Chapter 10
Data, Metadata, and Ted

Christine L. Borgman

10.1 Introduction

My conversations with Ted Nelson began in earnest in 2004 when we shared an office at the Oxford Internet Institute (OII). He was working on Xanadu, and I was working on *Scholarship in the Digital Age: Information, Infrastructure, and the Internet* [7]. My work was in conversation with Ted's since I was a graduate student, having read *Computer Lib* early on. Ted signed my copy of *Literary Machines* [25] at a talk in the mid-1990s, thus I was in awe of the man when Bill Dutton put us together as visiting scholars in the OII attic, a wonderful space overlooking the Ashmolean Museum.

Ted and I arrived at concepts of data and metadata from very different paths. He brought his schooling in the theater and literary theory to the pioneer days of personal computing. I brought my schooling in mathematics, information retrieval, documentation, libraries, and communication to the study of scholarship. While Ted was sketching personal computers to revolutionize written communication [24], I was learning how to pry data out of card catalogs and move them into the first generation of online catalogs [6]. Our discussions that began 30 years later revealed the interaction of these threads, which have since converged.

C.L. Borgman (✉)
Department of Information Studies, University of California, Los Angeles, CA, USA
e-mail: borgman@gseis.ucla.edu

© The Author(s) 2015

D.R. Dechow, D.C. Struppa (eds.), *Intertwingled*, History of Computing,
DOI 10.1007/978-3-319-16925-5_10

10.2 Collecting and Organizing Data

Ted overwhelms himself in data, hence he needs metadata to manage his collections. He drapes himself in data collection devices (Fig. 10.1). On any given day, he carries some combination of paper notebooks, a packet of colored marker pens draped on a string over his shoulder, a video camera, still camera, audio recorder, and other recording devices.

Ted's data immersion is not simply about recording one's life experiences, as in Gordon Bell's MyLifeBits project [5]. Rather, Ted's data collection encompasses information relevant to documentation, writing, networks, and hypertext – anything that could possibly inform the design of Xanadu and related technologies. The common thread of the data collection projects of Ted Nelson and Gordon Bell is that both acquire heterogeneous data types that must be integrated. Bell, a distinguished computer scientist at Microsoft, has the resources to build a testbed for studying and exploiting those data (Gemmell et al. [15]). Ted, for whom necessity is the mother of invention, takes a much more informal approach to capturing, describing, and integrating the content he gathers. One of our first conversations was about metadata – he asked me to explain it, and as I started to do so, he asked me to stop and wait a moment. He pulled an audiocassette recorder from his jacket pocket, turned it on, said "Christine Borgman on metadata." Then he turned to me and said, "now talk about metadata" … and we did! At the end of that conversation, he made an entry in his daily diary about the conversation and where it was located on which cassette. Thus, Ted created a document (the recording), assigned a subject heading ("metadata") and a personal name entry ("Christine Borgman") as metadata about

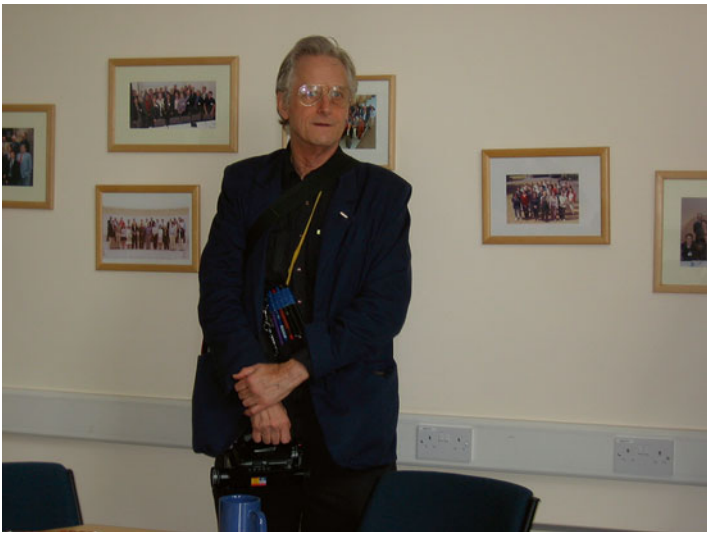

Fig. 10.1 Ted Nelson, 2005, carrying data collection devices at the Oxford Internet Institute (Photo by Christine L. Borgman)

the document, and created a catalog record (the entry in his notebook). In this case his action was recursive, as he created a metadata record about metadata.

10.2.1 Theoretical Traditions

Formally, metadata is "structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource" [23]. The NISO definition breaks metadata into the three general categories of descriptive, structural, and administrative. Other definitions of metadata make finer distinctions among types [2, 17].

Ted developed a fundamental understanding of data, metadata, and documentation through his work on hypertext and literary machines, despite his lack of familiarity with the field of information studies. He recognized that documents do not stand alone, even if they look like independent objects. Rather, they are deeply connected to many other objects. These relationships can be abstract, as in the influence of one text on the meaning of another – known as "intertextuality" in semiotics and literary studies. Relationships also can be explicit, when one document cites another, includes portions of other documents ("transclusions"), or makes any other direct link. These explicit relationships are the basis for hypertext and hypermedia, terms coined by Ted in the 1960s. The body of relationships among documents is sometimes known as "hypertextuality."

In documentation, usually dated to the Belgian, Paul Otlet, in the early twentieth century, texts are deconstructed into component parts and linked together. In the information sciences, Otlet's work is considered to be the precursor to hypertext [29–31]. Building upon the complex history of bibliography, documentation, identity, and philosophy of information, modern cataloging rules link together nodes of documents, authors, publishers, and other entities as a network [35]. The model known as FRBR, for Functional Requirements for Bibliographic Records, establishes four levels of entities: work, expression, manifestation, and item [36]. The *work* is the distinct intellectual creation, such as Shakespeare's play *King Lear*. The *expression* is the specific form, such as the text of the play as published in Shakespeare's First Folio. The *manifestation* is a physical embodiment of an expression, such as the Royal Shakespeare Company's 2007 production of King Lear in Stratford-upon-Avon starring Ian McKellen. The *item* is a single exemplar and a concrete entity, such as a specific copy of the program for a performance of that 2007 production. FRBR also establishes relationships among persons, corporate bodies, concepts, objects, events, and places.

10.2.2 Practical Consequences

Metadata, such as the familiar entities in a catalog record—author, title, publisher, date, place, physical description, subject, and classification—are essential descriptions of documents and other entities. Without metadata, a library would be no more than rooms full of books and documents shorn of their title pages. Metadata describes, enables access, and provides links to other documents. Some forms of metadata creation can be automated, such as extracting keywords and citations from a text, and others are created by human experts, such as descriptions of the intellectual content and history of an object.

Having stumbled upon the concept of metadata in our conversations, Ted was an eager student of knowledge organization. I introduced him to Ann O'Brien of the Department of Information Science at Loughborough University, one of Britain's experts on knowledge organization [20, 37]. Dr. O'Brien specialized in multi-media documentation, a particular challenge for Xanadu. While she was at first daunted by Ted's style of inquiry (Fig. 10.2), they quickly became able sparring partners. Ted, Ann, and I explored many aspects of metadata that might be applied in Xanadu.

Among the challenges that Ted encountered, long known to Ann and other experts in knowledge organization, is that the apparatus necessary to represent relationships between documents can be very large. Data, including texts, can be the tip of the iceberg. The metadata required to manage, to find, and to follow relationships amongst documents is often much more voluminous than the documents themselves. Furthermore, as networks grow in size, they become more complex, requiring other layers of representation and more sophisticated tools for navigation. Ted's concept of hypertext supports multi-directional links between documents (Fig. 10.3). His approach is aligned with semiotics, philosophy, and information science thinking about relationships between works [14]. However, multi-directional links are complex to implement computationally, which was especially true in the early days of personal computing. Technical compromises made in the early days of the World Wide Web undermined Ted's ability to implement hypertext on a large scale. He

Fig. 10.2 Ted Nelson and Ann O'Brien, Oxford, 2006 (Photo by Christine L. Borgman)

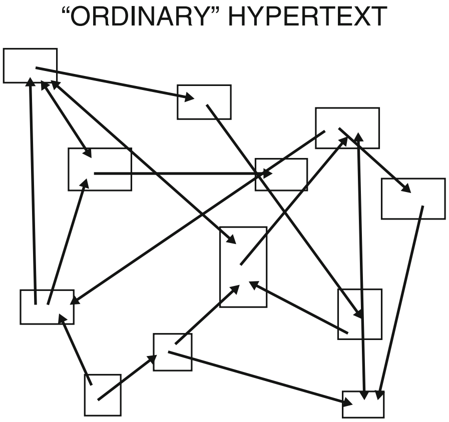

Fig. 10.3 Ordinary hypertext, with multi-directional links. From *Literary Machines* (Used with permission)

continues to rail at this constraint. Forty years after Computer Lib, computers are far more sophisticated and the networks among digital objects are much richer and more complex. It is time to revisit fundamental assumptions of networked computing, such as the directionality of links, a point made by multiple speakers at the symposium—Wendy Hall, Jaron Lanier, Steve Wozniak, and Rob Akcsyn amongst them.[1]

10.2.3 Managing Research Data

Managing research data is similarly a problem of defining and maintaining relationships amongst multi-media objects. Research data do not stand alone. They are complex objects that can be understood only in relation to their context, which often includes software, protocols, documentation, and other entities scattered over time and space [8]. The need to model these complex relationships stimulated technical research in persistence, identity, and linking of research objects [4, 26, 28, 38]. These approaches build upon—and are limited by—the technical capabilities of the World Wide Web.

As research data become valued as objects to be maintained, reused, and repurposed, many stakeholders are coming together to address questions of linking, identity, and stewardship. These concerns cross boundaries of scholarly communication, computer science, publishing, research funding, libraries, archives, data repositories, and education [8, 9, 13, 34]. Breakthroughs on these data problems may contribute to understanding hypertextuality, and vice versa.

[1] See in this volume Wendy Hall, Chap. 11: *Making Links: Everything Really is Deeply Intertwingled* and Rob Akcsyn, Chap. 15: *The Future of Transclusion*.

10.3 Provenance and Pluralism

Provenance, another fancy word that was unfamiliar to Ted but basic to his ideas, has meanings both narrower and broader than *metadata.* The term was borrowed from French in the eighteenth century to indicate the origin or source of something. It can mean simply the fact of the origin or the history of something and the documentation of that record. In the narrower sense, provenance can be a type of metadata that describes the origin of an object. Provenance on the World Wide Web includes aspects such as the attribution of an object, who takes responsibility for it, its origin, processes applied to the object over time, and version control [16, 21]. The ability to establish the provenance of a dataset, for example, may influence whether a result is deemed trustworthy, is reproducible, is admissible as evidence, or to whom credit is assigned [10, 22].

Provenance is particularly difficult in hypertext because it requires not only establishing authoritative links between objects, but also sustaining those links and information about the links over long periods of time. These links remain reliable only if the identity of the object can be established uniquely at the item level [1, 32, 33]. Unique and persistent identifiers need an institutional home, whether an International Standard Book Number, which is maintained by national libraries [19]; a Digital Object Identifier (DOI), which is maintained by the DOI Foundation and stored in interconnected registries ("Digital Object Identifier System" [11]); an Open Researcher and Contributor Identifier (ORCID) for author names, which is maintained by a non-profit foundation and stored in interconnected registries [18]; or domain-specific identifiers, such as those for genomics, chemistry, and so on. Lighter weight solutions, such as Linked Open Data, can be used to establish rich sets of relationships among objects, but these are not intended for long-term stability [3, 27]. In scholarship and in research data, stable linking is essential to follow chains of evidence. The apparatus to establish and to maintain those links cannot exist in a vacuum. Rather, it is part of a larger knowledge infrastructure, one that is now being imagined anew [8, 12].

Ted's notion of "pluralism" is that "anyone may revise anything – harmlessly" ([25], 2/61). Pluralism expresses today's notion of use and reuse of digital objects. The social movement toward open access is predicated on the ability to borrow and reuse content, with attribution to the original source. Authors and other creators are more willing to share their works openly if they can expect credit for that work. Both credit and harmlessness thus depend on provenance. The original object must stay intact and later references to those originals must be sustained.

10.4 Conclusion

Ted has tackled—head on—some of the thorniest known problems of information organization. He lacked the background in the information sciences to know how hard these problems were. Yet hard problems often are solved by those who approach unaware of the littered path of failure. Ted brought fresh ideas to knowledge organization and stimulated those inside the field to revisit fundamental premises. The challenges that have stymied Ted are those that frustrated many who came before. Ted, like Paul Otlet, tried to develop a pure new system that did not depend on the technologies and bureaucracies of the day. Reinventing infrastructure is even harder than reinventing literature, and he has tried to do both. Ted has a large following in the library world because he dared to reimagine the library. Everything is indeed intertwingled, another provocative term of Ted's invention. Xanadu, the hypertext system, is related to Samuel Taylor Coleridge's 1797 poem about the summer palace of Kublai Khan, is related to the Yuan dynasty, is related to the ruins of Shangdu in Inner Mongolia, is related to … the many other paths of inquiry to be pursued in the ideal world of comprehensively networked knowledge.

Open Access This chapter is distributed under the terms of the Creative Commons Attribution Noncommercial License, which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

References

1. Agosti M, Ferr N (2007) A formal model of annotations of digital content. ACM Trans Inf Syst 26(1). doi:10.1145/1292591.1292594
2. Baca M (1998) Introduction to metadata: pathways to digital information. Getty Information Institute, Los Angeles
3. Bechhofer S, Buchan I, De Roure D, Missier P, Ainsworth J, Bhagat J, Goble C et al (2013) Why linked data is not enough for scientists. Futur Gener Comput Syst 29(2). Special section: Recent advances in e-Science: 599–611. doi:10.1016/j.future.2011.08.004
4. Bechhofer S, De Roure D, Gamble M, Goble C, Buchan I (2010) Research objects: towards exchange and reuse of digital knowledge. Nat Proc. doi:10.1038/npre.2010.4626.1
5. Bell G (2001) A personal digital store. Commun ACM 44(1):86–91. doi:10.1145/357489.357513
6. Borgman CL (1977) Library automation at Dallas Public Library. In: Shepherd CA (ed) Information management in the 1980's: proceedings of the ASIS annual meeting, Chicago, vol 40. Knowledge Industry Publications for American Society for Information Science, White Plains, p 29 (2–A9–A–14 Microfilm)
7. Borgman CL (2007) Scholarship in the digital age: information, infrastructure, and the internet. MIT Press, Cambridge
8. Borgman CL (2015) Big data, little data, no data: scholarship in the networked world. MIT Press, Cambridge
9. Bourne PE, Clark T, Dale R, de Waard A, Hovy EH, Shotton D (eds) (2011) Force 11 manifesto: improving future research communication and e-Scholarship. Retrieved from http://www.force11.org/white_paper
10. Buneman P, Khanna S, Tan WC (2001) Why and where: a characterization of data provenance. Lect Notes Comput Sci 1973:316–330
11. Digital Object Identifier System (2009) Retrieved from http://www.doi.org
12. Edwards PN, Jackson SJ, Chalmers MK, Bowker GC, Borgman CL, Ribes D, Calvert S et al (2013) Knowledge infrastructures: intellectual frameworks and research challenges. University of Michigan, Ann Arbor. Retrieved from http://deepblue.lib.umich.edu/handle/2027.42/97552
13. Force11 (2015) Home page. Force11: the future of research communications and scholarship. https://www.force11.org/about

14. Furner J (2010) Philosophy and information studies. Annu Rev Inf Sci Technol 44(1):159–200. doi:10.1002/aris.2010.1440440111
15. Gemmell J, Gordon B, Lueder R (2006) MyLifeBits: personal database for everything. Commun ACM 89:88–95. doi:10.1145/1107458.1107460
16. Gil Y, Cheney J, Groth P, Hartig O, Miles S, Moreau L, Pinheiro da Silva P (2010) Provenance XG Final Report. W3C Incubator Group. http://www.w3.org/2005/Incubator/prov/XGR-prov-20101214/
17. Greenberg J, White HC, Carrier S, Scherle R (2009) A metadata best practice for a scientific data respository. J Libr Metadata 9(3/4):194–212
18. Haak LL, Baker D, Ginther DK, Gordon GJ, Probus MA, Kannankutty N, Weinberg BA (2012) Standards and infrastructure for innovation data exchange. Science 338(6104):196–197. doi:10.1126/science.1221840
19. International Standard Book Number (ISBN) Agency (2013) Home page. http://www.isbn.org
20. Ma Y, O'Brien A, Clegg W (2007) Digital library education: some international course structure comparisons. Joint Conf Digit Libr 490. doi:10.1145/1255175.1255289
21. Moreau L (2010) The foundations for provenance on the web. Found Trends Web Sci 2(2/3):99–241. doi:10.1561/1800000010
22. Moreau L, Groth P, Miles S, Vazquez-Salceda J, Ibbotson J, Sheng J, Varga L et al (2008) The provenance of electronic data. Commun ACM 51(4):52–58. doi:10.1145/1330311.1330323
23. National Information Standards Organization (2004) Understanding metadata. NISO Press, Bethesda
24. Nelson TH (1974) Computer lib: you can and must understand computers now/dream machines. Hugo's Book Service, Chicago
25. Nelson TH (1994) Literary machines, 93rd edn. Mindful Press, Swarthmore
26. Object Reuse and Exchange (2014) http://www.openarchives.org/ore/
27. Parsons MA, Fox PA (2013) Is data publication the right metaphor? Data Sci J 12:WDS32–WDS46. doi:10.2481/dsj.WDS-042
28. Pepe A, Mayernik M, Borgman CL, Van de Sompel H (2010) From artifacts to aggregations: modeling scientific life cycles on the semantic web. J Am Soc Inf Sci Technol 61(3):567–582. doi:10.1002/asi.21263
29. Rayward WB (1991) The case of Paul Otlet, pioneer of information science, internationalist, visionary: reflections on biography. J Librariansh Inf Sci 23:135–145
30. Rayward WB (1994) Visions of Xanadu—Paul Otlet (1868–1944) and hypertext. J Am Soc Inf Sci 45:235–250
31. Rayward WB, Buckland MK (1992) Paul Otlet and the prehistory of hypertext. Proc ASIS Annu Meet 29:324–324
32. Renear AH, Dubin D (2003) Towards identity conditions for digital documents. In: Proceedings of the 2003 international conference on Dublin core and metadata applications: supporting communities of discourse and practice. Dublin Core Metadata Initiative, Seattle, WA
33. Renear AH, Palmer CL (2009) Strategic reading, ontologies, and the future of scientific publishing. Science 325:828–832. doi:10.1126/science.1157784
34. Research Data Alliance (2015) Home page. https://rd-alliance.org/node
35. Svenonius E (2000) The intellectual foundation of information organization. MIT Press, Cambridge
36. Tillett BB (2004) What is FRBR?: a conceptual model for the bibliographic universe. http://www.loc.gov/cds/FRBR.html
37. Tinker AJ, Pollitt AS, O'Brien A (1999) The Dewey decimal classification and the transition from physical to electronic knowledge organisation. Knowl Org 26(2):80–96
38. Van de Sompel H, Sanderson R, Klein M, Nelson ML, Haslhofer B, Warner S, Lagoze C (2012) A perspective on resource synchronization. D-Lib Mag 18(9/10):1–6. doi:10.1045/september2012-vandesompel