Handbook of Technical Communication

Edited by
Alexander Mehler and Laurent Romary

In cooperation with
Dafydd Gibbon

De Gruyter Mouton

ISBN 978-3-11-018834-9
e-ISBN 978-3-11-022494-8

Library of Congress Cataloging-in-Publication Data

A CIP catalog record for this book has been applied for at the Library of Congress.

Bibliographic information published by the Deutsche Nationalbibliothek

The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at http://dnb.dnb.de.

Cover design: Martin Zech, Bremen
Typesetting: Dörlemann Satz GmbH & Co. KG, Lemförde
Printing: Hubert & Co. GmbH & Co. KG, Göttingen
♾ Printed on acid-free paper

Printed in Germany

www.degruyter.com

FSC
www.fsc.org
MIX
Papier aus verantwortungsvollen Quellen
FSC® C016439

15. Digital Curation as Communication Mediation

Christopher A. Lee

1. Introduction

Communication is a highly interactive and dynamic process. Parties to the process engage in continual efforts to revise their understanding, knowledge and assumptions.[1] Communication across space and time is often mediated through information artifacts. A journalist, for example, can communicate – through what is often called "mass communication" (Peterson 2003) – with her readers through the articles that she publishes in a newspaper. Those who read the articles can gain insights about both the subjects being discussed and the journalist herself. If the newspaper is preserved, this type of communication can occur over very long periods of time.

Humans have a long tradition of retaining information artifacts for future use. Collecting institutions – libraries, archives and museums – manage extensive collections of materials that can communicate events, insights, facts and perspectives to those who encounter them. Individuals, families, corporations, and various other types of organizations also generate and retain information artifacts across time. These activities can be seen as communication mediation.

As all sectors of contemporary societies have increasingly adopted digital technologies in order to carry out their activities, the communication mediation processes have dramatically changed. Supporting the meaningful use of digital objects over time – a set of activities that has recently come to be called "digital curation" – requires an understanding and appreciation of the various layers of representation through which meaning can be conveyed in digital systems. This chapter discusses the characteristics of those layers and strategies for ensuring perpetuation of meaning across time.

2. Digital traces as intentional and unintentional communication

Communication can be characterized as the conveyance of information from one party to one or more other parties (potentially including conveyance to him/herself). Information is "a detectable pattern on which action can be conditioned" (Cohen and Axelrod 1998: 38). Humans value information because they find it meaningful. Meaning never comes for free. Its value comes through its enactment in specific situations – and no two situations are exactly the same (Barwise and Perry 1983).

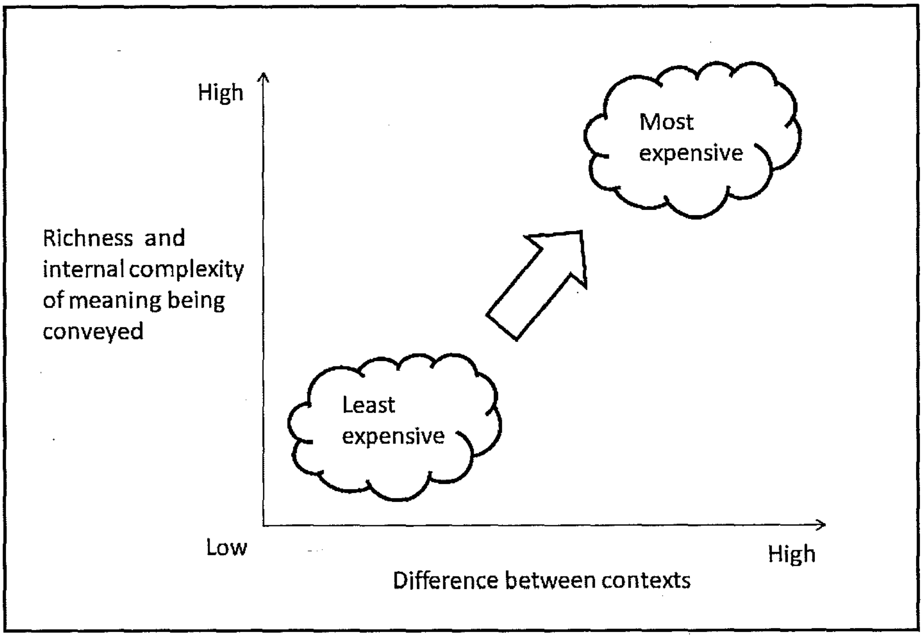
High
Richness and internal complexity of meaning being conveyed
Low
Least expensive
Most expensive
High
Difference between contexts

Figure 1. Factors influencing expense of conveying meaning across contexts.

Conveying meaning from one party to another – whether they are engaged in a real-time conversation or one is reading a document that another wrote hundreds of years ago – involves a “fusion of horizons” (Gadamer 1989). This requires some degree of “common ground” between them (Stalnaker 1978; Clark 1996).[2] For Alice to grasp a concept offered by Bob, Alice must devote effort and attention to the task (and Bob may also). For example, if Alice and Bob are sharing their views about a television show that they both saw the night before, they will exchange a variety of impressions and background assumptions – the process of sharing meaning is not complete just because they know that they both witnessed the same episode. As a more extreme example, developing a shared sense of political identity can take generations of interaction, deliberation, negotiation and shared experiences. Two factors that influence how difficult it is to convey meaning are: (1) the degree of richness and complexity of the meaning and (2) the degree of difference between the contexts in which the meaning is being enacted. Figure 1 provides a representation of these relationships.

Conveyance and enactment of meaning is not simply a matter of reliably transferring a discrete signal from one place to another. All entities involved in the process – conveyor, recipient and meaning of the information – are subject to transformation. If, for example, Jack knows a story, he can tell it to Nora. As a result, Jack, Nora and the meaning of the story may all be different from what they were before. If Nora then tells the story to Henry, a similar transformation can occur.

While there are some forms and aspects of communication that are completely ephemeral – such as unrecorded facial expressions or hand gestures – a great deal of human communication involves the generation of traces that have varying levels of persistence.[3] Human activities leave numerous traces, and all of the traces are potential conveyers of information. A door left open can signal that it is socially acceptable to enter the room. Footprints on a beach provide evidence of previous visitors and where they walked. Receipts convey information about transactions and the contexts in which they took place. Photographs of public officials shaking each other's hands indicate both that the two individuals met and that someone found the encounter to be noteworthy enough to warrant capture. The contents of a web browser cache reveal actions and behavior patterns of the browser's users. Extensive information about previous actions on a computer is also stored by the computer's operating system. The packets used to transfer data across the Internet include headers, which reflect the data's previous and intended path.

A great deal of the literature – in linguistics, communication and philosophy of language – related to processes for conveying meaning has focused on direct human discourse as the context of interaction. This has included definition of a speech act as a specific communicative event, whose meaning and appropriateness can depend significantly on the circumstances in which it takes place (Austin 1962). Speech acts that are carried out in face-to-face verbal conversation generally do not leave persistent documentary traces as evidence that they occurred. Many authors have attempted to formally represent the context of discourse in terms of the "body of information that is presumed to be shared by the participants in the conversation" (Stalnaker 1998; see also Bouquet, Serafini, and Thomason 2008). The given-and-take of direct human interaction is facilitated not only by shared knowledge and assumptions, but also by shared purposes. "Our talk exchanges do not normally consist of a succession of disconnected remarks, and would not be rational if they did. They are characteristically, to some degree at least, cooperative efforts; and each participant recognizes in them, to some extent, a common purpose or set of purposes, or at least a mutually accepted direction" (Grice 1975: 26). One can even characterize communication as a series of responses to "coordination problems" (Lewis 1969).

When conducted through fixed symbolic forms – such as electronic mail, telephone, video – communicative acts can yield persistent objects. The use of computer systems always involves the generation, copying and manipulation of strings of explicit symbolic representations; those representations are then potentially available for use in later interactions: a phenomenon that Shoshana Zuboff (1988) has called "informating." It is this "informating" of interactions that generates an unprecedented amount of explicit information; and facilitating long-term use of the information is the mission of digital curation.

One can intentionally generate traces or instead leave traces as unintended side effects of actions. Sara may put on a work uniform, for example, purposely to let others know that she is "on duty." If the uniform is looking worn and threadbare, that might also reveal that she has worked on the same job for many other days in the past, but that is not because she purposely distressed the fabric to convey such information. Likewise, composing and sending an email message is a deliberate act, but the creator is not usually making an intentional choice to also generate system log entries or temporary files that are associated with the message. The level of investment and conscious commitment to traces can fall along a spectrum from completely unreflective/incidental to highly crafted and constructed. When there are special moments in life, one often makes an effort to create much richer traces of them. For example, someone might take frequent low-resolution pictures of her everyday life using a cheap camera on her phone, but then arrange for a consciously orchestrated, well-lit and high-resolution photograph of her wedding. If the context in which the traces are accessed and used is significantly different from the context in which they were generated, then the interaction cannot rely on the same sorts of informal cues for understanding that would be available in a direct conversation.

Information, like money, is often given without the giver's knowing to just what use the recipient will want to put it. If someone to whom a transaction is mentioned gives it further consideration, he is likely to find himself wanting the answers to further questions that the speaker may not be able to identify in advance; if the appropriate specification will be likely to enable the hearer to answer a considerable variety of such questions for himself, then there is a presumption that the speaker should include it in his remark; if not, then there is no such presumption (Grice 1975).

The curation of information artifacts is fundamentally different from direct communication in that one cannot assume that the parties who generated the traces will be available to provide "answers to further questions." Instead, one must make educated guesses about likely uses of the traces and then pre-emptively respond to likely questions by embedding appropriate information in the mechanisms used to manage and provide access to the traces (e.g. repositories, collection descriptions, information packages).

There are three powerful strategies that one can adopt when his/her goal is to convey traces across contexts:

1. First, one can use fixity measures to increase the chances that the traces will persist across space and time. For example, low-acid paper will not deteriorate as quickly as paper with high-acid content.

2. Second, one can generate objects that are purposely self-describing. A book, for example, can include its own glossary of terms, and a business memorandum has elements to convey its context of creation such as to, from, subject and date.
3. Finally, one can embed traces into larger systems of traces that provide meaning to each other. For example, storing a whole series of correspondence together in the same folder will help to convey meaning that could be lost if letters were all separated and stored in different places.

I will discuss each of these strategies in Section 3 on digital curation as communication mediation.

In most of the world, human actions leave numerous digital traces. Curation of selected sets of those traces – by creators, users, information professionals and others – can enable numerous forms of inquiry, discovery and communication. Proper digital curation requires an understanding of the ways in which digital traces are both similar to and dramatically different from traces left in analog forms.

3. The nature of digital traces

Any use of digital resources is highly mediated. It involves the interaction of various hardware and software components. By definition, digital information is stored and processed as a series of discrete values. The values are represented through properties of physical media – usually through charged magnetic particles or tiny holes in disks. Hardware and low-level software detects the physical properties and interprets them as binary digits – i.e., digits that can take only one of two possible values – which are called *bits*. By convention, we say that the two possible binary values are 1 and 0 (Shannon 1948).

It is often useful to be able to identify and consistently reproduce a specific series of ones and zeroes, which is called a *bitstream*. The bitstream is a powerful abstraction layer, because it allows any two computer components to reliably exchange data, even if the underlying structure of the components is quite different (e.g. arrangement of sectors on a hard drive or pits and lands on an optical disk). In other words, even though the bits that make up the bitstream must be manifested through physical properties of computer hardware, the bitstreams are not inextricably tied to any specific physical manifestation. Another useful characteristic of bitstreams is that they can be reproduced with complete accuracy. By using well-established mechanisms – such as generation and comparison of SHA1 hash values[4] – one can verify that two different instances of a bitstream are exactly the same. This is more fundamental than simply saying that one has made a good copy. If the two hash values are identical, then the two instances are, by definition, the same bitstream.[5]

Humans are sometimes directly interested in bitstreams (e.g. when conducting detailed forensic analysis, security audits or debugging), but they are much more likely to care about the meaning that emerges from the dynamic interaction of hardware and software components that access and process numerous bitstreams.

3.1. Interacting components and interoperability

Users experience digital objects through human-computer interfaces. They handle and read documents that have been printed to paper or view and interact with representations of information on computer screens or hear it through computer speakers. In some cases, a user experiences a digital object as a relatively discrete entity, as when she reads an electronic journal article. In other cases, interaction with distinct digital objects is less direct, such as when a researcher is conducting an analysis through a system that pulls data from a variety of different sources. In the former case, the visual and/or auditory rendering of the digital object is likely to be important to the user. In the latter case, the user is likely to be more concerned with the accuracy and efficiency of access to and manipulation of the data contained in the digital objects (for more information on evaluation in technical communication see Menke 2012 in this volume).

Regardless of whether one is interacting with digital objects as distinct entities or interacting with representations that aggregate various sources, the process is highly mediated by hardware and software. In other words, any particular encounter with a digital object must occur within a specific technological environment. A specific stack of hardware and software that can be used to reproduce the digital object can be defined as one of the object's "view paths" (van Diessen 2002).[6] For example, a *Portable Document Format* (PDF) file could be read using the Adobe Acrobat Reader within the Windows 7 operating system running an Intel Core processor; or it could be read using Okular in combination with Poppler within the Ubuntu distribution of the Linux operating system running an AMD64 processor (on document formats see also Rahtz 2012 in this volume). The same hardware or software component can be shared by many potential view paths. For example, one could run the same Java application in either a Windows or Macintosh environment, as long as both systems are running a Java virtual machine.

For consumers, it is desirable to have numerous options for computer platforms that can be used to access any given digital object. If there are very few types of technical environments in which to use the digital object (i.e. a limited number of available view paths), then one can become overly dependent on the providers of those particular environments. Two major risks are (1) lock-in to specific vendors, which can require consumers to pay unreasonable rents to the vendors over time, or (2) vendors ceasing to provide support for the technology at all because they have gone out of business or changed their service/product offerings.

The flip side to dependency is interoperability. From an engineering perspective, interoperability is the ability for two or more systems or functional units to "exchange information and to use the information that has been exchanged" (IEEE 1990) or "communicate, execute programs, or transfer data ... in a manner that requires the user to have little or no knowledge of the unique characteristics of those units" (ISO 1993). Interoperability can greatly facilitate coordination and communication across systems. It is a concept that can be applied at many different levels of granularity, from the ability to connect two physical devices to the sharing of concepts and work processes across organizational boundaries (Tolk 2003). Two factors that have made interoperability increasingly important over the past several decades are (1) "distributed computing infrastructures" based on networked access to resources and (2) "increasing specialization of work, but increasing need to reuse and analyze data" (Sheth 1999) (for more information on distributed computing infrastructures see Heyer, Holz and Teresniak 2012 in this volume).

Early work on computers tended to focus on specific combinations of hardware and software that were configured to carry out tasks. However, by the 1960s, the computer science literature reflected various approaches to promote "machine independence" (Halpern 1965). This included, for example, computer-supported translation between low-level languages, in order to mitigate the effects of platform dependency (Gaines 1965; Wilson and Moss 1965). Computer scientists and engineers increasingly recognized the desirability of being able to use data in a diversity of computer environments, and they have developed numerous methods and mechanisms for doing so, including virtualization (Parmelee et al 1972), modularity (Baldwin and Clark 2000; Langlois and Robertson 1992), and standardization (Cargill 1997).

The initiative that most decisively introduced interoperability across generations of hardware into the computer industry was the development of System 360 by *International Business Machines* (IBM) in the 1960s. One of the most widely recognized innovations of the System 360 architecture was that it allowed IBM to release and support an entire family of computers – targeted at different segments of the market – that all interoperated with each other. However, System 360 also included code that could emulate hardware and thus supported software and data files created on earlier IBM hardware. Many software and hardware vendors have followed IBM's lead, building backward compatibility into their products, in order to perpetuate use of their line of products while also supporting an easy transition to their latest offerings. At the same time that producers were applying and refining the backward compatibility concept, expertise began to develop on the user side about how to address the problems associated with data residing on legacy systems within organizations.

Starting around the 1980s, there has been an increasing emphasis in the computer industry on portable and reusable code (Krueger 1992). For some types of software, developers are often willing to pay the price of increased development time or some degradation of performance in order to increase the chances that their software can be used on a wide variety of hardware platforms. A subset of system and database administrators – particularly those working for large organizations that have maintained collections of digital objects over many years – have also developed expertise in the recovery of both data and functionality from legacy systems of all sizes.

In order to address system integration, interoperability and data exchange, computer professionals frequently describe hardware and software as a stack of independent but interacting *layers* (Messerschmitt and Szyperski 2003). For example, systems for administering interactive web sites are often characterized in terms of layers (or tiers) dedicated to specific functions such as data storage, business logic, query processing, and user interface. Perhaps the most widely recognized example of such layering is the *Open Systems Interconnection* (OSI) Reference Model for network data transfer (ISO 1994). Although the OSI Model has been widely characterized as an approach that "lost" the standards war against competing ARPANET protocols, the OSI had a significant and lasting influence on how members of the computer industry characterized and conceived of networks. While the definitions of the specific OSI layers are outside the scope of this paper, the intended purpose of the stack of layers is very relevant to the discussion.

Figure 2 presents what I contend to be the implied communication dynamics of the OSI Model. Alice is sitting at a computer in Los Angeles and Bob is sitting at a computer in New York. Alice sends an email message to Bob. Because they are not sitting at the same computer, Alice cannot simply copy the file from one directory to another. Instead she must push the data down through several layers, starting at the level of the application that she is using and ending down at the level where physical movement of the data can occur. Bob's system receives the data through the layer at the bottom of his stack, and the data is then pushed up through the set of layers that ends at the application that he is using to view the email. If Alice and Bob are now both looking at the message on their screens at the same time, then we can say that they are sharing meaning based on the contents of the message. However, they are never directly moving data across the top part of the figure – i.e. the place that they care about because the meaning resides there. All interactions are mediated through a set of layers designed to support interchange across diverse and remote systems. And those applying this model will generally assume (for good reason) that users of the network do not want to see or think about data passing through any layer but the one at the very top.[7]

A result of portability of data across systems is the range of experiences that one can have when encountering the same digital object, depending upon the

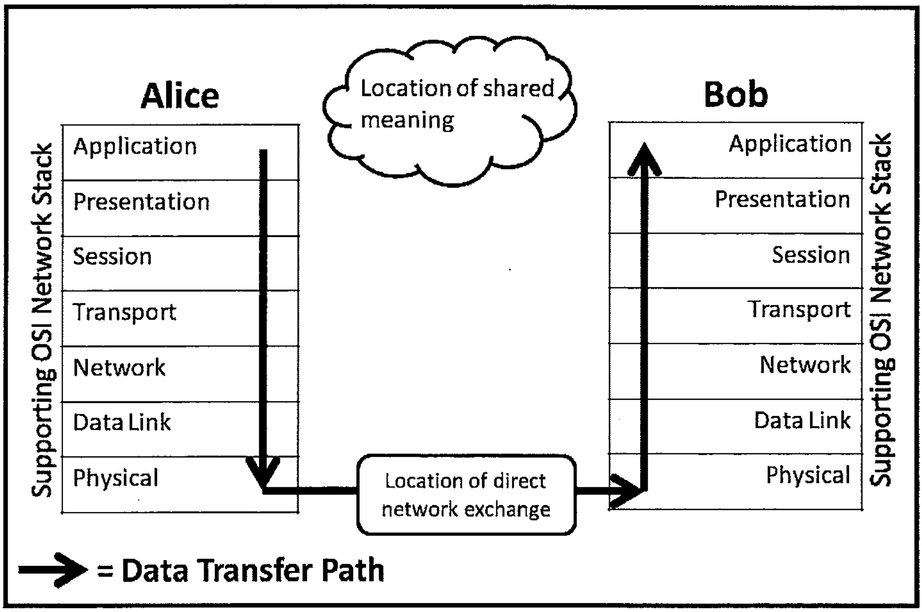

Figure 2. Implied communication dynamics of *Open Systems Interconnection* (OSI) network model.

characteristics and state of the computing environment being used to encounter it. This can be the result of fixed characteristics of the technology being used (e.g. screen size) or configuration settings. For example, when Alice and Bob view the "same" email message, they may see different fonts, lines breaking at different points, exposure of different elements of the email header, and different ways of representing embedded hyperlinks and text copied from other messages. This example demonstrates that many properties of digital objects are best understood as a range of potential values, rather than specific discrete values.

3.2. Levels of representation

Digital resources are composed of interacting components that can be considered and accessed at different levels of representation. Access to data from a storage device normally involves mounting a volume and then copying or opening files through the filesystem. There must be hardware to detect signals on the medium, hardware and software to translate the signals into bitstreams, and hardware and software to move the bitstreams into the current working computer environment. One can then interact with data as entire files or components of files. The filesystem usually plays a mediating role between the user and the underlying data, and it is designed to facilitate interaction at the file level (e.g. file naming, viewing timestamps, access controls). It serves to "hide" complicated information from the user about "where and how it stores information" (Farmer and Venema 2005). Those who are interested in the underlying data that is hidden by the filesystem can instead generate and interact with disk images, which are low-level, sector-by-sector copies of all the data that resides on the storage medium. Inspection of the disk image can reveal a significant amount of information that users of the drive did not consciously or intentionally leave there (Garfinkel and Shelat 2003). However, for most purposes, the filesystem is a very valuable abstraction mechanism, because it does not require users to understand or directly access the underlying data. Users can encounter the contents of a file as a bitstream by using a hex editor,[8] but they are more likely to render its contents within a particular application. At even higher levels of representation, one often encounters digital objects not as distinct files but as either discrete objects composed of multiple files or aggregations of such objects.

See Table 1 for a summary of several levels. One could alternatively propose a smaller set of levels that each span a larger set of properties and activities – e.g., stored data, software-readable information, and user experience – or a larger set of levels that reflect much finer distinctions – e.g., bitstreams with or without checkbits used for error correction, the work-expression-manifestation-item distinctions of the *Functional Requirements for Bibliographic Records* (FRBR), Panofsky's natural subject, conventional and intrinsic meaning levels (Panofsky 1955). However, I have defined the levels at a degree of granularity that I believe reflects specific and important implications for digital curation as communication mediation.

The properties of information at a given level of representation are directly based upon, but are not fully reducible to, properties of information at the level immediately below it. Each level has emergent properties, which convey potential meaning that is not available through any of the other layers. This is because moving between layers always involves a process of translation that both adds and removes information.

The rendering of a web page in a browser, for example, does not reflect any comments that are within the text of the HTML file. Viewing the HTML file through a text editor would reveal the comments (and other properties of the HTML markup such as whitespace and style sheet references), but it would not reveal specifically how the page is presented to users who visit the site through a browser. The developer of a web site who is trying to fix a "broken" page will routinely shift between these two levels of representation, in order to see specifically how the HTML code is expressed and how changes to the code affect the appearance and behavior of the page in one or more web browsers. A given HTML file can be rendered quite differently in two different browsers, because the software interprets the code differently.

Table 1. Levels of Representation.

Level	Label	Explanation	Interaction Examples
7	Aggregation of objects	Set of objects that form an aggregation that is meaningful encountered as an entity	Browsing the contents of an archival collection using a finding aid
6	Object or package	Object composed of multiple files, each of which could also be encountered as individual files	Viewing a web page that contains several files, including HTML, a style sheet and several images
5	In-application rendering	As rendered and encountered within a specific application	Using Microsoft Excel to view an .xls file, watching an online video by using a Flash viewer
4	File through a filesystem	Files encountered as a discrete set of items with associate paths and file names	Viewing contents of a folder using Windows Explorer, typing "ls" at the Unix command prompt to show the contents of a directory
3	File as "raw" bitstream	Bitstream encountered as a continuous series of binary values	Opening an individual file in a hex editor
2	Sub-file data structure	Discrete "chunk" of data that is part of a larger file	Extracting a tagged data element in an XML document (see Stührenberg 2012 in this volume) or value of a field in a relational database
1	Bitstream through I/O equipment	Series of 1s and 0s as accessed from the storage medium using input/output hardware and software (e.g. controllers, drivers, ports, connectors)	Mounting a hard drive and then generating a sector-by-sector image of the disk using Unix dd command
0	Bitstream on a physical medium	Physical properties of the storage medium that are interpreted as bitstreams at Level 1	Using a high-power microscope and camera to take a picture of the patterns of magnetic charges on the surface of a hard drive or pits and lands on an optical disk

Similarly, someone using a hex editor to view a disk image bitstream – a copy of the disk that reflects the contents of every storage sector rather than just the files – will be able to see many types of information that are not visible to someone who mounts the drive and views it through Windows Explorer. Examples of such information are contents of unallocated sectors, deleted files, and "hidden" system files. However, someone who only had the hex editor view of the disk image would not be able to experience the WIMP (windows, icons, mouse and pointer) interactions with folders, files and applications to which current users are so accustomed. Someone trying to fix a corrupted file or correct a bug in a viewing application is likely to shift back and forth between the hex editor representation and in-application rendering representation, in order to determine specifically how properties in the former are reflected in the latter and vice versa.

4. Digital curation as communication mediation

As described above, use of digital resources is a process that spans multiple levels of representation. This often involves a process that is similar to the one reflected in Figure 2, if one were to replace the OSI layers with the levels of representation that are elaborated in Table 1. Let us return to the example of Alice's email. When using her email client, Alice is operating at the top of the stack. She is interacting with aggregates of objects, in the form of folders and email discussion threads. If she views an individual message, she can experience it as a coherent object, but it may be composed of multiple files. When she makes any changes to the contents of her email account, the changes are "pushed down" into lower levels of representation, saving files to the filesystem and data onto storage media. If Alice sends a message to Bob, his encounter with the message will require a "pushing up" from storage (on his hard drive or a remote server) to a filesystem, application and an object or aggregation-level view.

The exchange between Alice and Bob leaves numerous traces at multiple levels of representation. Further use of those traces does not require that one uses the exact same access mechanism that Bob used. Someone with an interest in Alice's digital objects – whether that is Alice herself, a family member, coworker, auditor, journalist or scholar studying her work – could potentially interact with them at any of the levels (see Figure 3).[9] One could use or imitate an email client in order to view her email within an application environment that is similar to the one that she used. Alternatively, one could investigate email messages and attachments as individual files. Yet another approach would be to treat all of the bits on her hard drive (or the hard drive of a server that held her email) as one large bitstream and use search tools to find patterns of interest – such as email addresses and headers – within the stream.

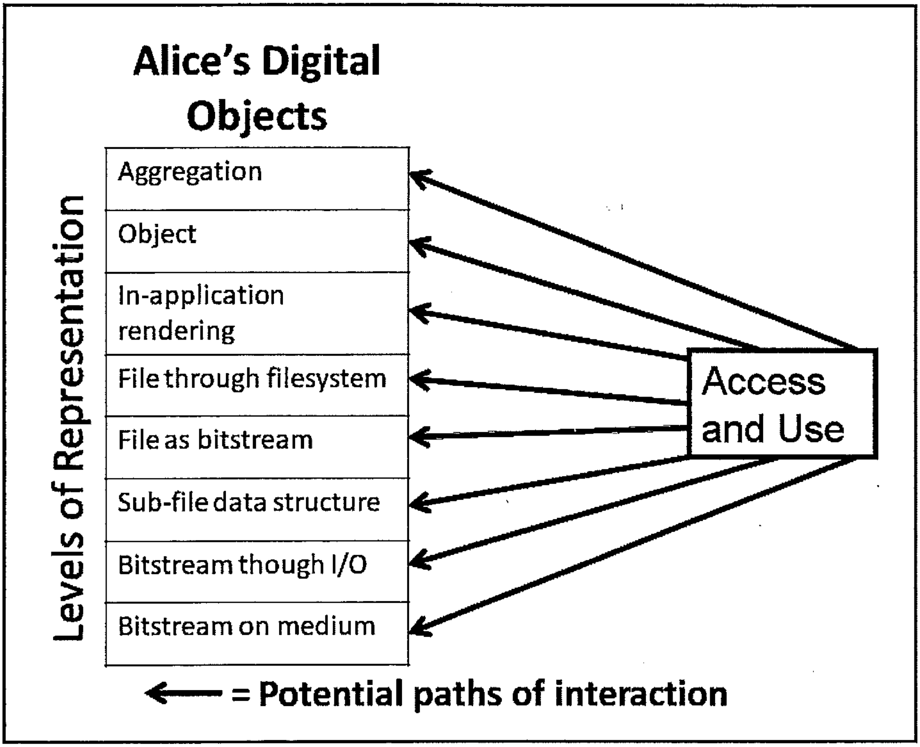

Figure 3. Deriving meaning from digital objects at multiple levels.

Digital curation is the set of activities required to ensure that digital objects can be meaningfully used over time (Lee and Tibbo 2011). Responsible care for digital traces is a highly shared and distributed endeavor. Important elements of this work are often carried out by information professionals, such as archivists, librarians, museum curators, records managers and data managers (Hedstrom and King 2006). There are also many other parties who can have an interest in and influence over the trajectory of digital traces. Those directly engaged in a communication exchange are obvious examples, because their interactions are being recorded, and they can often decide how, where and in what ways the traces of those interactions are retained. However, others can also have a major stake in how information is managed and conveyed, e.g. others who are discussed in the exchange or whose lives are significantly impacted by it (e.g. citizens of two countries whose leaders are hashing out foreign policy issues). In the following discussion, I will present several fundamental considerations for digital curation.

4.1. Reflecting context

A long-standing tenet of literature on communication is that context matters. One cannot determine the meaning or communicative effect of a message or utterance by analyzing it as a discrete entity in isolation (Dewey 1931; Rommetveit 1979). Characteristics of context are essential to the communication process. Context is inherently relational; it is always context of, about, or surrounding something, which I will call the *target entity* (TE).

There are three broad ways in which we can characterize the context of a TE (Lee 2011):

- *Context$_1$*: the set of symbolic expressions or representations that surround a TE and help one to express, make sense of, translate or otherwise act upon or within it.
- *Context$_2$*: objective or socially constructed characteristics and conditions of the situation in which a TE is, appears or occurs.
- *Context$_3$*: aspects of the mental or physical state, disposition, intentions, identity or recent experiences of an actor that bear upon how she interprets, understands, acts within, or what she notices of, the situation at hand.

Those responsible for designing and implementing information systems use symbolic representations or collections of symbolic representations (a form of context$_1$) in order to capture and maintain relevant aspects of context$_2$ and context$_3$. This process is never comprehensive or complete. There are limits to what any representation system can reflect about the environment in which it was originally embedded (Shanon 1990). "Context, in principle, is infinite. The describer selects certain layers for inclusion, and decides which of those to foreground" (Duff and Harris 2002).

The purposes and intentions of the actors associated with the resources constitute an important part of the resources' context. Identify the purposes can be essential to determining what should be done with the resources. For example, the creator of a Microsoft Word document might intend to share with future users the text that is immediately visible when opening the document in Word, but she might not intend to disclose "hidden" information embedded in the file, such as tracked changes or embedded spreadsheet data. Similarly, one might identify numerous properties that are supported by a given file format, but then realize that the creator of a set of files in that format did not purposely set any of those properties or care whether they are reproduced over time.

4.2. Capturing the state of information from dynamic environments

Digital preservation is essentially ensuring that important characteristics and values of digital objects can be consistently reproduced over time within an acceptable range of variability. As discussed above, the generation of the characteristics and values at any time requires the interaction of numerous hardware and software components. Depending on the particular set of components and settings one is using, one's experience can vary dramatically. The variability can increase over time, as both the underlying technology and user behaviors evolve. Creators and users of digital objects will value and attempt to ensure the relative fixity of some properties and values more than others. For example, Alice might care deeply about the accurate reproduction of the diacritics that appear within the text of a document that she wrote, but not particularly care about where the individual lines break, or vice versa. Bob might not care whether he is looking at the TIFF or JPG version of an image but be very concerned that they have the same persistent identifier to ensure that they are associated with the same digital object within a repository that he trusts.

There are two fundamental options for capturing and persistently reproducing information from a dynamic, distributed environment. First, one can create a snapshot of the entire state of the information at a point in time. Examples of this approach include generating a backup of a database and harvesting a snapshot of a web site. The other approach is to account for changes to information over time (i.e. capture change logs and audit trails). Examples of this approach are the revision logs available through Wikipedia and the records that are generated within a transaction processing system. Appropriate digital curation involves the identification, capture, perpetuation and reproduction of properties and values at one or more levels of representation.

4.3. Avoiding unnecessarily tight coupling to specific technologies

Digital objects are inherently dependent on technological components. However, it is possible to design, create and manage them in ways that minimize their dependence on specific technologies. As discussed above, interoperability across systems has long been a goal of computer system designers. However, there is a frequently competing goal of hardware, software and service vendors to lock consumers into their specific offerings. If Alice has created a large set of digital objects in a format that can only be read within a specific application, and that application can run on only one operating system, then Alice is very likely to purchase future releases of the application and operating system, i.e. she is locked in. If she has only one copy of the objects on her computer's hard drive or in a hosted space on the Web, she could lose the objects if a hard drive crashes. Relying completely on one hosted service provider to maintain her digital objects can also result in loss if the provider goes out of business, changes its offerings, cuts off her service, falls victim to a security breach or storage failure, or decides to delete the objects for some reason.

Digital curation is well served through "robust design" (Hargadon and Douglas 2001), which is effective in the short-term but also sufficiently flexible to remain effective in a wide range of possible future contexts. Limiting the interdependencies between subsystems can also make a design more robust against disruptions from the environment (Simon 1962). To avoid lock-in to particular combinations of hardware and software, individuals and organizations can make use of:

- redundancy (Maniatis et al. 2005);
- storing information in multiple ways (e.g., online services, formats, systems);
- diversity in technological approaches (Rosenthal et al. 2005) and business models;
- abstraction;
- virtualization (Moore 2008);
- detailed descriptive and administrative metadata beyond that which is required for immediate use and
- the development and adoption of open standards in ways that are attentive to the need for flexibility (Hanseth, Monteiro, and Hatling 1996; Egyedi 2001).

System evolution, sustainability and innovation can also be greatly facilitated through modularity (Langlois and Robertson 1992; Baldwin and Clark 2000).

4.4. Data extraction and recovery

Digital curation activities by individuals and organizations can significantly reduce the risk of information loss or corruption, but they will never entirely eliminate the need to extract or recover information from problematic hardware and software. The extraction and recovery can occur at any of the levels of representation discussed in this chapter.

Recovery of data from physical media has been a topic of discussion in the professional library and archives literature for several years (Ross and Gow 1999).There is also an active, expanding industry associated with digital forensics, which focuses on the discovery, recovery, and validation of information from computer systems that is often not immediately visible to common users. Several projects and authors have recently investigated the use of forensic tools and techniques for acquiring digital collections in libraries and archives (John 2008; Kirschenbaum, Overden and Redwine 2011; Woods, Lee and Garfinkel 2011).

4.5. Promoting discovery and sensemaking through associated metadata

Like context, metadata is a relative concept (Huc, Levoir and Nonon-Latapiem, 1997; for more information on metadata and other controlled languages see also Trippel 2012 in this volume). In relation to the body of an email message, the message header can be considered metadata. However, if one is managing each message as a distinct object, then the header might be considered part of that object's content (i.e. data rather than metadata). Similarly, one could treat the cover of a book as either metadata about the book's main content or instead part of the book's content. From the perspective of digital curation, such distinctions can be important when determining what will get retained over time and what will be exposed to users. Contemporary digital images, for example, often contain significant amounts of metadata within the headers of the files. If one were treating the files (Level 3 in Table 1) as the targets of preservation, then the header information would be preserved along with them. However, if one were treating the rendered, viewable representation of the image's content as the target of preservation, one might decide to transform the image into a new file format, which could fail to include all of the same header information.

In order for the digital objects to serve as conveyers of meaning across contexts, it can be important to generate surrogates of the objects, which can be discovered and experienced along with the digital objects themselves. For example, building an index of the contents of a set of text files will allow users to efficiently search over the content of those files. Creating thumbnails of larger images can allow users to browse through, compare and select images of interest. An archival finding aid that characterizes a whole collection of papers from an individual can provide a more holistic and richly contextualized view into the collection than one would get solely from encountering the objects on their own.

A broad category of activities by information professionals is called intellectual control (Pearce-Moses 2005). This involves mechanisms and conventions for bringing additional useful order to materials. This can involve naming conventions, authority control, and a variety of mappings across inconsistent terminologies.

4.6. Acting locally, but thinking globally

Most communication takes place within relatively well-bounded contexts. An email exchange between two individuals involves those two people and their associated expectations, background and understanding. Public broadcasts usually conform to well-established genre conventions and reach audiences that can be predicted with a fairly high degree of confidence. The exceptions – such as a private email exchange that is forwarded accidentally or "leaked" to a third party or an online story that "goes viral" and reaches an unexpectedly large and diverse audience – are noteworthy precisely because they are exceptional. However, one could argue that the reach, volume and potential persistence of digital communications makes such cases increasingly likely.

There is a more intermediate zone of interaction in which people have become quite accustomed to use of their traces beyond the original communication context. Family photos, meeting minutes and slides from conference presentations are all artifacts that serve an immediate purpose but are also often retained in order to serve further purposes in new contexts. The archives and records management literature characterize such phenomena as cases of "secondary use," while the literature on *personal information management* (PIM) investigates them as cases of re-finding and re-use. In both cases, there is an acknowledgement of the value that can be derived from perpetuating meaningful information beyond the context in which it was initially created or encountered.

No proposal for digital curation is likely to be viable if it requires individuals to devote substantial attention to secondary use of their information. People are simply too busy getting on with their lives to focus heavily on future use scenarios. However, a little bit of attention to digital curation within a local context can potentially go a long way toward more global goals. If Bob creates a collection of photos in a way that is not locked into one specific, proprietary application, this can increase the chances that both he and others will be able to make meaningful use of the photos in the future. If Alice uses simple, but consistent file naming conventions, both she and others will be able to make better sense of her files in the future. Whether one is operating in a home environment or a large bureaucracy, the questions to ask are similar: To whom might I hand off these digital traces in the future? How would that work? What are the likely motivations and needs of the recipient?

5. Conclusion

Digital traces can convey a diversity of information across space and time. They are recent additions to the repertoire of human communication mechanisms. I have argued that one of the fundamental considerations of digital traces is that they exist and can be encountered at multiple levels of representation. There is no single, canonical representation of digital resources, because the salience of a particular level of representation will vary by context. Parties responsible for digital collections should be attentive to digital traces' multiple levels of representation, making informed decisions about which levels should be encountered by future users and in what ways. Digital curation – conscious management of digital resources in order to ensure appropriate and meaningful use of the resources over time – can make the difference between accidental and ad hoc communication with the future, on the one hand, and conveyance of rich and responsible social memory, on the other.

Acknowledgements

This chapter has benefited from valuable comments and suggestions from Alexander Mehler and an anonymous reviewer.

Notes

1. For discussions of these interactions within the context of specific conversations, see Austin (1962) and Grice (1975). For a broader investigation of the relationships between the meaning of "texts" (e.g., documents, statements) and language systems, see Halliday and Matthiessen (2004).
2. As pointed out by a reviewer of this chapter, the concept of common ground is often formalized as a set of propositional assertions or attitudes. For example, A and B have common ground, because A knows something that B knows and B knows that A knows it. Digital objects cannot themselves have knowledge of propositions. However, meaningful use of a digital object requires that the user of the object has access to many of the same facts and assumptions as those who engaged in the creation of the object. Flouris and Meghini (2007) offer the concept of "Underlying Community Knowledge" as the language and theory associated with a digital object that are necessary for deriving its meaning.
3. Paul Ricoeur (2004) distinguishes between cerebral, affective and documentary traces. My use of the term is most closely aligned with, but somewhat broader than, his third category.
4. SHA-1 (NIST 2004) is one of many algorithms that can be used to generate cryptographic hashes. It can take any possible bitstream as an input and then generate a fixed-size bitstream (hash) as an output. Cryptographic hash algorithms are designed so that it is relatively easy to compute the hash of a given input but extremely difficult to 1) reconstruct the original input based on knowing the value of its hash output or 2) generate a different bitstream (including an alteration of the original bitstream) that can result in the same hash output.
5. Strictly speaking, there is always some chance that when two non-identical bitstreams (that differ by at least one of the 1 or 0 values) are run through the same hashing algorithm, they will generate the same hash value. However, if one uses a sufficiently robust hashing algorithm, the probability of this happening is so unlikely as to be effectively zero.
6. Brian Carrier (2006: 60) presents a similar idea within the context of digital forensics, which he calls an "observation tree."
7. There are many other models of communication that emphasize representation at several layers, see e.g. Strawson's three senses of meaning (1973), Levelt's "tiers" (1989) and Gumperz's "levels of signaling" (1982).

8. The hex editor itself provides a level of mediation, because it presents a *bytestream* as a set of hexadecimal values representing bytes and usually also the ASCII text values associated with the byte values. For example, the bitstream 01000001 would be seen on the left side (hexadecimal view) of a hex editor as the value "41" and on the right side (ASCII text view) of the hex editor as the value "A." Strictly speaking, humans very rarely interact with bitstreams represented as series of 1s and 0s.
9. This assumption of potential access to any level of representation is not supported by the model of "digital preservation as communication with the future" presented by Mois, Klas and Hemmje (2009), which is composed of six layers: data, representation, preservation, presentation, knowledge and content. According to Mois et al., "actual transmission" occurs only through the bottom (data) layer. The access process is assumed to always work up through the layers from the bottom to the top: "A future system is able to unpack each of the packages and gets with each of these steps the necessary information to process the information on the next higher level" (116).

References

Austin, J. L.
1962 *How To Do Things with Words*. Cambridge, MA: Harvard University Press.

Baldwin, Carliss Y. and Kim B. Clark
2000 *Design Rules. Vol. 1: The Power of Modularity*. Cambridge, MA: MIT Press.

Barwise, Jon and John Perry
1983 *Situations and Attitudes*. Cambridge, MA: MIT Press.

Bouquet, Paolo, Luciano Serafini and Richmond H. Thomason (eds.)
2008 *Perspectives on Contexts*. Stanford, CA: Center for the Study of Language and Information.

Cargill, Carl F.
1997 *Open Systems Standardization: A Business Approach*. Upper Saddle River, NJ: Prentice Hall.

Carrier, Brian D.
2006 *A Hypothesis-Based Approach to Digital Forensic Investigations*. Ph.D. dissertation, Center for Education and Research in Information Assurance and Security, Purdue University.

Clark, Herbert H.
1996 *Using Language*. Cambridge, UK: Cambridge University Press.

Cohen, Michael D. and Robert Axelrod
1998 Complexity and Adaptation in Community Information Systems: Implications for Design. In: Toru Ishida, J. G. Carbonell and J. Siekmann (eds.), *Community Computing and Support Systems: Social Interaction and Collaborative Work*, 16–42. Berlin: Springer.

Dewey, John
1931 Context and Thought. *University of California Publications in Philosophy* 12(3): 203–24.

Duff, Wendy and Verne Harris
2002 Stories and Names: Archival Description as Narrating Records and Constructing Meanings. *Archival Science* 2(3–4): 263–285.

Egyedi, Tineke

2001 Infrastructure Flexibility Created by Standardized Gateways: The Cases of XML and the ISO Container. *Knowledge, Technology & Policy* 14(3): 41–54.

Farmer, Dan and Wietse Venema

2005 *Forensic Discovery.* Upper Saddle River, NJ: Addison-Wesley.

Flouris, Giorgios and Carlo Meghini

2007 Terminology and Wish List for a Formal Theory of Preservation. In *Proceedings of PV 2007 – Ensuring Long-term Preservation and Value Adding to Scientific and Technical Data,* Weßling, Germany, October 2007.

Gadamer, Hans Georg

1989 *Truth and Method.* 2nd rev. ed. New York, NY: Crossroad.

Gaines, R. Stockton

1965 On the Translation of Machine Language Programs. *Communications of the ACM* 8(12): 736–41.

Garfinkel, Simson L. and Abhi Shelat

2003 Remembrance of Data Passed: A Study of Disk Sanitization Practices. *IEEE Security and Privacy* 1: 17–27.

Grice. H.P.

1975 Logic and Conversation. In: Peter Cole and Jerry L. Morgan (eds.), *Syntax and Semantics, vol. 3*, 41–58. New York, NY: Academic Press.

Gumperz, John Joseph

1982 *Discourse Strategies.* New York, NY: Cambridge University Press.

Halliday, M.A.K. and Christian M.I.M. Matthiessen

2004 *An Introduction to Functional Grammar*, 3rd edition. London: Arnold.

Halpern, Mark I

1965 Machine Independence: Its Technology and Economics. *Communications of the ACM* 8(12): 782–785.

Hanseth, Ole, Eric Monteiro and Morten Hatling

1996 Developing Information Infrastructure Standards: The Tension between Standardisation and Flexibility. *Science, Technology and Human Values* 21(4): 407–426.

Hargadon, Andrew B. and Yellowlees Douglas

2001 When Innovations Meet Institutions: Edison and the Design of the Electric Light. *Administrative Science Quarterly* 46(3): 476–501.

Hedstrom, Margaret and John Leslie King

2006 Epistemic Infrastructure in the Rise of the Knowledge Economy. In: Brian Kahin and Dominique Foray (eds.), *Advancing Knowledge and the Knowledge Economy*, 113–134. Cambridge, MA: MIT Press, 2006.

Heyer, Gerhard, Florian Holz, and Sven Teresniak

2012 P2P-based communication. In: Alexander Mehler, Laurent Romary and Dafydd Gibbon (Eds.): *Handbook of Technical Communication.* Berlin/Boston: De Gruyter.

Huc, Claude, Thierry Levoir and Michel Nonon-Latapie

1997 Metadata: Models and Conceptual Limits. In: *Proceedings of the Second IEEE Metadata Conference.* Silver Spring, MD: IEEE Computer Society, 1997.

IEEE
1990 *IEEE Standard Computer Dictionary: A Compilation of IEEE Standard Computer Glossaries*. New York, NY: Institute of Electrical and Electronics Engineers.

ISO
1993 ISO/IEC 2382–01, *Information Technology Vocabulary, Fundamental Terms.*

ISO
1994 ISO/IEC 7498–1, *Information Technology – Open Systems Interconnection – Basic Reference Model: The Basic Model.*

John, Jeremy Leighton
2008 Adapting Existing Technologies for Digitally Archiving Personal Lives: Digital Forensics, Ancestral Computing, and Evolutionary Perspectives and Tools. In: *iPRES 2008: The Fifth International Conference on Preservation of Digital Objects, London, UK, September 29–30.*

Kirschenbaum, Matthew G., Richard Ovenden and Gabriela Redwine
2010 *Digital Forensics and Born-Digital Content in Cultural Heritage Collections*. Washington, DC: Council on Library and Information Resources.

Krueger, Charles W.
1992 Software Reuse. *ACM Computing Surveys* 24(2): 131–183.

Langlois, Richard N. and Paul L. Robertson
1992 Networks and Innovation in a Modular System: Lessons from the Microcomputer and Stereo Component Industries. *Research Policy* 21(4): 297–313.

Lee, Christopher A.
2011 A Framework for Contextual Information in Digital Collections. *Journal of Documentation* 67(1): 95–143.

Lee, Christopher A. and Helen R. Tibbo
2011 Where's the Archivist in Digital Curation? Exploring the Possibilities through a Matrix of Knowledge and Skills. *Archivaria* 72: 123–68.

Levelt, Willem J.M.
1989 *Speaking: From Intention to Articulation.* Cambridge, MA: MIT Press.

Lewis, David. K.
1969 *Convention: A Philosophical Study.* Cambridge, MA: Harvard University Press.

Maniatis, Petros and Mema Roussopoulos, T.J. Giuli, David S. H. Rosenthal, and Mary Baker
2005 The LOCKSS Peer-to-Peer Digital Preservation System. *ACM Transactions on Computer Systems* 23(1): 2–50.

Menke, Peter
2012 Evaluation of technical communication. In: Alexander Mehler, Laurent Romary and Dafydd Gibbon (eds.): *Handbook of Technical Communication.* Berlin/Boston: De Gruyter.

Messerschmitt, David G. and Clemens Szyperski
2003 *Software Ecosystem: Understanding an Indispensable Technology and Industry.* Cambridge, MA: MIT Press.

Mois, Martin, Claus-Peter Klas, and Matthias L. Hemmje
2009 Digital Preservation as Communication with the Future. In: *DSP'09 Proceedings of the 16th International Conference on Digital Signal Processing*, 112–119. Piscataway, NJ: IEEE Press.

Moore, Reagan
2008 Towards a Theory of Digital Preservation. *International Journal of Digital Curation* 3(1), 63–75.

NIST
2004 *Secure Hash Standard*, Federal Information Processing Standards Publication 180–182.

Parmelee, R. P., T.I. Peterson, C. C. Tillman and D.J. Hatfield
1972 Virtual Storage and Virtual Machine Concepts. *IBM Systems Journal* 2: 99–130.

Panofsky, Erwin
1955 *Meaning in the Visual Arts: Papers in and on Art History.* Garden City, NY: Doubleday.

Pearce-Moses, Richard
2005 Intellectual Control. In: *A Glossary of Archival and Records Terminology.* Chicago, IL: Society of American Archivists.

Peterson, Mark Allen
2003 Anthropology and Mass Communication: Media and Myth in the New Millennium. Oxford: Berghahn.

Rahtz, Sebastian
2012 Representation of Documents in Technical Communication. In: Alexander Mehler, Laurent Romary and Dafydd Gibbon (eds.), *Handbook of Technical Communication.* Berlin/Boston: De Gruyter.

Ricœur, Paul
2004 *Memory, History, Forgetting.* Chicago, IL: University of Chicago Press.

Rommetveit, R.
1979 Words, Contexts and Verbal Message Transmission. In: R. Rommetveit and R. M. Blakar (eds.), *Studies of Language, Thought and Verbal Communication*, 13–26. London: Academic Press.

Rosenthal, David S. H., Thomas Robertson, Tom Lipkis, Vicky Reich, and Seth Morabito
2005 Requirements for Digital Preservation Systems: A Bottom-up Approach. *D-Lib Magazine* 11(11). Online available at http://www.dlib.org/dlib/november05/rosenthal/11rosenthal.html, retrieved on May 31, 2011.

Ross, Seamus and Ann Gow
1999 *Digital Archaeology: Rescuing Neglected and Damaged Data Resources.* London: British Library.

Shannon, C.E.
1948 A Mathematical Theory of Communication, *Bell System Technical Journal* 27: 379–423, 623–656.

Shanon, Benny
1990 What Is Context? *Journal for the Theory of Social Behaviour* 20(2): 157–166.

Sheth, Amit P.
1999 Changing Focus on Interoperability in Information Systems: From System, Syntax, Structure to Semantics. In: Michael Goodchild, Max Egenhofer, Robin Fegeas and Cliff Kottman (eds.), *Interoperating Geographic Information Systems*, 5–29. Boston, MA: Kluwer Academic Publishers.

Simon, Herbert A.
1962 The Architecture of Complexity. *Proceedings of the American Philosophical Society* 106: 467–482.

Stalnaker, Robert C.
1978 Assertion. *Syntax and Semantics 9*: 315–322.

Strawson, P.F.
1973 Austin and 'Locutionary Meaning.' In: *Essays on J. L. Austin*, 46–68. Oxford: Oxford University Press.

Stührenberg, Maik
2012 Foundations of markup languages. In: Alexander Mehler, Laurent Romary and Dafydd Gibbon (eds.), *Handbook of Technical Communication*. Berlin/Boston: De Gruyter.

Tolk, Andreas
2003 Beyond Technical Interoperability: Introducing a Reference Model for Measures of Merit for Coalition Interoperability. *Paper presented at the 8th International Command and Control Research and Technology Symposium, Washington, DC.*

Trippel, Thorsten
2012 Controlled language structures in technical communication. In: Alexander Mehler, Laurent Romary and Dafydd Gibbon (Eds.): *Handbook of Technical Communication*. Berlin/Boston: De Gruyter.

van Diessen, Raymond J.
2002 *Preservation Requirements in a Deposit System*. Amsterdam: IBM Netherlands.

Wilson, Donald W. and David J. Moss
1965 CAT: A 7090–3600 Computer-Aided Translation. *Communications of the ACM 8(12)*: 777–781.

Woods, Kam, Christopher A. Lee, and Simson Garfinkel
2011 Extending Digital Repository Architectures to Support Disk Image Preservation and Access. In: *Proceedings of the 2011 Joint Conference on Digital Libraries*, 57–66. New York: Association for Computing Machinery.

Zuboff, Shoshana
1988 *In the Age of the Smart Machine*. New York, NY: Basic Books.