Bibliographic control in the fifth information age: Slide notes

Gordon Dunsire, February 2021

These notes accompany the presentation of the same name.

What is bibliographic control?

"On the record", the report of the Library of Congress Working Group on the Future of Bibliographic Control published in 2008, defines bibliographic control as "the organization of library materials to facilitate discovery, management, identification, and access".

The "IFLA Library Reference Model: a conceptual model for bibliographic information" (LRM) covers "everything considered relevant to the bibliographic universe, which is the universe of discourse …".

Dictionary definitions for "discourse" emphasize written or spoken communication, and some specify a scholarly or "serious" context.

The LRM clearly intends a broader definition, by giving examples of image, map, and music resources. The LRM also restricts the definition to recorded communication: a resource is assumed to be embodied in a persistent carrier.

This presentation will therefore use a definition of bibliographic control that includes all forms of recorded human communication. The bibliographic universe is the set of all products of human discourse that forms the collective memory of Homo sapiens.

Why control and how?

The bibliographic universe requires control because the organization of human memory is necessary for social cohesion and cultural evolution. Recorded discourse is communication through time and across distances greater than the range of human senses. It carries the information that allows humans in different household groups to cooperate with each other.

The persistence and accumulation of recorded memory drives culture and its evolution. The inheritance of recorded memory is essential for cultural identity; the bibliographic universe is cultural heritage.

The management of recorded memory improves its utility and functionality.

Recorded memory is an intermediary stage in the communication of a message from one person to another. The message is transmitted and then frozen in time; the message waits to be received at some unknown time in the future by an unknown person. The focus of bibliographic management is therefore the connection between the message and the receiver: what happens after the memory is recorded.

In classical library theory, according to S.R. Ranganathan, the message, the recorded memory, the product of human discourse, is a book, and the receiver is a reader. This terminology reflects a narrower focus on written communication, but the model is readily extended to all recorded memory.

The primary factors affecting the delivery of the book to its reader, the recorded message to its recipient, are its portability, reproducibility, and findability. Portability determines if the book is taken to the reader, or the reader to the book. Reproducibility determines if the book can be accessed by more than one reader at a time. Findability determines if the book exists and how it is to be accessed by the reader. This is the realm of metadata: a book that describes other books so that readers can access their contents, data about data, the organization of recorded memory.

Information ages

The continuous evolution of human society and culture is punctuated from time to time by an innovation in communication technology that has a revolutionary impact. The innovation is followed by a significant increase in the complexity of interactions and activity across all social groups world-wide. Profound changes take place in commercial, legal, religious, and other cultural systems that affect all aspects of personal life.

Four specific innovations have had the greatest impact: Writing, printing, telecommunication, and the Internet.

Each innovation provides a fundamental change in one or more of the basic aspects of recording human memory and providing subsequent access to it.

This results in a significant change in basic cultural and social concepts and processes; a paradigm shift.

The innovation changes and continues to change everything until the next innovation. It is useful to categorize the timespan between innovations as an “age”, and specifically as an “information age”. Four innovations denote five information ages; the present is the Fifth Information Age.

First Information Age

The First Information Age is the timespan before the invention of writing. It is defined as pre-literate, and labelled “prehistoric” despite the existence of products of human discourse in the form of images and objects.

The manufacture of a painting or sculpture takes time and requires specialist skills and tools, so such products are expensive and rare. Social and cultural memory is conveyed into the future, beyond the individual memory of a person, through an oral tradition that cannot be recorded until the invention of writing.

The content of the recorded discourse is mostly representational, depicting the things of interest in the local environment. Some content is symbolic and abstract, but the context is unknown. The meaning or intention of recording the content cannot be determined; only the “art” can be appreciated in the context of modern aesthetics.

Reproduction of the recorded memory is as expensive as manufacturing the original. Each carrier of the content is a one-off, a singleton manifestation in LRM terms.

Access to the recorded discourse is very limited. Images carried by cave paintings are often located in the furthest reaches of the cave. The reader must be taken to the book to access it, and this seems to have been a religious or ritualistic activity. Portable sculptures must be small and light enough to be transported along with the other possessions of hunter-gatherer social units. Fragile carriers such as wood and soft stone are easily destroyed; small objects are easily lost.

What has survived is now curated in museum collections.

Second Information Age

The Second Information Age begins with the invention of writing, the symbolic representation of language.

Writing allows the recording of linguistic discourse. The act of speaking is readily transferred to the act of writing, and recording discourse in specific aspects of human culture becomes common-place.

Content is descriptive and much more expressive than images and objects. There is immediate benefit in recording the “word” in commercial, legal, and religious systems. Peer-to-peer communication over long distances between persons who are known to each other, the writing of letters, becomes possible.

Carriers remain singletons, but reproduction requires only the skills of the scribe. Reproduction has the same costs as the manufacture of the original manuscript, but this is less expensive than manufacturing a painting or object.

Access to recorded memory becomes easier. Writing is applied to flat surfaces, and the third dimension of the cave or figurine is not required. This allows and encourages portability by embodying the message in materials such as clay, bark, bone, and textiles. Some writing is monumental, and the reader must go to the book, but most products of discourse can be carried by hand to the reader.

Third Information Age

The Third Information Age begins with the mechanization of printing.

Printing is the mechanical reproduction of writing and images. Development of the technology begins in the Second Information Age with seals for stamping text onto clay or paper. The content is usually a name that confers ownership or authority on an accompanying manuscript. The technique evolves to cover the content of the manuscript text or drawing in a larger stamp made of wood, stone, or some other hard material that can be sculpted.

The Second Information Age ends with the development of movable type and printing presses.

Recorded discourse becomes more common-place, but it is mediated by the printer who has the skills to set the type and operate the press.

There is an immediate and significant increase in the range of persons whose memory is recorded. A greater proportion of depictive content is manufactured and distributed using the new technologies. Scholarly communication becomes industrialized with the development of printed journals.

Manufacture and reproduction of the products of discourse becomes much less expensive, and there is a corresponding increase in the quantity of such products. Reproduction becomes part of the process, and the existence of multiple identical copies becomes the norm.

Access becomes easier. The reader has a choice of copies of the book, located in multiple places, and the book is easy to transport.

Fourth Information Age

The Fourth Information Age begins with the invention of telecommunication.

Most forms of telecommunication require the message to be encoded so that it can be transmitted. The message is decoded back into its original form when it is received. Application of telecommunication technologies in discourse usually requires the discourse to be recorded as part of the encoding and decoding processes.

Encoding allows all forms of content to be transmitted, including music, speech, and static and moving images. The range and quantity of recorded discourse increases again.

Electromagnetic media become available for the persistent storage of memory. Digital encoding allows the content and carrier of the book to created, manufactured, distributed, and accessed in an integrated, seamless, and intangible infrastructure. Reproduction is unavoidable and invisible; a temporary copy of the product of discourse is automatically created in every encode/decode transaction and it is trivial to make that copy persistent.

There are no physical barriers to access, and access becomes localized; the book always goes to the reader, wherever the book and the reader may be. Transportation is instantaneous; the reader gets the book when the reader wants it.

Fifth Information Age

The Fifth Information Age begins with the invention of the Internet.

The Internet globalizes digital telecommunication networks and allows the participation of nearly every living human in discourse over a distance.

Digital encoding and decoding are a necessary process for discourse using the Internet. All discourse is recorded on persistent digital media. The deletion of recorded memory, “the right to forget”, has become a cultural and social issue, in a complete reversal of the First Information Age and “the right to remember”.

The World-Wide Web is an application of the Internet that allows any person to take on and combine the roles of author, publisher, printer, distributor, and reader. The book includes every email, social media post, Skype or webinar conversation, blog, website, or search ever made by every reader.

Reproduction is a built-in automatic feature. Overt reproductions of recorded memory are made to ensure persistence of cultural heritage, improve access, and retain evidence of discourse.

The “Internet of things” is a result of the miniaturization of computer chips as digital encoding, storage, and decoding devices. The reader and the book exist in the same local space and time. The reader is every individual human; the book is a collection of all digital human memory.

Metadata

The development of metadata for bibliographic control arises in the Third Information Age.

The quantity and availability of printed products stimulated an increase in collections of recorded memory by social groups and individuals. Such collecting began in the Second Information Age with the development of libraries of manuscripts, but these were rare because of the expense of obtaining or reproducing hand-made products. Printing allowed wealthy individuals to accumulate private collections for pleasure, research, and status, and for a greater range of commercial, legal, religious, and scholarly organizations to develop repositories of information to support their activities.

As collections grew in number and size, it became necessary to record the collector's memory of what the collection contained, and to organize access to the collection to find and select a specific product of discourse. Is the item in the collection, and if so, where is it located?

The content of metadata is essentially descriptive, and therefore linguistic in form. Textual metadata can be sorted and ordered using the syntax of the language of description, and it is much easier to formulate search and retrieval queries in the same syntax. The reader reads metadata in order to find the book.

Depictive metadata content is of limited utility. A thumbnail image is a representation or depiction of the whole image, not a description of it. Textual metadata can be transformed into spoken word, using a screenreader, or visual symbols such as colour-coded categorizations.

The Fifth Information Age allows the reader to be the author and publisher of metadata, the cataloguer, as well as the being the author and publisher of a book that is being described.

Current approaches to metadata are rooted in the paradigms of the Third and Fourth Information Ages. The impact of the Fifth Information Age on bibliographic control is at its beginning and the detail belongs to the unknown future, but it will be profound.

Identity management

The management of identity is essential to the functionality of metadata. An identifier is a label that distinguishes the referent from other things. Effective information retrieval processes require that the subject of a metadata description is identified: what individual book or associated entity is being described?

Identity management is the basis of classical authority control. The nature of discourse, and human culture itself, results in the same individual being labelled with different identifiers, and the same identifier being used for different referents. Much of this diversity is driven by local context and the difficulties of assigning identifiers that are agreed at global level.

The Fourth Information Age stimulated the development of global approaches to identifier management, generally limited to the book. Examples include the International Standard Bibliographic Number and International Standard Serial Number systems. The beginning of the Fifth Information Age saw the development of similar approaches to the identities of persons, including the author and therefore the reader.

However, the Fifth Information Age eliminates the problem of the same identifier being used for different referents. The Internationalized Resource Identifier (IRI) system, based on the Uniform Resource Identifier (URI), is applicable to anything that can be described; that is, any thing that is the subject of bibliographic metadata. This is one of the necessary and fundamental aspects of the Internet, the World-Wide Web, and the linked open data of the Semantic Web. It is managed independently of any cultural application or context.

The assignment of more than identifier to an individual thing cannot yet be eliminated. That would require the thing being described to be described as well as labelled.

In the Fifth Information Age, authority control evolves into the management of linked data identifiers.

(meta) Data provenance

The Semantic Web is a globalized metadata retrieval system built on the World-Wide Web. It is based on description logic and has no intrinsic accommodation of "truth". The Semantic Web adheres to the AAA Principle: anybody can say anything about any thing. What is said in metadata may be true or false, in the same way that the content of a product of discourse may be true or false relative to the context in which it was created.

"This statement is true" may be fake, and the author a liar. This is not just a cultural phenomenon. Discourse itself has in-built paradox, ranging from the "impossible" images of M.C. Escher to the linguistic paradox of Epimenides. "This statement is false" is false if it is true, and true if it is false.

These uncertainties mean that effective bibliographic control requires data provenance for metadata. Provenance is a means of quality control. Knowing who created metadata helps to distinguish high-quality data created by trained professionals with ethics from low-quality data created by amateurs with bias.

It is also important to know when metadata was created and what standards were used. Metadata theory and practice evolve just as much as any other form of discourse. How things were described in the past may be useless or misleading in a contemporary context.

It's an open world

The Semantic Web also makes the Open World Assumption (OWA).

The assumption is that the absence of metadata is not a description of absence, but simply a description that has not yet been made. Metadata may be added in the future, and there is no expectation that future metadata will be objectively or subjectively true. This is a consequence of the AAA principle and the paradox of discourse: there cannot be a complete description of a thing because an infinite number of false or unprovable statements can be added.

Applications based on closed-world assumptions become less efficient in the Fifth Information Age.

A bibliographic record is no longer a fixed and complete description of a book or the entities associated with it. Metadata will always accumulate, so the size of the "record" increases through time. It is unlikely that any single application will need or want to use the whole set of metadata that describes an entity.

The closed-world practice of updating erroneous or incomplete metadata is no longer tenable. Instead, it is assumed that the original metadata is "out in the field" where it is not feasible to update every copy. Revisions are made with new statements; erroneous statements are assigned appropriate data provenance.

Wikis that share data from multiple authors without central mediation have been involved in conflicts where statements are updated by one author and "updated" back to the original statement by another author. Each author wants their version to be published and the other's version to be discarded. As a result, data provenance and version control systems built-in to wiki software has become an important tool in quality control and assurance. Nothing can be truly deleted in a wiki, and amendments can be "rolled-back" to a previous version. Similar systems are required for metadata.

Imposing fees for the use of metadata in wide-area applications or for the copying of metadata to use in local applications is a barrier to the utility of metadata and discourages the reader's contribution of metadata to the global pool.

Consensus

If any reader can make any metadata statement they want, with no distinction between “fact” and “fiction”, how can any consistency be found?

In the Fifth Information Age, recorded discourse is cultural memory, and metadata is the organization of culture itself. What makes local culture consistent is local consensus. A social group agrees to a particular set of truths, reflected in recorded memory, to maintain a consistent world view.

Consensus in metadata can be determined through analysis by machine and by human being.

Statistical analysis of large sets of metadata accumulated from multiple sources can calculate consensus by matching similar statements and by using data provenance to detect bias from particular sources. This is basically how search engines work; relevance is determined by the automatic analysis of the links on a webpage, where the focus of the page is assumed to be the subject of the link, and the links to a webpage, where the page is the target of the link. The link itself is metadata; the subject and target are associated in some way.

Linked open data in the Semantic Web can be processed using semantic reasoning, a standard set of algorithms that can derive metadata statements from metadata statements. These algorithms are simple, reflecting the simple “atomic” structure of the linked data subject-predicate-object triple. They are not a substitute for human intelligence and culture.

These automated techniques are a tool for cataloguers, not a substitute for cataloguers or other humans.

Human analysis of metadata may be conscious or subconscious. The reader carries out such analysis throughout their information seeking and retrieval activity.

The conscious analysis of the relevance of data is a form of “ask the audience” in a quiz show. This is a core feature of social media in the Fifth Information Age, where the audience is invited to like or dislike a piece of data, a mini-book. Consensus is reflected in the numbers of persons who like or dislike the information and the balance between them. This is a very broad measure of data/metadata. A more refined approach is to crowd-source contributions for specific sets of books by specific sets of readers.

Subconscious analysis is now possible using eye-tracking technologies. The reader has no control of how their eyes read a book or description of a book, and it is not the linear scan that it appears to be in the conscious mind. The development of virtual reality, mimicking the immersive cultural memory of the Fifth Information Age, will stimulate the use of subconscious feedback technologies.

Conclusion

The future of bibliographic control is as unpredictable as the future of writing, printing, telecommunication, or the Internet. In every case, there has been an immediate impact on human discourse and recorded memory, followed by a slower but profound impact on every aspect of human culture. The timespan of each Information Age decreases by at least an order of (decimal) magnitude, from tens of thousands of years through a few thousand and a few hundred years to a few decades.

Bibliographic control is likely to be based on the Open World Assumption. It will involve the coordination of metadata created by professionals and amateurs with metadata created by machine analysis. Data provenance is essential to achieve this by providing context and supporting the management of quality control. Metadata is common and necessary in the Fifth Information Age, and is a social and cultural “good” that should not be controlled by commercial interests.

The purpose and function of bibliographic control is to manage cultural identity in a global framework. The distinction between data and metadata is no longer useful, and bibliographic control will become indistinguishable from culture. The Fifth Information Age is the technological extension and immersion of personal and social mind.