Reassessing Conventional Paradigms for Document Description

by

Tom Delsey

We are currently in a transitional state, adapting cataloguing conventions and rules to accommodate the description of newly emerging forms of digital resources. In many respects, we are dealing with a moving target. Digital technologies are relatively new and continue to evolve at a rapid pace. The application of digital technologies to the production and transmission of information resources is even newer than many of the technologies themselves. There is still a great deal of experimentation and innovation in the application of the new technologies, and that is not likely to diminish any time soon. It is difficult to predict what transformations we may see in either the near term or the longer term. It is becoming increasingly apparent, nonetheless, that document production and document transmission have already undergone changes of a quite fundamental nature as the result of digital technologies. In light of those changes, a reassessment of our approach to document description is inevitable. A question remains, however, as to how far-reaching a reassessment is needed.

What I would like to do in this paper is to highlight a number of what in my view are among the most significant effects to date of the application of digital technologies to document production and transmission, and to outline the implications of the resulting changes for the approach we currently take to document description.

In assessing the effects of the new technologies, we need to look both at the digital document and at the digital network.

The digital document

Digital technologies have transformed the production of documents by making possible the full integration of what we have conventionally regarded as distinct and separate media. Before the introduction of digital technologies, the choices made by the producer of a document as to the form in which its content was to be expressed and the medium in which it was to be transmitted were interdependent. Each medium—print, audio, audio-visual—had specific limitations with respect to the expression of content. Digital technologies, however, can be used to produce a single document that seamlessly integrates symbolic notation (such as text), graphics, recorded sound, and moving image, and may even incorporate added features such as a structured database and applications software. To the extent that the producer utilizes the technology's capacity for incorporating multiple forms of expression and the integration of media, the document produced will be that much more complex than the conventional analogue document, both in its structure and in its content. Digital technologies have also made documents more complex with respect to technical specifications for their transmission and use.

Secondly, digital technologies make possible both the modification of the document's form and the revision and extension of its content in ways that are considerably more transparent than is possible with conventional media. The form of the document may be subject to mediation through communications and applications software. The document's form may be determined either by choice or by a set of defaults linked to the network and/or the device used to access and "serve" the document. In other words, external factors related to the transmission of the digital document can be as significant in determining its ultimate form as factors related to its production. Likewise, digital technologies make it possible to revise and/or extend the content of the document in ways that may be much less readily apparent to the user than is the case with the conventional reissue

of a document in a revised edition or the sequential release of parts, issues, updates or supplements. It may also be more difficult to detect unauthorized alterations or false attributions within the digital document. From the user's perspective, the net result of the transparent nature of changes in either the form or the content of the digital document is that it can be considerably more difficult to distinguish different "versions" of a document than is normally the case with documents produced in conventional media.

Digital technologies also make possible new modes of interaction between the user and the document. The user may have the option of interacting with the digital document in a non-linear mode. The user may also be able to set up a personalized "view" of the document, which effectively makes that instance of the document unique. Annotation of the document may also be possible, either on an individual or group basis. In effect, then, the digital document has much greater potential to function as an "organic" entity than does the conventional analogue document.

The digital network

The application of digital technologies to communications networks has served to create a virtual space in which the description of a digital document and the document itself are equally accessible. Digital networks enable the user to move seamlessly from the description to the document described. The connection between description and document may be activated through a "hot" link in the form of a uniform resource identifier recorded in the description. Alternatively, data stored with or derived from the document itself may function as the description. The net result is a blurring of the conventional distinction between searching document surrogates as a means of accessing documents and searching documents themselves. The digital network is, however, a transitory space in which the association of a document description with the document described is more tenuous than is generally the case in an analogue setting.

The digital network also functions as a shared virtual space for document descriptions from various sources. Library catalogues, bibliographies, abstracting and indexing tools, archival finding aids, and publishers' catalogues (all of which until recently have occupied what in a figurative if not a literal sense could be called separate spaces) now share what is effectively a common space within the network. That common space also houses new tools designed to facilitate "resource discovery" that have emerged with the introduction of network technologies. In some respects, this sharing of common space has prompted a repositioning of traditional bibliographic tools. To a certain extent, there has also been a displacement of those traditional tools by the newer resource discovery mechanisms. On the other hand, network technologies provide opportunities for more effective interfacing of both traditional and newer forms of resource discovery tools. Such interfaces can be used either as a means of importing and exporting data from one source to another, or as a means of providing seamless links through which the user can move from one tool to another.

It is also important to highlight one other characteristic of digital networks, and that is their global scope and reach. The volume and diversity of materials available through digital networks is far greater than what has been and continues to be made available through conventional communications media such as publishing, sound recording and film distribution, and broadcasting. That is due in part to the fact that the technology used to produce and distribute material through digital networks is so widely available, and, relatively speaking, so inexpensive that virtually anyone can become a "publisher" or "record producer". Digital networks have also had the effect of diminishing, if not eliminating altogether, differences in level of visibility and accessibility of a product that are characteristic of conventional communications media. Material that in conventional publishing terms would be considered "grey literature" can have the same level of visibility and accessibility in the digital network as a "trade" publication would have in the

conventional print media. Of equal importance is the fact that digital networks have the potential to reach a much broader and more diverse audience than conventional communications channels. That breadth and diversity of reach has multiple dimensions: geographic, linguistic, cultural, and demographic.

Implications for document description

The transformations we are witnessing in the production and transmission of digital documents—their increased complexity of structure and content, the transparency with which they can be changed, and their organic nature—raise a number of questions with respect to the way we approach document description. Key among those questions are the following:

> What it is that we are referencing when we describe a digital document?
> How do we define the boundaries of the digital document?
> How can we most effectively reflect document relationships?
> Is there a need for additional data to support digital resource discovery and digital resource management?

The repositioning of document description within a networked environment raises another set of questions:

> How can we ensure valid and permanent links between document descriptions and the digital documents they describe?
> How can we adapt document description practices to function more effectively in a global context?
> How can we facilitate the interfacing of document descriptions derived from different sources?

The remainder of this paper will explore those questions in further detail and suggest avenues that we might pursue in adapting current descriptive conventions to the emerging digital environment.

Identification and description

With conventional analogue materials, we tend to view the document as a physical object. However, the description *per se* normally centres on those characteristics that are (or at least assumed to be) common to the larger set of objects that comprise an "edition". In effect, the document we describe is an abstraction deduced from the characteristics of a single exemplar that we assume to be representative of a set of identical (or at least similar) objects. Nonetheless, for the purposes of description the physical object and the abstraction are effectively viewed as one and the same thing. The reason we are able to view the two as equivalent is that the constraints inherent within the physical processes used to produce and transmit conventional analogue documents allow limited scope for variation between the individual objects that form the larger set we define as an "edition".

With digital documents, the exemplar on which we base the description is perceptible only through the mediation of the technologies embedded in the networks and/or devices used to transmit and "serve" the document. As noted earlier, those technologies open the possibility of alterations being made to the form and even the content of the document during the course of transmission and serving in ways that may be transparent. Likewise, digital production technologies make it possible for alterations in form and content to be made from the document's point of origin at any given time, and again in ways that may be transparent. Consequently, equating the description of the object as perceived via a specific transmission channel at a specific point in time with the descrip-

tion of the document in its abstract sense (i.e., the abstraction we might identify as an "edition") becomes problematic.

In many cases, of course, the digital document will be not unlike its analogue counterpart. It may comprise, for example, nothing more than a representation of text that from a technical perspective is straightforward, and the content may be in a form that the producer of the document treats as final and "permanent". In such a case, we effectively have the digital equivalent of the conventional analogue document, and we can apply our current model for document description without any significant re-adjustment.

At the other end of the spectrum, however, we may have a digital resource that exhibits a multiplicity of characteristics made possible through the transformative effects of digital technologies. Take for example a web resource that incorporates text, graphics, sound, moving image, a database and database software. The originator of that resource may revise, update, extend or change its content in other ways on either a scheduled or an unscheduled basis. In addition, that resource may accommodate customized user views reflecting language preferences, display modes, etc. Short of creating a "snapshot" description of the resource as viewed by a specific user at a specific point in time via a specific transmission channel, trying to apply our current model for document description to such a resource is simply not workable.

What we need is a more flexible approach that would accommodate the need to adjust, when necessary, the point of reference for the description. The conventional approach that uniformly treats the abstraction we define as an "edition" as the primary point of reference for the description may be adequate for describing digital documents that mimic their analogue counterparts. Creating a viable description for a complex, organic resource that is subject to transparent changes of form and content, on the other hand, requires a significant adjustment in our point of reference. One possible approach to describing a resource of that kind would be to shift our focus from the "edition" (or the "manifestation", as defined in the IFLA *Functional Requirements for Bibliographic Records* (FRBR)) to the "work" (again, as defined in the FRBR).[1] In practice that would entail centring the description on attributes and relationships associated with the abstraction we define as the "work" (e.g., title of the work, form of work, intended audience, context, persons and bodies responsible for the work, subject of the work, etc.). Attributes and relationships associated with the "expression" (as defined in FRBR) might be incorporated into the description as well. However, where the resource incorporates multiple or variable "expressions" of its content (such as multiple language versions) such information would be included primarily for the purpose of indicating the range or nature of the resource's content rather than for the purpose of identifying a specific "expression". Similarly, attributes associated with the "manifestation" might be included, but again primarily for purposes of indicating the range of output or display formats available rather than for purposes of identifying a specific "manifestation".

We still, however, have to address the issue of recording within the description attributes of the document that may be subject to change over time or may vary as a result of the interaction between the document *per se* and the networks and/or devices used to transmit and "serve" the document. Ultimately the description of the document does have to be cast in the form of a "snapshot". That is to say that certain key elements of the description such as the title of the document have to reflect the corresponding attribute of the document as viewed through a given channel at a given point in time. We have established certain conventions for dealing with time-related change in the attributes of analogue documents that serve the purpose reasonably well. Those conventions, however, were designed in the context of traditional modes of issuing and updating analogue documents. In adapting them to the description of digital documents, we need to carefully assess their applicability in order to ensure that they will serve the intended purpose.

Defining document boundaries

The absence of the physical boundaries that have served to delimit the extent of the conventional analogue document raises a question as to where, for the purposes of description, the digital document begins and ends. The nature of digital transmission is such that, in effect, the document is assembled at the point of "serving". A certain portion of that assembly may be initiated automatically in the course of opening the document. Subsequent stages of assembly may be triggered by user interaction (e.g., clicking on a chapter heading in a table of contents or an icon on a web page). User interaction may also activate links to content that we would normally regard as something outside the document itself (e.g., pulling up a related article by clicking on a citation in a list of references). Because the technology used in a networked environment to link content that is notionally "internal" to the document is the same as the technology used to link to "external" content, and because the linking mechanisms are largely transparent, the line between what is "internal" and what is "external" becomes blurred. Therefore, to the extent that the description of the document reflects the nature, structure and extent of its content, that blurring of boundaries is problematic. We need to establish criteria that will clarify, at least for the purposes of description, the outer boundaries of the digital document.

A similar problem arises with respect to delimiting hierarchical divisions within a document and hierarchy within a set of documents. For the purposes of description, we have traditionally made distinctions between the document as an integral entity, a document component, and an aggregation of documents. With conventional analogue documents, those distinctions are based on both the physical and the intellectual characteristics of the object or objects being described. In determining whether a set of physical objects constitute a single document we take our cues from physical aspects of the packaging as well as from intellectual elements such as the wording of titles. Similarly, in determining whether a group of physical objects constitute a multipart document or a "collection", we rely on both the physical and the intellectual characteristics of the objects in the group. Those distinctions then come into play in the conventions we use to describe the objects. Rules for describing single-part documents differ from those for describing multipart documents; rules for describing physical components of a document differ from those for describing intellectual components of a document; and rules for describing a series or set differ from those for describing a "collection". With digital documents, the physical characteristics on which we base the distinction between a single-part and multipart document are absent. Distinctions between physical and intellectual components of a document cannot be made. The distinction between a group of documents forming a series or set and a group that qualifies as a "collection" may be entirely arbitrary. We need, therefore, to reassess the relevance of those distinctions to the description of digital documents and the implications of modifying the provisions within our current rules for description that are based on those distinctions, at least insofar as their application to digital documents is concerned.

Reflecting relationships

The way in which we currently reflect relationships through document description is closely tied to what it is that we are referencing when we describe a document and how we delineate boundaries within and between documents. Inasmuch as the abstraction we define as an "edition" is our normal point of reference for the description, the relationships that are reflected through notes and added entries tend to be cast in the form of what we might call "edition-to-edition" relationships. Using the more specific terms defined in FRBR, those relationships generally fall into two broad types—"expression-to-expression" relationships and "manifestation-to-manifestation" relationships. Relationships of the first type are normally reflected through notes identifying revisions, transla-

tions, adaptations, etc., and added entries that identify the "original" on which the expression is based. Relationships of the second type are normally reflected through notes identifying alternate formats available, and reproductions or the "edition" used to produce a reproduction.

If we find, as suggested earlier, that we need to shift our point of reference when describing digital documents that are multi-layered and organic in nature, our perspective on many of those conventional relationships changes. The relationships between various language versions of a text, for example, must be viewed differently when those language versions are incorporated into a single digital resource and can be accessed simply by clicking on a button or icon designating language preference than they would be if they were produced and released through conventional media as separate editions of the text. Similarly, the transparent or organic revision or extension of content embodied in a digital document cannot be reflected in the same way as the relationship between two conventional documents, one of which contains a revised and/or extended version of the content embodied in the other. In effect, many of the relationships that have been reflected as external or "edition-to-edition" relationships between conventional documents may have to be reflected as attributes that are internal to the digital document. While that adjustment may seem straightforward, it does raise a question as to how specific we need to be in referencing those internal relationships and what kind of coordinates we might use when specificity is required.

Hierarchical relationships within and between digital documents raise an issue of a different sort. Current conventions for document description provide a number of alternatives for reflecting hierarchical relationships. The simplest form, perhaps, is the use of a series statement and series added entry. Normally that technique is used to relate descriptions for individual volumes within a series to the series as a whole (and indirectly to each other). The same technique is sometimes used to relate descriptions for individual physical components of a multipart document to the document as a whole. An alternative technique for reflecting hierarchical relationships within a series or within a multipart document (or both) is the multilevel description, in which descriptions for each of the volumes and/or parts are brought together and displayed hierarchically within a single record. As currently designed, the multilevel record technique can be applied to the description of any level of component within a hierarchical grouping, regardless of whether the component has a separate physical identity or whether it has only a discrete intellectual identity. Series statements and series added entries, on the other hand, are designed for use exclusively with components that have a separate physical identity. The corresponding techniques used to describe components that lack a separate physical identity are analytic added entries and "in analytics" or component part descriptions. However, with digital documents, as noted earlier, making the distinction between physical components and intellectual components within a document or between parts of a multipart document and documents within a multi-volume series or set is problematic. There are two questions to be addressed, then. The first is whether the techniques we use currently for reflecting hierarchical relationships can be applied to digital documents regardless of whether meaningful distinctions can be made between the physical and intellectual components of a document or set of documents. The second is whether any significant adjustments are needed to the techniques themselves for their effective application to digital documents.

Resource discovery and resource management

Document description operates within a functional context. We may define the functions of the descriptive record in the traditional terms used by early cataloguing theorists such as Cutter, or we may define them in a more structured form, as was done by the IFLA Study Group on the Functional Requirements for Bibliographic Records. Regardless of how we define the functions of the record, we nevertheless recognize both implicitly and

explicitly that they play a central role in determining the content and structure of the record as a whole as well as the form of individual data elements within the record. It is, therefore, essential to assess periodically our conventions for document description to determine whether the data we record fulfils those basic functions. An assessment of that kind is particularly necessary at this juncture because of the transformations that have occurred with the application of digital technologies to document production and transmission.

Take, for example, the traditional "finding" and "collocating" functions of the record (those that support the user task defined in the FRBR as "find"). Our interpretation of those functions has shaped conventional practice pertaining to the selection and formulation of "access points" for the descriptive record. Traditionally, our focus has centred on access points representing individuals and groups associated with the content of the document, titles associated with either the document itself or the content of the document, and the subjects reflected in the document's content. While those three broad categories of access points are as valid a centre of focus for the description of digital documents as they are for conventional documents, there are aspects of each that merit closer attention. With the shift toward the production of documents that incorporate and interweave multiple forms of expression and involve individuals and groups playing new roles, it is perhaps time to take another look at the criteria we apply to determine which of those individuals and groups will be represented by means of an access point. Similarly, given the fact that digital technologies have resulted in the virtual reinvention of the document itself, it is probably time to revisit our traditional notions of what constitutes a document title and our criteria for determining precedence among variant titles. We also need to consider the need for access points that may fall outside the three traditional categories, such as the "domain" names associated with web resources.

On another level, we need to reassess the adequacy of data that is recorded for purposes of assisting users in "selecting" resources that are appropriate to their needs. Given the user's dependency on the technologies employed to transmit and "serve" digital documents, data on technical specifications and systems requirements has become increasingly important for purposes of selecting appropriate materials.

Supporting "access" to documents also takes on a new dimension in a digital environment. The data used traditionally to identify sources from which a document might be acquired may not be the most appropriate given the new business models that have emerged in a networked environment. Perhaps even more important from a library perspective is the fact that the conventional practice of adding a call number to the document description will not serve as an adequate means of providing the user with the information needed to access a document if the document is part of the library's virtual collection. In that situation, a whole new category of information may have to be invoked in order to process the user's request against the licensing agreement or other access arrangements the library has with the document provider. While in the final analysis it may be determined that some or all of that additional information will be stored outside the descriptive record, it is nevertheless necessary to look at the access requirements in greater detail and assess the implications for the descriptive record as such.

We also need to look beyond the traditional uses of descriptive data in the context of "resource discovery" to data requirements for "resource management". Our dependency on the technologies employed to transmit and "serve" digital documents has significant implications for libraries and other organizations with respect to the management of those resources. As migration and reformatting become increasingly critical from both a service and a preservation perspective, there is an increasing need for detailed technical data pertaining to the digital document. Data pertaining to the scheduling or completion of preservation or other actions affecting the document is also becoming increasingly important from a resource management perspective. Again, it may be determined that

some or all of that additional information will be stored outside the descriptive record, but the requirements need to be analysed and the interface between what we define as the descriptive record and any other records containing data relevant to the document described needs to be fully understood.

Persistence

The conventional library catalogue has functioned, in effect, as an inventory control tool. Inasmuch as the catalogue and the collection have both been under the direct control of the library, maintaining the link between a description contained in the catalogue and the associated document contained in the collection has been a relatively straightforward matter. However, as libraries have moved toward the creation of virtual collections, incorporating into their catalogues descriptions for networked documents that reside outside their direct control, maintaining the link between a description in the catalogue and the associated document within the network has become more problematic. Given the evanescent nature of networks of digital resources, the relationship between the library catalogue and the library collection takes on a new significance.

Within a networked environment, the challenge of maintaining a valid link between a document description and the document described has two dimensions. First, there is a requirement to ensure that the reference mechanism or identifier used to link the description to the document described remains valid over time. In order to guarantee the validity of that link, it is essential that the identifier used as the linking mechanism be "persistent" (i.e., that it be kept up to date or that it be supported by a resolution service that can always direct a "look up" to the current address of the document). Ultimately responsibility for the persistence of the identifier lies with the originator or custodian of the document. The second requirement is to ensure that the description remains a valid description of the document with which it is associated. Given the potential for various attributes of the digital document's form and/or content to change over time, and the transparency with which those changes may occur, ensuring the continued validity of the description is problematic. As responsibility for the accuracy of the description *vis à vis* the document it references lies with the originator or maintainer of the description, this second dimension of persistence becomes a significant issue in the reassessment of our current approach to document description. In the final analysis, it may be necessary to rethink the basic relationship between description and document. For certain types of networked resources, we may in fact have to consider the idea of fully integrating the document description function with the document archiving function.

The global context

Developing effective resource discovery tools for networked resources also requires a renewed focus on the importance of providing context for the user. Because the range of materials is so broad and diverse, knowing the source from which the material originates has become an increasingly important factor for users endeavouring to assess the relevance of a particular document to their needs. Because of the increased availability of materials in multiple versions and a variety of formats, links between related materials are increasingly important to users. Because of the global reach of digital networks, ancillary aspects of document description such as authority files, classification schemes and subject thesauri all have to be viewed in a new light, given that the document descriptions they are aligned with are increasingly being accessed by an audience of users that is multilingual and multicultural in scope.

In practical terms, there are a number of questions that need to be addressed. We need to ask ourselves whether the kind of information we currently provide on the document's source—names of authors, publishers, sponsoring organizations, etc.—is sufficient, given

the evolving needs of users for context to assist them in evaluating the relevance of networked resources to their specific needs. Likewise, we need to ask whether the information we provide about relationships between one resource and another is sufficient, and whether the conventions we use to facilitate access to those related resources are effective in supporting "navigation" of the network. We need to ask also whether there are ways in which authority lists, subject thesauri and classification schemes could be adapted to function more effectively in a networked environment where the users cannot be assumed to share one language and one cultural context.

Interfacing with other sources of descriptive data

Finally, we need to address the issue of interfacing various sources of descriptive data within a networked environment. We are inclined to think of the document descriptions created by libraries as functioning within a closed environment. However, the fact is that there has always been a significant level of interdependence between the descriptive function fulfilled by the library catalogue and parallel functions fulfilled by other descriptive tools, particularly those that we commonly refer to as abstracting and indexing tools. The realities of library economics are such that the catalogue *per se* can provide only one level of access to the library's collection. For the most part that access centres on the document as such (i.e., the document as a monograph, a serial, or a series). Traditionally, we have relied on the producers of abstracting and indexing tools to provide analytical access to the content within those documents (i.e., to the articles within journals, the papers within the published proceedings of a conference, etc.).

Within a networked environment, those analytical tools, as noted earlier, occupy the same virtual space as the library catalogue. There are opportunities, therefore, to enhance the efficiency and effectiveness of the interfaces between the library catalogue and the analytical tools that provide an essential functional complement to the catalogue. In order to exploit those opportunities, however, it is essential that we re-evaluate descriptive practices on both sides to identify ways to improve the interface between the two. Key areas of focus in that effort would include the design and use of identifiers (both at the document and at the analytic level), and the correlation of access points (particularly those for the names of authors and terms used to identify subjects).

On a broader level, there is an emerging need to address a number of issues related to interfaces between what we have traditionally regarded as descriptive data from a library perspective and a range of document-related data that in a networked environment is becoming increasingly important from a resource management and resource use perspective. Two examples of that kind of data were noted earlier in connection with data requirements for the support of access to networked resources obtained under licence and for the management of the technical processes involved in servicing and preserving digital documents. Those examples serve to highlight the emerging importance of data associated with a broad range of functions from intellectual property management to document archiving that in the broader scheme of things are integrally related to the more circumscribed set of data that we have traditionally associated with document description.

In some respects, the key challenge we face in reassessing our current paradigms for document description lies in redefining the boundaries of document description itself, and designing effective mechanisms for interfacing the data that falls within those boundaries with data stores external to the description that support document access, management and use in a digital environment.

Notes

1. IFLA Study Group on the Functional Requirements for Bibliographic Records. *Functional requirements for bibliographic records: final report*. München: Saur, 1998.