The Cutting Edge:
The Next Generation Digital Library

Sayeed Choudhury
Johns Hopkins University

Let me begin by referring to Paul Conway's taxonomy of preferences related to technology. He referred to the adventurous types who bought the iPod Nano the day it came out, and now wear it around their necks like a badge of honor.

I bought the iPod Nano the day it came out.

But you'll note that I'm not wearing it in an effort to show it off, and I've also learned some interesting lessons for early adopters. First, you may have heard that the Nano screens are easily scratched; my Nano is no exception. So I'm finding that I need to buy a protector. Second, my wife asked me where I keep my Nano most of the time; I told her I keep it in my pocket. She immediately asked me why, because she is sure that my pocket is making the scratching problem worse. I told her I keep my Nano in my pocket because that's the most convenient place to keep it. Sadly, both she and I have to adjust our expectations, a sobering reality to consider for any early adopter.

As the title of my talk indicates, I'd like to consider the next generation digital library. Of course, this begs the question: what is the current generation digital library? This question isn't as obvious as it might seem. I've often noted that if you want to start an argument, ask a group of people to define "digital library." You may have heard the phrase "the good thing about standards is there are so many of them." Well, I also believe the good thing about digital libraries is there are so many of them. This diversity of opinions and perspectives is healthy, but for the purpose of this talk, I ask that you accommodate me by agreeing with a few definitions. If we don't do this, we could spend the entire session arguing about terminology, scope, etc. While that's an interesting discussion in and of itself, I have been asked to provide a specific set of observations and ideas, and I can't do this unless we agree on a few terms.

At Johns Hopkins University , in 2003, the Mellon Foundation awarded the Sheridan Libraries a grant for a strategic planning effort related to digital programs.[1] Through a series of visits to peer institutions, and discussions within Hopkins' libraries and central IT unit, we developed a framework for identifying appropriate elements for digital programs. During this strategic review, I wasn't surprised to find out that every institution felt they had inadequate resources to deal with digital programs. I was surprised to hear them assert that external funding often introduced complications. It seemed strange that the presence of funding in a resource-constrained environment would be viewed as anything but a blessing. This realization made me focus even more

so on core principles or elements of a framework for digital programs. What are some of the core elements of this framework?

First, libraries are service organizations and our primary mission is to serve our customers. Please note that "customers" is the first term I ask you to accept. When I say customer, I mean the faculty, students, staff, individuals whom your institution serves. You may prefer the term "patron" or "client" or "user" or prefer to state directly "faculty" or "students." Broadly speaking, I encompass all of these concepts in the term "customer."

Second, our Dean of Libraries, Winston Tabb, has correctly asserted that libraries are built on three pillars — collections, services and infrastructure. While each is essential and important, one could assert that collections are the core pillar — what services would one provide without collections?

Finally, Don Waters from the Mellon Foundation has described John Henry Newman's "Idea of a University" that considered the essential or core elements of a University. From this essay, one could submit that the four core elements of a University comprise research, teaching, dissemination and preservation. Today, we might use substitute "learning" for "teaching" to emphasize a more learner-centered perspective, and we might adopt the term "scholarly communication" to connote a broader sense of dissemination. But these four elements remain largely valid, even in the digital age. Perhaps the principles for libraries and universities remain valid, but the practices might need to change.

My educational training lies within the fields of engineering and economics, so I tend to think in terms of numbers, tables, lines, charts, etc. I felt that an interesting and useful way to combine these principles was a matrix as follows:

Library Digital Programs

	Research	Teaching	Dissemination	Preservation
Collections				
Services				
Infrastructure				

This representation provides a high-level depiction of what might comprise library digital programs. Quite simply stated, if a particular activity does not fit into this matrix, one might question why that particular activity is being considered or undertaken.[2] As a community, we need to consider this entire matrix of needs, but individual institutions might choose to focus on specific aspects or elements. Other institutions might then consider complementary activities. As an example, Anne Kenney provided an excellent presentation of notable work at Cornell related to strategic perspectives or models for digital preservation. Before Hopkins , or any other institution, considered similar work,

we would be wise to examine Cornell's work. This isn't to say that multiple institutions should not focus on common problems; we need a diversity of strategies and approaches for all elements of digital programs. However, we should acknowledge that others may have forged ahead in meaningful, portable ways, and perhaps we might serve our local and global customers more effectively if we focus on other pieces of the puzzle.

So what about our customers? We can develop digital collections, services and infrastructure to support the research, teaching, dissemination and preservation needs of our customers, but we need to think carefully about the changing, fluid, dynamic needs of our customers. I read two books that relate to this topic, ***Growing Up Digital*** by Don Tapscott and ***Playing the Future*** by Douglas Rushkoff. It's been several years since both books were published, but I still find them useful and relevant perhaps because both authors focused on trends and principles, rather than specific predictions of the future.

In his book, Tapscott identifies characters of the so-called "net generation." In particular, he asserts that they exhibit fierce independence, free expression and strong views, which are often supported by collaborative query-based reasoning and learning. For their information seeking behavior, there appears to be a shift toward non-textual or multimedia content, and they expect seamless access to content and services. When I read about these characteristics, I was struck by the question of how similar these trends might be for our faculty.

Rushkoff takes a more free-flowing approach in his book; it's almost as if he's trying to provide an analysis in a style that respects the cyberpunk culture. When thinking about how to characterize his book, I ended up picking this quote from the Amazon.com review:

> Do "The Simpsons" represent a leap forward in media consciousness? Do Sega video games and channel-surfing offer new strategies for coping in a world fraught with unpredictability? Can raves, snowboarding, or online chatting teach us something about adapting to cultural change?

There's an interesting discussion about snowboarding in Rushkoff's book. I should mention that I've never skied or snowboarded so I'm out of my element. Nonetheless, as far as I can tell, snowboarders look down on skiers because they follow pre-defined paths or tracks, and success often comes from emulating the "right" approach or technique. Snowboarders, by contrast, often seek new paths, or define new tracks simply by virtue of mastering them for the first time. When a snowboarder was asked about jumps, he used the following phrase: "…it is probably always a good idea…" I don't even want you to think about what else he might have said; just think about the use of the term "*probably always*." What does this mean? I have interpreted it to mean that snowboarders seem to approach each jump with a sense of what needs to be done, but they are also quite aware that each jump may require a special adaptation or trick. One has to be ready to change from what "always" works.

What jumps might we have to make? There's a participatory movement afoot. If you

haven't had a chance to examine [Wikipedia](#), [Open Directory](#), [Flickr](#), or [del.icio.us](#), I would encourage you to do so. Each of them reflects a desire from our community to get involved in organizing information. I realize this raises all sorts of implications, but shouldn't we tap into this energy, even enthusiasm? Should we build a wikibrary?

I mentioned the shift toward the non-textual or multimedia. According to Nielsen, 76 percent of active Web surfers access the internet using a non-browser based application such as a media player or a file sharing application — and this was the case in December 2003. If you want an interesting search for Google, just type "Indian hole in the wall." The links you'll find describe an experiment by an Indian researcher to embed a computer and monitor into a wall in a poor neighborhood in India. The children, who were almost certainly illiterate, learned to point and click within minutes. Soon thereafter, they were browsing the web (one sobering observation is that the Disney website was one of the most popular destinations). Once the researcher had installed speakers, the children figured out how to play audio and video files. Are we beginning to deal with a generation of individuals who consider reading a quaint habit of the past?

I also want to mention data more to the point, a data explosion. The Principal Investigator for the Virtual Observatory (VO), Alex Szalay, is a Professor in the Physics and Astronomy department at Hopkins . Alex likes to say that he's "living in an exponential world" when considering data. The data that VO acquires or processes doubles every year; it currently comprises terabytes of data, but it will soon comprise petabytes. This exponential growth in data has launched new lines of inquiry, collaboration and discovery. But it has also raised significant challenges for curating the data. Astronomy, in many ways, is the vanguard in this regard, but other scientific disciplines are also experiencing this data-driven scholarship. I would assert that the humanities and social sciences might also move in this direction.

The last trend I'll mention relates to rich media by which I mean simulations or virtual reality-like environments that provide "immersive" experiences. I should admit that I've been fascinated with video games since I was a child. I can recall games that were text-based and quite crude. My friends and I used to play a game that used ascii and line art, with a question and answer format. Interestingly, it did not have a manual or instruction set. So we had to find our way, often with great frustration. In one encounter, I approached a room with an ogre in it. My exchange with the computer was something like:

> **Me:** Kill Ogre.
> **Computer:** You can not kill the Ogre.
>
> **Me:** Attack Ogre.
> **Computer:** You can not attack the Ogre.
>
> **Me:** Enter room
> **Computer:** You have entered the room. The Ogre is attacking you.

Me: Kill Ogre.
Computer: You can not kill the Ogre.

Me: Attack Ogre.
Computer: You can not attack the Ogre. The Ogre has killed you.

While this may seem frustrating, even ridiculous, we enjoyed many hours of entertainment and, dare I say, learning and critical thinking through such exchanges. Games have come a long way since then, but I would submit that the same characteristics that made these early games so compelling are still present today. Yes, I know that the "first person shooter" game probably offers little in terms of educational value, but even those games can highlight opportunities for exploration and collaboration. The "serious games" movement is examining the potential for games as a means for learning. What are the implications for libraries in this type of an environment? There are other efforts such as [Croquet](#) and [Geowall](#) that don't require the same overhead as a fully developed immersive game, but may offer the same potential for learning and exploration.

What does all this mean? I would argue that collections (content) are becoming recombinant, by which I mean that the "fixed" notion we've become familiar with is being challenged. People think of content as malleable, something that can be broken into smaller chunks, shared and repurposed or transformed. Projects such as the VO provide evidence that collections and services are merging. That is, astronomers think of particular services when they acquire new collections, and they think of what "other" collections might benefit from a set of services that were developed with a particular collection in mind. In such a fluid environment, I would argue that we need infrastructure that is open and modular, allowing us to identify specific components — either open source or vendor-based — which can be replaced as we need. The line between research, learning and dissemination is blurring — or has always been the case and we're only returning to the original vision of higher education?

I have to mention a few words about preservation. I really enjoyed and appreciated Anne Kenney's presentation. She mentioned that a vast majority of institutions considering digital preservation are using in-house systems or software. One could assert that this reflects a relative paucity of options in this regard. That is, if one wants to manage and provide access to content, there are several paths, including commercial options, to consider. There are considerably less commercial choices for digital preservation. I'm struck by the amount of attention the library community spends on access to digital content. Of course, this is an important issue. But isn't the "core" mission of so-called memory institutions the preservation of knowledge? One of the most useful aspects of the digital world is that the mechanisms we might adopt to support digital preservation (e.g., the repository) also offers channels for access. When it comes to digital preservation, I don't think there's much competition for libraries and there's compelling need. We might engage our customers more readily if we offer to preserve their content, and then offer services that provide access.

In the abstract for this presentation, I mentioned Google Print (now known as [Google Book Search](#)), so you're probably wondering how I think it fits into the picture. One question I often like to ask relates to Gopher. For those of you who may not recall Gopher, it was a text-based hypertext system. About ten or twelve years ago, many institutions were running Gopher servers. Today, I would be surprised if any institution in this room continues to run a Gopher server. Gopher was a wonderful resource, but it became completely superceded by the Web, which offered enhanced functionality. There's an interesting lesson to consider from this shift. Gopher used open standards and protocols, so it was very easy to migrate Gopher-based content onto the Web. This is good! But transferring Gopher content verbatim onto the Web usually resulted in a fairly "boring" representation because it was textual only. Since most of us didn't imagine sharing images, we developed Gopher sites without visual considerations. So, it was possible to migrate content because it was open in nature, but I'm willing to bet that most people added images or other visual elements to websites that featured migrated Gopher content. Or perhaps built new websites that simply substituted for previous Gopher sites.

I have a theory that Google Book Search may be the Gopher of the current age. I do not doubt that Google has lots of ideas for services that can be built on top of a large corpus of digitized text, but if you take the project at face value — a large-scale book search of digitized text — it may have a similar effect to Gopher. It will excite us, prompt us to think of interesting and useful services, but it will only whet our appetite for even more. I also think of Google Book Search as a hurricane — something potentially disruptive, even destructive, but it can be tracked reasonably well and we can consider approach responses or measures to deal with it. I'm worried about earthquakes — sudden, unpredictable shifts in the landscape. I can't resist mentioning the engineers' take on this. For wind events, one designs stiff, rigid structures that remain upright, but fail in catastrophic ways if overwhelmed. For earthquakes, one designs ductile, flexible structures that can ride out the waves, acknowledging that some damage might occur but success constitutes survival and the structure maintaining its integrity.

In closing, I offer the following advice — Don't Panic. For those of you who are fans of Douglas Adams, you'll recognize this phrase immediately. If I had to offer a "Hitchhiker's Guide" to digital libraries, I would offer the following "chapters":

> *Focus on customer needs, don't argue with them*. I believe that librarians are amongst the most dedicated individuals I've ever met. I'm consistently impressed by the dedication to public service. As a technology person, I've learned a great deal about focusing on customer needs from librarians. Having said this, I have noted, on occasion, we're talking past our customers. When someone says "I used Google for my information needs" and we respond, "You should use the library catalog" it seems a bit like someone saying "I had a great pizza" and responding "You didn't have a full-course Italian meal." We are talking about different aspects of service. The question I would ask everyone to consider is whether Google is moving more quickly to providing access to "good" content, or whether the library community is moving

more quickly to making access to good content as easy as using Google.

Preservation is a cornerstone for libraries. If libraries do not make significant contributions, through their own efforts and collaboration with others, toward digital preservation, we run the risk of ceding one of the core services that libraries offer. Our customers are looking for help in this area, and we need to respond accordingly.

Repositories are the beginning the process, not the end. I worry when I hear people assert that they're offering digital preservation by installing a repository. One of my colleagues and I had the pleasure of being part of the NDIIPP technical architecture planning meetings, which resulted in the mantra "machines store bits, but institutions preserve." There are aspects, including policy issues, in addition to the storage of content that must be considered. Additionally, the current suite of options for repositories represents our best efforts for today and the near future. In some years, maybe even as soon as five or ten years, I plan on asking the "Gopher question" with regard to DSpace or Fedora or other repository software. The most important aspect of these software systems is that they allow us to manage content in ways that support long-term preservation, which will almost certainly include large-scale bulk export and migration into other systems.

Don't try to predict the future, embrace it. We've heard the expression that the best way to predict the future is to invent it. I have never accepted this perspective. I think predicting, or trying to invent, the future might be entertaining, but it's not very productive. I believe it's better to pay close attention to the environment in which we find ourselves, and realize, embrace, that it's fluid, unpredictable and evolving. If we keep an open posture, while retaining an understanding of our core principles, we'll probably always be OK.

---

1. The results of this planning study are available at:
http://www.library.jhu.edu/departments/librarydean/integration.html

2. At Johns Hopkins, we have defined a set of prioritization criteria for digital programs, which are described at: http://ldp.library.jhu.edu/documents/criteria