Bibliographic Control and Content Management on the Web

Debora Seys
HP Laboratories Palo Alto
HPL-2004-2
January 2nd , 2004*

Web Content Management is examined in order to understand what kind of changes this type of publishing will bring to our ability to provide bibliographic control and access for these new forms of text on the web. As authors create 'chunks' of content and software distributes it in various forms to multiple interfaces, each reader/end-user can now see a unique text. This creates what can be called a three-dimensional text. As this type of publishing becomes the norm, users will need less help in navigating to an individual bibliographic entity and more help in navigating the relationships between the content 'chunks' and the ideas and meaning they embody. In addition, some interesting bibliographic problems and solutions are also presented that can help us to understand and solve some of the questions that arise as a result of the revolutionary nature of publishing on the web.

Bibliographic Control and Content Management on the Web
Debora Seys
January 2, 2004

Introduction

We live in a three-dimensional world where objects take up space and have depth, yet this is not the way we usually think of written text, such as books or serials. Even after the move to a digital format, we still think of them as having only two dimensions, as flat and linear objects. A new process of managing information and publishing on the web, loosely called Content Management, has the potential for radically changing the format of written materials and moving us towards what might be called a three-dimensional text, one that is non-linear, assembled rather than constructed, with depth and multiple perspectives or instances, depending on the point of view of the user. We may find answers to the questions surrounding bibliographic control for these types of materials by looking at the characteristics of serials and ongoing materials as a starting point, and thinking about how a new approach might take us even into a fourth dimension.

Three-Dimensional World

In a fascinating little book written in 1884 entitled Flatland, (1) Edwin Abbott tells a story designed to illuminate the concept of physical dimensions. The protagonist, a simple square, lives in a two-dimensional world and shares his life with other flatland objects such as a triangle or a line. One day the square meets a sphere and begins the intellectual journey that introduces him to the concepts of a third dimension of space and

a fourth dimension of time. Here the Flatland square speaks of his encounter with the sphere.

> He then introduced me to the Cube, and I found that this marvelous Being was indeed no Plane, but a Solid; and that he was endowed with six plane sides and eight terminal points called solid angles; and I remembered the saying of the Sphere that just such a Creature as this would be formed by the Square moving, in Space, parallel to himself: and I rejoiced to think that so insignificant a Creature as I could in some sense be called the Progenitor of so illustrious an offspring.

The square then takes what he has learned a bit further and begins to speak with the sphere of a fourth dimension of movement and time.

> Or if it indeed be so, that this other Space is really Thoughtland, then take me to that blessed Region where I in Thought shall see the insides of all solid things. There, before my ravished eye, a Cube, moving some altogether new direction, but strictly according to Analogy, so as to make every particle of his interior pass through a new kind of Space, with a wake of its own-shall create a still more perfect perfection than himself, with sixteen terminal Extra-solid angles, and Eight solid Cubes for his Perimeter.

Borrowing from the book, we can create a mental picture and shorthand way of speaking of physical dimensions that can help us as we discuss web Content Management and the fundamental changes it brings to publishing on the web. See Table 1.
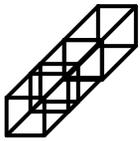
No Dimension	Pointland	●
One Dimension *(Length)*	Lineland	▬
Two Dimensions *(Length and Height)*	Flatland	■
Three Dimensions *(Length, Height and Width)*	Spaceland	
Four Dimensions *(Length, Height, Width and Time)*	Thoughtland	

Table 1 *Physical Dimensions*

This paper will attempt to make some useful analogies between the concepts underlying the physical notion of single and multi-dimensional worlds and the concepts that we use to build our procedures for the bibliographic control of static, ongoing, Internet and Content Managed publications.

Content Management

It is possible to find a variety of definitions in use for the term "Content Management," though it is generally agreed that it refers to the management of web-based materials. It is a current "buzzword" for vendors selling information and document management systems for the web and even some library software vendors have begun to modify their systems to meet the needs of a web environment. (2) The term is general enough to cause some confusion. See Table 2 for a brief comparison with other terms in current use.

Knowledge Management	The broad process of enabling the systematic use of the information and expertise within an organization by turning personal knowledge into corporate knowledge that can be widely shared.
Content Management	The process of developing, maintaining, organizing and deploying web content, in order to efficiently support the collaborative activities of content creators, site administrators and users.
Information Management	The application of information science principles to the administration of information resources to ensure that they are captured, organized, maintained and accessible.

Table 2   *Definitions*

Table 3 takes a closer look at the processes and technologies involved in each step of a Content Management effort. The primary purpose of Content Management is to control the unique needs of the lifecycle of information on the web.

Deploy	Publishing, Dissemination, Presentation, Aggregation, Push/Pull, Personalization, etc.
Organize	Directories, Categorization, Metadata, Markup, Taxonomies, Controlled Vocabularies, etc.
Maintain	Storage, Version Control, CRUD (Create, Read, Update, Delete) Access, Retention, Integrity, Security, etc.
Develop	Creation, Authorship, Selection, Generate/Prepare, Design, Approve, etc.

Table 3 *Content Management Processes*

As I have noted before (3), Internet materials are revolutionary in nature, changing rapidly and sometimes violently. Content Management technology and processes have developed in response to the labor-intensive work required to keep up with this rapid change. Content Management enables and facilitates the "revolution". By providing a set of technologies and a process, Content Management standardizes the

work and maintenance required by Internet materials. In fact, rather than maintain static pages, most Content Management systems create materials on demand, responding in real time to requests for text, data and formatting. By using technology to identify the user making the request, this 'on-the-fly' response can also be targeted to a particular user or class of users, presenting them with personalized text or material.

In an article entitled "Dynamic Content is King," Allen Elliott writes,

Furthermore, databases are evolving into Web-content repositories to replace simpler but more chaotically managed file-based Web servers. The content in all these systems requires new approaches to obtaining strategic Web value; providing static content in increasingly huge documents in an ad hoc manner is no longer viable. Users must be empowered to assemble and reuse streamlined chunks of content in ways that earn their loyalty and build awareness, and IT organizations must be able to manage content centrally and dynamically across the enterprise. (4)

Reporting on a recent meeting of publishing executives, the ContentManager email newsletter reports,

Content is increasingly being used for two different purposes: editorial and advertising, blurring the line between these traditionally opposed groups into products that can best be described as 'advertorial.' Publishers are using digital technologies to break down content into granular 'chunks' for multiple re-use, extending the life and value of content in ways they had never conceived of even a year ago. (5)

There are two fundamental changes described here. The first is the use of "Web-content repositories" and the second is the use of "granular chunks" of content. By

creating content in small pieces and then storing these pieces as if they were bits of data

in a database, a "text" can then be created at will, by producing a report from the database

that essentially glues together the pieces into a whole. Jon Bosak and Tim Bray write

about the "Second Generation Web" and declare, "This is actually an advantage for

publishers, who would often like to 'write once and publish everywhere' – to distill the

substance of a publication and then pour it into myriad forms, both printed and

electronic." (6)

Outsell, Inc., a research and advisory firm specializing in information content,

released a recent briefing on "Virtual Sourcing" describing the advent of content services

that are provided by remote access on the Internet. In it they describe the key

components of this model and what makes it work. One of these is the "Use of the URL

as a Universal API."

> One of the most significant features developed through browser technology is the ability to initiate
>
> programs, such as database queries, from the same mechanism that could be used to show simple pages
>
> of information. The Universal Resource Locator (URL), which we see at the top of most browser
>
> windows when we retrieve information, is extremely easy to format and program for communication
>
> with any information source. (7)

This delivery of individualized, assembled "content chunks" in response to a demand

from a specific user, shifts the publication process even further past what I've previously

described as revolutionary and takes us into a new dimension. (9)

Three-Dimensional Text

In this environment, we need to think about electronic text in a much more flexible way, in a three or even four-dimensional way (9), as we are no longer constrained by the flatland of paper or its digital equivalent, but have the freedom of space and time as creators and publishers take full advantage of the technology of the web. By these means, we can also think of the creation and publication process as moving from a flat and linear process to a lively, multi-faceted world.

Another way of saying this is that while static, ongoing (for example, most journals or regularly published proceedings, etc.) and most Internet materials might appear either at a single moment in time or evolve over a period of time, they both exist (physically and virtually) in a two-dimensional way. In contrast, the materials that are "Content Managed" take up space and time in a three dimensional way. Content chunks can be used to publish an unlimited number of different texts in a faceted way over an unknown span of time. Each piece of content has multiple possible relationships to the other pieces of content. It is the depth of the collection of content, the real-time publishing and the published material that together make up the three or four-dimensional text. The depth within the text can be seen, like the cube first described by the flatland square, in the unique perspective on the content provided for the user by the personalized publishing process. The fourth dimension of time is found in the potential of the ever-changing text as it is formed on demand from the raw material of the content – which is also likely to change over time. Table 4 captures this analogy.

Publication	Dimension and Quality
Static ⇩ Ongoing ⇩ Internet ⇩ Content Management	One Dimension *(Length)* Fixed
	Two Dimensions *(Length and Height)* Movement
	Three Dimensions *(Length, Height and Width)* Depth
	Four Dimensions *(Length, Height, Width and Time)* Thought

Table 4 *Publications and Dimensions*

Creating the Three-Dimensional Text

There is another, even more dramatic and fundamental change that occurs when materials are published through the Content Management process. Traditional and even Internet materials generally exist in an environment where the creator and publisher retain complete control, perhaps surprising the user with the speed and quantity of change, but still maintaining the same relationship to the user and the material. With the new technology of Content Management, the relationships change and the control shifts.

The changes can be seen in Diagrams 1 and 2, below. For static, ongoing and most Internet materials, the creator and publisher are the initiators, placing the text into the world to be used *as is.* In the new Content Managed environment, the user is the initiator, in a sense creating the text by requesting it and providing the context for how it will look and what "content chunks" will be assembled to create it. As we've seen, the creator writes the content, not the text. In addition, stylistic or formatting elements can

be included in the delivery of the content through the automated publishing process, but these are not necessary when text is created using a standard markup language and so the user may even have a great deal of control over the look of the final product. (10)
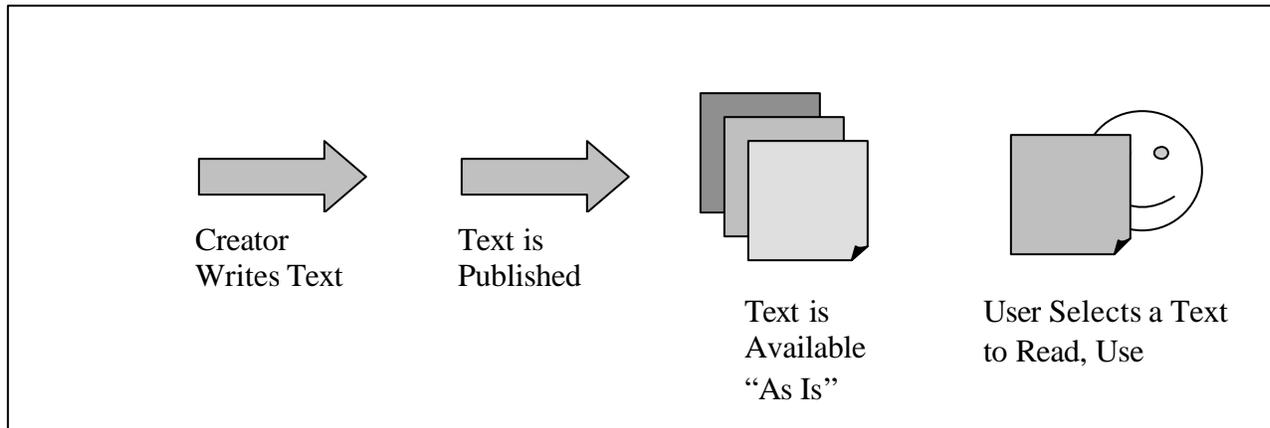
Creator
Writes Text

Text is
Published

Text is
Available
"As Is"

User Selects a Text
to Read, Use

Diagram 1 *Static, Ongoing and Internet Materials*

Creator Writes Multiple
Pieces of Content

Content Repository
("Content Chunks")
+
Personalization
Rules

User ID and Request

Publishing
Process/Tools

User ID +
Request

Users Receive, Read,
Use Personalized Text
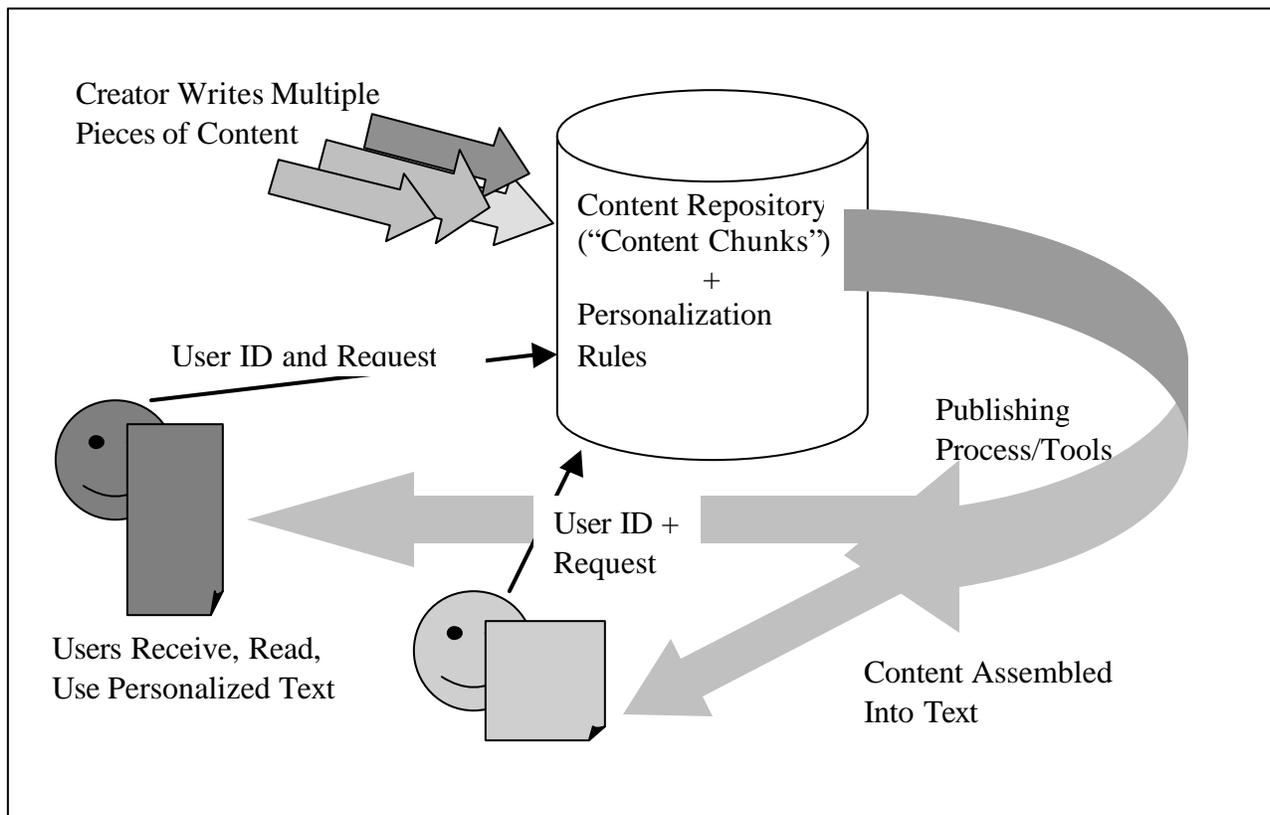
Content Assembled
Into Text

Diagram 2 *Content Managed Materials*

In fact, the future of digital information and text includes a variety of media and technologies that will increase the need for this assembly-line or self-service approach to publishing. For example, an article in Web Techniques describes HDML, which is the Handheld Device Modeling Language. As opposed to the web site 'page' metaphor, this information delivery model for a PDA (Personal Digital Assistant) uses a 'deck of cards' model, with each consecutive visible piece of information being a card and the entire set of information being sent over the network in a discrete, self-contained set or 'deck' of cards to be used as needed by the PDA. (11)

Bibliographic Control

Previously I've discussed the building blocks of the cataloging process (examination, description, transcription and access) and how these apply to the cataloging of static, ongoing and Internet materials (3). I showed how each step in the process is adapted from static to ongoing materials and how this adaptation provides a conceptual bridge to the radically different treatment required by Internet materials. See Table 5. Content Management as a process for managing information and publishing on the web will take these changes even a step further, yet I think we can continue to look to the characteristics and to the treatment of ongoing materials as a conceptual bridge to the future of bibliographic control. (12)

Static	Ongoing	Internet
Examination	Extrapolation	Creation
Description	Accumulation	Repurposing
Transcription	Supervision	Markup
Semantic Access	Evolution	Revolution
Structural Access	Relationship	Integration

Table 5 *Bibliographic Control*

In spite of all these changes, bibliographic control and access has generally continued to provide the user with a way of obtaining information about a specific bibliographic or two-dimensional item. How will this process need to change when we apply these same concepts to a three or four-dimensional text? To answer this question, let's take a moment to look at the function(s) of a traditional library catalog.

The Library Catalog

The traditional library catalog is the cumulative result of descriptive records for static and ongoing materials (and now also Internet materials) with the intended goal of collocating like items, by means of structural access points, such as author or uniform title, or by semantic meaning, such as classification or assigned subject headings. Library catalogs are created with conceptual elements in mind such as audience or community needs, purpose and relevance, giving the catalog an internal logic and making it more than just the sum of the catalog records.

For example, the accumulation of semantic access points in a catalog creates a useable and meaningful, coherent map to the knowledge contained in the resources described in the catalog. One could argue that this map or 'layer of meaning' is the true value of the catalog.

So, how do we retain this value when there is no text or material to catalog? A three or four-dimensional text is by its very nature impermanent and ephemeral, too slippery for transcription or structural access. It is clear that we will need to move to a process that is not dependent on the description and cataloging of an individual item yet does not disappoint the goal of creating a map of the knowledge or meaning contained in the content of the collection. Steven L. Hensen is quoted in a recent American Libraries,

> "Going back to Panizzi and Cutter, it has been axiomatic that bibliographic control was a matter of keeping track of, or inventorying, a specific physical object in a specific physical place. Today, those 'objects' are as likely as not to be in a variety of both physical and evanescent formats and in no specific physical place as we have been used to understanding these things. Given then that the physical nature of the work is increasingly meaningless or difficult to define, the focus should now shift to where it should have been in the first place – and indeed the focus from which the various Internet search engines derive their usefulness – and that is the intellectual content and substance of the work itself." (13)

While Hensen argues that the focus should 'shift to where it should have been in the first place,' I believe that libraries and other implementers of bibliographic control have been doing it all along, we just haven't been articulate about the theory or methods or very intentional about the way in which we create the results.

The Layer of Meaning

A recent article in Information Outlook on the changing tools and role of the Information Professional comments on several convergent trends around the growth of information and the Internet and predicts that the product of new software processing will not be fewer hits, but instead new forms of information that add value for users. The authors then recommend that Information Specialists should

> … analyze text not just to point us to precious nuggets (to extract a few really good articles), but to illuminate patterns in the full body of information. The issue here is not information retrieval – there is too much good information for that to suffice. The issue is to perceive the patterns – e.g., associations among particular concepts within the domain, producer emphases changing over time, new techniques entering the "fringe" of the domain with promise to change it greatly. (14)

I would argue that we can also gain a view of these 'patterns' of the meaning contained in the collection by moving away from the function of the catalog as finder for individual bibliographic entities and towards a vision of the catalog as a map or web of meaningful relationships as embodied in the content of the 'collection.' (15) In the case of Content Managed materials, we could imagine that the 'collection' is the set of content chunks that work together with a particular publishing process to create multiple faceted texts as we've discussed above. A user seeking information would use the 'map' to gain insight to the potential meaning of the entire collection without even needing to see any particular instance of a text at all. In this sense, cataloging becomes a task of identifying the potential of the aggregate. Bibliographic relationships as we have known them in the

catalog (16) would now include the relationship of one idea to another as well as one text to another. See Table 6 for some speculation on what this means for cataloging Content Managed materials.

Static	Ongoing	Internet	Content Managed
Examination	Extrapolation	Creation	Intention
Description	Accumulation	Repurposing	Potential
Transcription	Supervision	Markup	Publishing Process
Semantic Access	Evolution	Revolution	Layer of Meaning
Structural Access	Relationship	Integration	Instance

Table 6 *Bibliographic Control – Next Steps*

I've placed Internet materials, which represent a wide range of types, and Content Managed materials very close together because there is some overlap between them. So, for three and four-dimensional texts, what was once examination now becomes the need to understand the purpose of the collection of content, description of a specific item becomes the description of the potential of the aggregate of the content, the publishing process replaces the need for transcription and provides the structural access at the time of publication or instance of the text. Semantic access is broadened to include the entire potential content and becomes a meaningful layer or map to this meaning for the user.

We've seen before how the cataloging of ongoing materials provided a bridge to understanding how we might begin to control and provide access to Internet materials. Are there other characteristics of the type of solutions that have been developed for

ongoing publications (like serials, journals, proceedings, etc.) that can help us understand how to provide access and control over texts in the third and fourth dimensions?

Conclusion

One could say that the ongoing, continuous nature of serials and their frequency of publication are an attempt to provide a textual container and history of a type of professional communication and conversation. The good news is that I believe the changes to the publication process brought about by Content Management can move us closer to this ideal of text as a communication tool. In fact, paper has always been a handicap – too slow and too rigid to change with the kind of flexibility that serials need.

In fact, serials on the web are already using the flexibility of the new medium. The ways in which publishers of journals are taking advantage this new freedom from print, include daily file updates, cumulative publications and access to articles prior to formal publication. Eric Jul writes, "These examples are hardly exhaustive, but they do illustrate the new boundaries of seriality and periodicity. … If technology will allow it, serials will do it." (17)

Web technology can respond quickly to creators and publishers, providing the ability to modify content and text as needed to follow their shifts in thought or purpose. If we look to serials, we can see that they provide a kind of conceptual bridge to this new environment by the way in which an ongoing publication establishes a framework or boundary (title, statement of purpose, intended audience) and then allows for the interior contents (articles, editorials, essays, etc.) to shift freely. This framework captures a

moving target, so to speak, both of form and meaning. We can use this same concept of the frame, with a shifting interior view, to begin to define a conceptual basis for cataloging content and creating a 'layer of meaning' on the web. The frame in this case could be equivalent to a collection boundary.

There is a significant difference, though, that needs to be addressed. The typical serial or TV delivers prepared content that the user simply receives (albeit with some choice, but no ability to interact or to change). In contrast, as we have already noted, Content Managed materials are 'created' by the user who provides the context and the motivation for the creation of the text. This situation would need a new kind of frame or concept of a collection that could handle unpredictability, subject relationships (not entity relationships), a single user perspective, and a way for the user to navigate through to the content meaning rather than to the bibliographic entities.

This begs the question – are all Content Managed materials also serials because of their 'ongoing' like nature? In the future this distinction might be irrelevant. If most electronic text now is more serial-like than not – and if the "medium is the message" as Marshall McLuhan says, then what is the message these days? I don't presume to know, but I think the answer can be found by looking more closely at the function of text as communication. The immediacy and responsiveness, the flexibility and personal nature of the publishing process provide all the tools for an online four-dimensional conversation. How we capture the meaning and provide a meaningful context for what is being communicated will be a difficult, but not impossible question to answer.

Notes

(1) Edwin A. Abbott, 1952, *Flatland: A Romance of Many Dimensions*. New York: Dover Publications, 6th Ed., p. 89 and 91. An online version of the 2nd edition can be found at http://eldred.ne.mediaone.net/eaa/FL.HTM (May 5, 2001).

(2) David Dorman comments on the entrance of library software vendors into the "Content-Management Game", in "Technically Speaking" American Libraries, 32(3) March 2001:67-8.

(3) Debora H. Seys, "Speaking a Serials Cataloging Tongue: Lingua Franca for the Web?" Proceedings: Making Waves: New Serials Landscapes in a Sea of Change. P. Michelle Fiander, Joseph C. Harmon, editors. New York: Haworth Press, 2001. Also in *The Serials Librarian*, 40 (2001).

(4) Allen Elliot, "Dynamic Content is King" Intelligent Enterprise, Sept. 14, 1999:38-46.

(5) Content Management Network http://www.content-net.com Email newsletter. Tuesday, March 13, 2001.

(6) Jon Bosak and Tim Bray, "XML and the Second-Generation Web" Scientific American, May 1999. Online version of the article can be found at http://www.sciam.com/1999/0599issue/0599bosak.html (May 5, 2001)

(7) Outsell, Inc. 2001, "Virtual Sourcing, Who's Got the Content?" Information About Information Briefing, 4(9) March 30, 2001.

(8) One could write another entire paper here on the difference of writing for the web vs. writing for a traditional (be it print or digital) type of text. There are fundamental differences in the creation process for writers to create reusable content chunks vs. entire

Page 17

texts. Peter Morville warns of the difficulty in creating what he calls "virtual documents" in his article entitled "Virtual Documents: The Challenges of Chunking" in Webreview http://www.webreview.com/1999/01_01/strategists/01_01_99_3.shtml (May 5, 2001)

(9) For some interesting discussions of the fourth dimension, see Charles H. Hinton, "Fourth Dimension Writings" http://www.eldritchpress.org/chh/hinton.html (May 5, 2001).

(10) In the case of markup languages, the text can be formatted and styled at will by anyone who understands the syntax of the markup.

(11) Molly E Holzschlag, "Clip Art" Web Techniques, December 2000:32-35.

(12) As I've written previously (3), "For Internet materials, the moment of examination becomes instead a moment of creation as metadata and/or markup are an intrinsic part of the text. Description must bow to the needs of repurposing as it is impossible to accurately or usefully describe something that can change at any time. Embedded markup or metadata in the text eliminate the need for transcription. Semantic or subject access becomes a dynamic process of continual change or revolution and structural access becomes a tool for creating relationships in an interdependent environment. Meaning can no longer simply be examined or extrapolated; it must be based on potential or possibility much like one would follow a conversation in progress."

(13) Steven L. Hensen, "Quote of the Month" American Libraries, 32(1) January 2001:86.

(14) Nils C. Newman, Alan L. Porter, and Julie Yang, "Information Professionals: Changing Tools, Changing Roles" Information Outlook, March 2001:24-30.

(15) Hur-Li Lee, "What is a Collection?" Journal of the American Society for Information Science, 51(12) 2000:1106-1113. In an extensive discussion, Lee examines what he describes as the "four major presumptions associated with collections: tangibility, ownership, a user community, and an integrated retrieval system."

(16) Barbara Tillett has written exhaustively on this, see for example Barbara Tillet, "A Taxonomy of Bibliographic Relationships" LRTS, 35(2):150 – 158.

(17) Eric Jul, "Revisiting Seriality and Periodicity" Serials Review, Spring 1996:70-71.