of the digital library funders*

The primary funders of digital library research in the U.S. were the NSF, within the Directorate for Computer and Information Science (CISE), DARPA, NASA, and NIH, with lesser contributions from NEH, IMLS (Institute of Museum and Library Services), and the Library of Congress. The notable characteristic of all the primary funders is their focus on technology-oriented science, in contrast to social science, humanities, or arts. This focus is reflected in a research program that funded mainly core computer science and its applications to networked information, with very little attention to network information as a sociotechnical phenomenon. In fact, the first phase of Digital Library Initiative (DLI) funding [379] was exclusively technical. This was moderated somewhat in the second phase [380]. Influenced by the results of a 1996 NSF workshop that called for increased research on the social aspects of digital libraries [83], the NSF included social science research in the DLI-2 solicitation.

An examination of documents published early in the digital library effort reveals the underlying assumptions and biases of the lead agencies that shaped the nature of the funding programs and their vision of the digital libraries that would emerge from it.

Although it appeared in a visionary 1988 document from Kahn and Cerf, the term “digital libraries” was introduced into the national research agenda in February 1994 in a report from a task force on Information Infrastructure Technologies and Applications (IITA) [257]. This report was commissioned by the newly funded High Performance Computing and Communications program, which was formed to leverage advances in computing and networking for the general social benefit [109].

The report defines digital libraries as follows:

[Digital libraries are] both technologies and applications which will lead to significant advances in the generation, storage, and use of digital information of different kinds across high speed networks. *A digital library is a knowledge center without walls, open 24 hours a day and accessible by way of a network.* Research areas range from advanced mass storage, online capture of multimedia data, intelligent filtering, knowledge navigation, effective user interfaces, system integration, to prototyping and technology demonstration. [257] (emphasis added)

The list of research areas enumerated in this statement is notable for its omission of the implications of eliminating the “walls”, or a being open “24 hours a day.” Clearly, the impression of the task force, or at least the only concern, was that the transfer of information to an online form raised only technical issues and that the larger social issues raised by this transfer were either inconsequential, unforeseen, or not worthy of study.

Another report from the same era, authored by Gladney and Fox, demonstrates prevailing thinking of the time that perhaps underlies this decision by the funders to focus only on technical issues.

The concept "library" has been refined over several centuries. It would be injudicious to depart from what people expect merely because a digital

service is replacing a material one. Except where explicit reasons suggest an improvement that is easily explained to ordinary users (e.g., in query services), library services should implement a familiar model<sup>16</sup>. [211]

Implicit in this text is the assumption that the nature of the institution and the information model it entails should remain as a stable overlay on a changed technical foundation. A digital library should, by nature, imply the same notions of integrity, trust, and quality, historically associated with libraries. Books might turn into bits, the catalog might become an online database, and shelves might turn into repositories, but the values, structures, and practices of the “institution” should be based on a “familiar model.”

Even as the web continue to grow in importance and scale, the notion that digital libraries were the focus for “serious” information-oriented activities persisted. The web was relegated to a more lowbrow status – an unfiltered mishmash of questionable and frequently objectionable content. For example, a 2001 (U.S.) President's Information Technology Advisory Committee report called the web a “rudimentary” information environment that “only hint[s] at the future of digital libraries” [409].

Even if the funding agencies had decided that the broader implications of the transfer of information to the online environment deserved investigation, it is doubtful whether they were structurally configured to handle such investigations. In a retrospective on the Digital Libraries Initiatives, Griffin [219] takes note of the problem that agencies such as the NSF have with research that is by nature long-term: “The program funding

---

<sup>16</sup> It is interesting to contrast this quotation with one from the position statement of noted futurist Esther Dyson in a workshop at roughly the same time: ‘What is the digital library? That term smacks of “filmed play,” “horseless carriage,” and the like. The digital library will be less like a library than we think, and more like itself. [191]’

models did not work optimally, particularly for the mid-size, longer-term, interdisciplinary research and test bed projects.”

*Perspective of the computer science research community*

The computer science research community had every incentive to follow the funding agencies in this selective interpretation of digital library research. The DL initiatives were a new and relatively large stream of funding for extending their pre-existing database and information retrieval research into a new application area [82]. As stated by Paepcke, et al.: “[The computer scientists] could see, or at least imagine, *how current library functions would be moved forward by an injection of computing insight*” (emphasis added).

The computer scientists who dominated DL research had little interest in examining the nature of “current library functions” or in understanding how the “injection of computing insight” might affect the foundations of these functions. The library was really only a convenient platform for technically focused work.

Indeed as Paepcke, et al. note the computer science researchers had little patience for the less technically manageable aspects, and “nagging downsides” of digital library research. For example, issues related to copyright and intellectual property were perceived as an annoyance that interfered with work on more interesting technical problems. Furthermore, the work often required collaboration with librarians who seemed overly focused on metadata “that the computer scientists felt would be replaceable by just another clever search algorithm improvement” [398].

In hindsight, the attraction of the computer science research community to the field of digital libraries was really not based on special allegiance to the library notion, but was just a case of following the funding. When the funding disappeared and it became

obvious that the web was a more attractive, and less restrictive environment for studying and exercising emerging computer science techniques such as machine learning, many of the former prominent members of the digital library community disappeared [398].

### *Perspective of the library community*

The “real” librarians, those with over a century-long tradition collecting, curating, and preserving books and other materials, entered the realm of digital libraries with a considerably more institutionally-focused definition of the library and vision of what the digital library would look like. From their perspective the library, as an institution, had successfully managed previous transitions to new media (the transition of the printed form from scrolls to the codex book to the printed book [117, 389], the inclusion of recordings, etc.) and had a track record of incorporating new technology into established practices, such as the computer-based catalog [78]. The “digital” library would be just another library and through all, the venerable institution would prevail:

The functions of the librarian have always been to select the material that his constituents will require; to catalog it so that those who would use it can know what is available and where it is; and to preserve it so that both contemporary readers and those who will follow will be able to use it...none of these tasks will disappear with the emergence of the electronic library. Somebody will have to perform them: if not the librarian, then his replacement. The anarchy of the Internet may be daunting for the neophyte, but it differs little from the bibliographic chaos that is the result of five and a half centuries of the printing press. [332]

Because of this allegiance to the institutional basis of the library and the belief that it was a necessary component of a useful information environment, librarians vigorously resisted the encroachment of the web on the domain formally dominated by the library. As noted by Paepcke, et al. “For librarians the intrusion of the web into the work on digital libraries was much more difficult to integrate” [398]. Initially, they

were largely dismissive and disdainful of the web as a serious information space, declaring that the “web is not a library” [220] and likening it to a bookstore in which “the entire stock is just piled up in the middle of the floor” [140]. As the amount of valuable content increased on the web, they responded with efforts to fold the web into standard operations, such as cataloging [391]. While these efforts to catalog the web were ultimately abandoned, they demonstrate how persistent traditional practices can be even in the face of a rapidly changing technical landscape.

In summary, the application of the library concept to the uncharted and unruly context of networked information reveals the distinctly narrow and flawed assumptions of the three parties responsible for its origin and use. The funding agencies mistakenly assumed that they could fund (and shape) the development of a new information infrastructure as a mainly technical endeavor. The computer scientists followed suit by framing digital libraries by-and-large as applications of familiar distributed database problems in which “predictable, repeatable ... access and retrieval is a prime value. [398]” In fact, the unpredictability of the web and its seemingly autonomous dynamism has not only affected our perceptions of information use and management, but it has had a far-reaching effect on computer science shifting it from its deterministic, algorithmic foundations to a more probabilistic and socially-oriented focus [268]. Finally, the librarians assumed that they could safely wrap radically new technology and traditional organizational values and structures in the same embrace.

In combination, these flawed assumptions lead to a research area that by and large treads the middle ground. At a fine granular level, it produced a number of interesting research results and applications of those results. But, at the higher level, it failed to explore the more far-reaching questions of how putting information online and giving people power over their information might change the nature of the information and

the way people use it. It is useful in closing this section to quote Agre [17] who, in his essay about “Information and institutional change”, spoke of the dangers of naively mixing historical forms with innovations:

A concept of “library” that is too fully rooted in past historical forms will make innovation impossible, but a superficial concept of “library” that draws out only a few aspects of those past historical forms (for example, a library as a big container of documents) will pass over phenomena whose absence in a newly designed system may be fatal. The middle ground between the maximal and simplistic conceptions of “library” is enormous and is not easily mapped.

### **Influence of the library on digital library technology**

Table 1 - Comparison of digital library and web architectures

	Digital Libraries	Web Architecture
Core Architecture	Repository-centric	Resource-centric
User Model	Portal Searching	Browsing
Content Model	Digital Objects	Resources
Indexing Model	Surrogates	Full-text and links
Identification	Persistent IDs	URIs
Federation Model	Federated Search, Metadata Harvesting	Centralized Indexing

The previous section described the set of assumptions about libraries and networked information that led to the choice of the term “digital libraries.” This section describes the manner in which that term and the presumptions underlying it have affected the nature of the technical artifacts produced by digital library research. This effect is reflected in both the overall architectural framework and on the individual architectural components of that framework. The contents of this section are summarized in Table 1.

### *Core Architecture: Repository-centric versus Resource-centric*

Digital library systems are by and large based on the notion of the institutionally-managed *repository* as the central architectural entity. The repository acts as the container for storage of and access to “digital objects” [255], the content “within” the library. In this manner, the repository is a virtual boundary defining the locus of institutional management, curation, and preservation of the contained digital objects. This virtual boundary is the functional equivalent of the physical boundary in the “bricks and mortar” library in which the physical structure defines the limits of library curation and stewardship of the information resources within it.

In contrast, the *resource* is the central entity in the web architecture [246]. Uniquely identified resources are the nodes in a virtual directed graph, in which the edges are the hyperlinks that connect resources. Notably absent from this graph model is the notion of containment or location. There is no first-class entity that corresponds to the repository in digital architecture. Although repositories are sometimes compared to websites, the comparison is incorrect due to the nature of the latter. A website is an ambiguously defined, second-class technical artifact – it may be all the web pages served within the same DNS domain, or those accessible through a single server. Technically, it has no identity (URI) and therefore it cannot be the target of any protocol requests. Conceptually, it does not imply control or management in the same manner as a repository.

The remainder of this section describes the major components of digital library systems that support this repository-centric architectural core.

### *Portals*

The portal, or the “front door of the digital library”, serves the same purpose as the physical entry to the traditional library. It provides the user with the clear notion of

being “inside” the digital library. Services and content within the portal are thereby blessed with the imprimatur of the library, endowing them with a level of trust and integrity. This is commonly known as “branding”. Correspondingly, most digital library applications clearly indicate to the user when they are “leaving the library”, for example by traversing a hyperlink to a page outside the boundary of the library.

The focus of a portal is usually a search interface, that in most cases is field-based, providing more functionality than the single text box search paradigm employed by most web search engines. This allows users to search on specific bibliographic fields such as title, author, or subject. This search paradigm reflects the influence of the library cataloging tradition [452], which eschews simple keyword searching that is predominant in mainstream crawler-based search engines (e.g., Google) in favor of more targeted search capabilities. Metadata, which is the basis of this field-based searching, is described in the next section.

In contrast to this “front-door” paradigm, the web user metaphorically “surfs” among linked information resources without regard for their location on the network. The informal notion of a “homepage” for a website does exist, but there is no presumption or enforcement of this as the uniform entry point to the collection of pages of that site. The notion of uniform, location-independent sources has proven to be quite powerful. In its simplest form it makes it possible to aggregate information from multiple sources in a single webpage, in the manner that a page may include an image that is stored in some other location on the net. As I will describe later, location independence is leveraged in Web 2.0 in a much more powerful manner in the form of “mash-ups.”

### *Metadata – cataloging in the digital context*

The shaping effect of the library tradition on digital libraries is perhaps most evident in the focus on descriptive metadata. This focus has its roots in cataloging, one of the core functions of the modern library [146, 158, 201, 213, 336, 452].

The traditional catalog developed for a number of reasons. At the simplest level, in a library of physical resources it gave users an easy and compact tool for finding information resources without having to traverse the shelves. However, describing the catalog as merely a compact shelf list trivializes its complexity and intellectual content. Underlying cataloging is the concept of information entities having uniform attributes, such as author, title, or subject classification, and the utility of those attributes for logical organization of those entities. This organization presents multiple *access points* based on those uniform attributes, and allows users to search and browse within those access points [452]. For example, a user may search for information by author name and then traverse the resources associated with that author, or alternatively they may search for information by subject classification and traverse the resources associated within that class. As a result, the organization of the catalog and the manner in which it is made available to the user (e.g., cards in drawers or screens in an electronic catalog) is independent of the manner in which the physical, or digital, resources are organized on shelves, or in repositories. Furthermore, an individual information resource (e.g., a book) may have multiple catalog instances, each accessible through specific access points (e.g., title, author, subject, etc.).

Efforts to extend the practice of cataloging into the context of online information reflects an ongoing belief that in a world where even the books that are part of library collections have been digitized and are available for full-text search [261], structured search over surrogates is more functional and ultimately preferred by users. This is

despite empirical evidence that users seem to prefer the “one text box” search paradigm of Google to the fielded-search paradigm employed in most digital library portals and online catalogs, and decades-old evidence of the frequent superiority in recall and precision of automated full-text search to human-assisted indexing and cataloging [128, 129].

Traditional library cataloging is both complex and expensive, especially when applied to the rapidly expanding and diverse set of digital resources. The notion of metadata emerged as a simpler and less expensive alternative to traditional cataloging records, perhaps making it possible for nonprofessionals to create structured bibliographic information. The predominant digital library metadata effort is the Dublin Core Metadata Initiative<sup>17</sup>, which is described in considerable detail in later chapters of this dissertation.

Ironically, the origins of Dublin Core lie in improving search and retrieval on the general web [483]. However, this effort to develop easy-to-use bibliographic standards for networked information objects has been deemed irrelevant, ill-conceived [165], or even counterproductive [168, 222], by the mainstream web community and most notably the search engines that dominate it. As I describe later, the attempts to translate the benefits of cataloging to the online domain via metadata have been compromised by problems with ensuring the quality of the metadata records that are produced by non-professionals and preventing so-called “metadata spamming” by unscrupulous agents trying to falsely lead information consumers to their sites [149].

---

<sup>17</sup> <http://dublincore.org>

*Digital objects – containers for complex data and metadata*

The content model of most digital library architectures is based on the notion of a *digital object* [255, 357, 403]; an identified (first-class) information resource that is an aggregation of multiple information units consisting of multiple formats, multiple subsidiary units (chapters of a book, issues of a journal), versions, or document components (e.g., the text, data, images, etc. of a scholarly paper). These are generally known as *compound objects*.

These object models reflect an ongoing effort by the library community to account for the complexity of information in both its abstract form and the physical or digital manifestations of it [103, 104, 336, 340]. These efforts have focused on mechanisms to represent the various relationships among information resources [452]; and to describe those resources at multiple levels of granularity [329], for various purposes, and in various descriptive formats [286].

Access to compound objects and their components is frequently mediated by protocols unique to the particular repository architecture. These protocols are usually embedded in the URLs that carry user requests from the digital library portal to the repository. These protocols allow operations such as access a digital object in a specific form, access a portion of a digital object such as its descriptive metadata, and the like. The proliferation of these architecture-specific access protocols has spawned a virtual cottage industry of repository interoperability initiatives [32, 234, 354, 393-396] in the digital library community.

In contrast, the web architecture [246] includes a quite simple information model based on the atomic resource. “Interoperability” is defined in the simple terms of the web architecture [246] – resources, URIs, and HTTP. There is no architectural notion of a compound object, or aggregation of resources. Ad hoc and de facto aggregations

exist, for example a logical document split into a set of interlinked web pages. However, these aggregations are not first-class objects; they do not have a unique identity and are essentially ephemeral. There is, in fact, an increased awareness in the web community that more complex information models are appropriate in a number of instances; for example, scholarly publishing. Chapter 12 describes our work to define one that is grounded in the principles of the web architecture.

### *Federation*

The issue of federation [330] arises because digital library systems are conceived as discrete institutionally-managed entities with distinct boundaries accessible to the user through branded portals. Sometimes, a user might want to search across multiple digital libraries when a selected resource is not available in their “local” library. In the physical library domain, this problem is solved by union catalogs such as WorldCat<sup>18</sup> and by interlibrary loan. Digital library applications employ two mechanisms to allow users to search for and access information outside the confines of a single digital library.

The first is federated searching or meta-searching, in which a single search query is multicast to several digital library search engines. The query is then individually processed at those search engines; the individual result sets are then returned and integrated at the site from which the original query was multicast. Federated searching was the subject of substantial work in the early years of digital library research [169-171, 194, 214, 393, 397] and Chapter 7 describes our own work in this area. Although instances of federated search still exist, the technique has fallen into some disfavor

---

<sup>18</sup> <http://www.worldcat.org/>

because of problems with dependence on the reliability of multiple search sites and the problems with the ranking of search results from several sources.

The second is metadata harvesting, in which bibliographic records from several distributed institutional sources are combined at a single indexing site, which provides a search interface across the resulting “union catalog”. The most widely deployed mechanism for metadata harvesting is the Open Archives Protocol for Metadata Harvesting (OAI-PMH) [314], which is described in greater detail in Chapter 10. That same chapter describes complications with metadata harvesting.

Neither of these techniques has achieved widespread deployment in the general web information space. As described in an earlier section, boundaries and the repositories that implement them are not a part of the web architecture. Web search engines such as Google crawl the web via graph traversal, ignoring the notion of the location of a webpage, except as a tool for optimizing graph traversal strategies [84]. In addition, ranking algorithms such as PageRank [95, 96] are designed to operate over a centralized index, and are difficult if not impossible in the context of distributed methods such as federated search,

#### *Persistent identity for network information*

A final example of the difference between digital library and web architecture is the notion of “persistent identity” for information stored in digital repositories. The attention to this issue in the digital library contexts reflects concerns about both preservation and control of intellectual property [400]. The Handle System [449] is the best known of this class of technologies. Like many persistent naming systems, the Handle System depends on a hierarchy of identity resolvers, and therefore the notion of a central *root* name server. These efforts have gained little traction in the mainstream web community, which has historically resisted centralization and has

comfortably adapted to the fragility of URLs, deeming identity persistence as a policy problem rather than a technical problem [51].

### **Coexistence of digital libraries and the web**

The previous section described the distinction between the repository-centric digital library architecture and resource-centric web architecture. In addition, it described how these different core architectural principles affected the technical components of each architecture. The digital library applications that have been assembled from these components are indeed quite powerful and include advanced searching capabilities, complex information models, and rich user interfaces. Ironically, the same architectural features that enhance their functionality have often interfered with the interoperability of digital library applications with the mainstream web and thereby mitigated the impact of these applications in the broader web context.

The problem comes from the fact that the specialized, repository-specific access protocols that provide access to these digital library resources often do not follow the conventions of mainstream HTTP access methods. For example, in many cases the URLs used to access objects are conflated with query predicates, the syntax of which is unique to the digital library and is hardcoded into portal/repository interaction. This is not a problem when access to the digital library resources occurs through the “front door” portal and through its respective search user interface that generates these query-based access URLs.

However, mainstream crawler-based search engines, such as Google, do not access objects through the front door, but rely on generalized graph traversal. The nature of the access URLs in digital libraries and their interdependence on the respective digital library search interface often makes these URLs unreachable via these graph traversal techniques. This is because the URLs of the digital objects in the repository are not

explicitly linked to, but are generated by the digital library based on search engine queries. These query-generated URLs are not visible in the web graph traversed by mainstream search engines and, as a result, the digital library resources are not crawled and are subordinated to an information black whole – the so-called “deep web” [49]. They fail to appear in result lists returned by mainstream search engines, which have emerged as the universal tool for discovery of information (much to the chagrin of the library community).

In an effort to increase search engine visibility, digital library providers frequently generate special link pages that expose the individual URLs of repository contents to crawlers as conventional hyperlinks. Digital library resources then appear as search results in Google and similar search engines. As a result, a steadily expanding amount of access to digital library resources occurs through these commercial providers<sup>19</sup>. But this reverse engineering to increase visibility of contained resources subverts the role of the digital library as a control zone, and the intention of the portal as a branded entry to that control zone. The “digital library collection” becomes just another set of web resources, with no joint identity or imprimatur. The digital library becomes “invisible infrastructure” [81], barely evident through a web-dominated information paradigm.

### **Digital libraries and the evolving web**

In addition to the technical incompatibilities between the digital library and web architectures that were described in the previous section, there is a widening gap between their underlying information models. The web that Tim Berners-Lee invented in 1989 has undergone an explosive growth in scale, measured in terms of number of

---

<sup>19</sup> Personal communication, J. Blake (NSDL) and S. Warner (arXiv).

URLs, servers, and traffic. At the same time, it has experienced a radical change in form and impact, referred to as Web 2.0 [390]. In contrast to the relatively passive and transactional search/access paradigm characteristic of the library and Web 1.0 in which the delineation between authors and consumers was relatively distinct, information interactions in Web 2.0 are highly interactive and participatory. Rather than just browsing and reading web pages, web users, acting as both authors and readers are writing reviews on Amazon, annotating and tagging pictures on Flickr, writing and updating articles on Wikipedia, publishing observations and research results in blogs, and mashing up online content into new content. This section examines the evolving web and the coexistence of digital libraries within that context.

The metaphor of versions – Web 1.0, Web 2.0, and the recently coined Web 3.0 – is obviously artificial and overly simplistic. However, it is a useful rhetorical device.

The predominant “features” of these versions are as follows.

*Web 1.0* – called the document web, the “web of cognition” [412], or the “read-only” web [39]. The time span of this version roughly extends from the invention of the web until 2000. It primarily consisted of hyperlinked, semi-static, atomic documents (HTML, PDF, GIF or JPEG images). Interaction and collaboration were minimal except for document authoring and querying. Content creation required specialized tools and, as a result, was restricted to a small subset of web users.

*Web 2.0* - called the “web of communication”[412] or the “read/write” web [420]. This “version” includes participation-oriented tools such as wikis, blogs, social applications like Flickr, and instant communication tools like Twitter. Another prominent feature is the notion of a “mash-up” whereby new information objects are created via the dynamic combination of existing information resources [432]. These features enable a phenomenon that Engestrom calls “object-centered sociality” [181,

182], in which information objects, people, and social exchanges are linked together in web space. This has effected a phase transition in the web's impact on economics, scholarship, learning, and most recently politics. The effect of this impact on national politics is exemplified by the recent observation that "... Barack Obama's victories in the Democratic primary and in the presidential election would not have been possible without Internet-empowered fund-raising and social networking" [126].

*Web 3.0* – called the “web of meaning” or the “contextual web” [130], this currently emerging web functionality incorporates concepts of the semantic web [22, 56, 185, 360], the underpinnings of which Tim Berners-Lee and the W3C have been developing since the late 1990's. The key features of the semantic web include machine readability and interpretation of web data and the ability to reason over that data. The technological foundation of the semantic web is the Resource Description Framework (RDF) [273] a data model for expressing statements about entities (web resources) and their properties (ontologically defined relationships).

Raffl et al. [412] adopt the language of Evolutionary Systems Theory [144] to illustrate how the features of these versions are cumulative: “[E]ach new layer is built upon a preceding one and ... the new stage comprises not only the new layer, but parts of the old one”.

Figure 1 illustrates the changing nature of the web through these versions. As shown, Web 1.0 was primarily a one-way channel from producers to consumers. In Web 2.0, the bifurcation of web participants blurs into consumer/producers who collaboratively author, manage, and annotate content. This is enhanced in the Semantic Web (Web 3.0) in which machines (agents) process and interpret this collaboratively produced content and contribute new content back on the web.

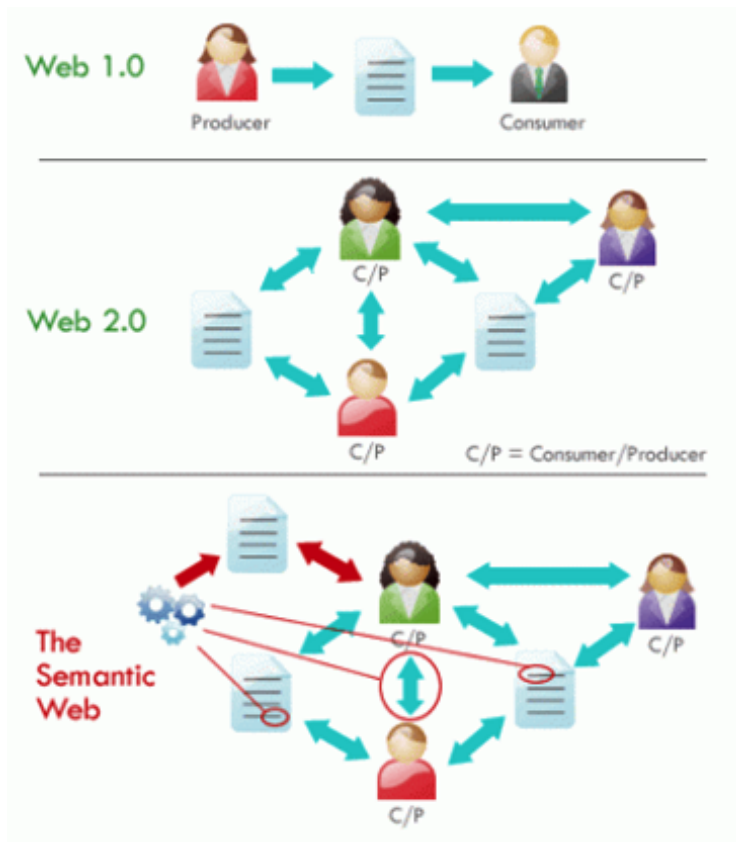


Figure 1 - Expansion of web functionality (from [130])

The participatory nature Web 2.0 should not be dismissed as just a popular phenomenon, manifested in increased use of mainly social sites like FaceBook. According to many experts in the field, we are witnessing a fundamental change in the prevailing information paradigm that is transforming all aspects of our culture. Notably, this impact extends to the nature of scholarly research and communication, one of the backbones of the research library. As pointed out by Paul Ginsparg, who create arXiv and is considered one of the icons of Internet-based scholarship:

... there are ... objective reasons to believe that we are witnessing an essential change in the way information is accessed, the way it is communicated to and from the general public, and among research professionals - fundamental methodological changes that will lead to a terrain 10-20 years from now more different than it was 10-20 years ago than in any comparable time period. [210]

Timo Hannay of Nature, one of the most prestigious scientific journals, makes the following observation of the impact of the current and future of the web on science scholarship:

For all but a very small number of widely read titles, the day of the print journal seems to be almost over. Yet to see this development as the major impact of the web on science would be extremely narrow-minded – equivalent to viewing the web primarily as an efficient PDF distribution network... Though it will take longer to have its full effect, the web's major impact will be on the way that science itself is practiced. [227]

In addition, there is mounting evidence that, for a number of important information-oriented activities such as education, the participatory information paradigm in Web 2.0 has advantages over the traditional consumption-based model. Fuchs and Raffl et al. [200, 412] argue that Web 2.0 paradigms of collaboration, construction, and participation are more closely aligned with recognized models of human cognition and knowledge development than the more restrictive and controlled library model. Downes [167] and Ullrich et al. [458] describe the utility of the Web 2.0 model for education because of the manner in which it facilitates activities such as group collaboration, exploration, and manipulation that are key to learning according to cognitively-oriented constructivist theories. Gee argues that the “affinity spaces” facilitated by the Web 2.0 environment are powerful tools for learning [206, 207]. Black identifies the notion of “beta-reading” in online fan communities where contributors grow as readers and writers based on mutual feedback [65, 66]. Finally, others see the general benefits of the “wisdom of crowds” [266, 450] that is enabled by the collaborative nature of Web 2.0.

### **Chapter Wrap-up**

This chapter described the origins of digital library research, the manner in which those origins shaped the technology produced by that research, and the compatibility that technology and the assumptions underlying it with the evolving web information

context. As described, digital libraries began with a rather simple assumption: the attributes, information model, and practices of the library could be translated relatively unscathed to the online environment. The prevailing belief was that the library would benefit from and be enhanced by the new technical developments, and the users of those libraries, the public, would in turn benefit from these “libraries without walls or operating hours”.

This early digital library work leveraged the concurrent development of the web. In its initial Web 1.0 form, it was primarily a technical system – a set of protocols and standards that enable browsing over a network of documents. In this form it provided an inert technical foundation for Digital libraries due to the fact that there was reasonable convergence between the web document-centric paradigm and the digital library document collection-centric paradigm.

However, as the web morphed into its 2.0 form it adopted a significantly more complex social nature overlaid on the core technologies introduced in Web 1.0. In this new social guise, the web was no longer inert, but profoundly active, embodying participation-based information interactions incompatible with many of the established concepts of the library. These concepts – institutionally-based boundaries or control zones, the document as a fixed unit of information, the unidirectional flow of information, and intermediation – are described in Chapter 3.

Whether the library as a meme or institution is flexible enough to adapt to these new paradigms and leverage their benefits is a matter of speculation. A set of initiatives known as “Library 2.0”, which intermingle traditional library services with Web 2.0 features, are now gaining in popularity in some elements of the library community [113, 358, 372]. Rather than comment on Library 2.0 in particular, I refer back to Christensen [122] who notes the difficulty of instituting radical transformations within

entrenched corporations and institutions [123]. Too often, these legacy institutions are burdened with continued support of legacy practices and with the demands of pre-existing customer communities. Furthermore, as mentioned earlier, the future viability of the library, digital or physical, rests not only in institutional changes, but also in modifications to public perceptions of it as outdated. Certainly, accomplishing both, especially the latter, is a formidable challenge.