

Provided by the author(s) and University College Dublin Library in accordance with publisher policies. Please cite the published version when available.

Title	Dilemmas in information science (IS) and information retrieval (IR): recurring challenges or new solutions?
Author(s)	Thornley, Clare V.
Publication date	2009
Publication information	Aslib Proceedings, 61 (3): 323-330
Publisher	Emerald
Item record/more information	http://hdl.handle.net/10197/3605
Publisher's version (DOI)	http://dx.doi.org/10.1108/00012530910959853

Downloaded 2018-05-27T23:26:33Z

The UCD community has made this article openly available. Please share how this access benefits you. Your story matters! (@ucd_oa) 

Some rights reserved. For more information, please see the item record link above.



Dilemmas in information science (IS) and information retrieval (IR): recurring challenges or new solutions?

Clare Thornley

School of Information and Library Studies, University College Dublin, Ireland

clare.thornley@ucd.ie

Abstract

Purpose - This paper analyses the extent to which understanding IS and IR as disciplines characterised by intractable dilemmas is a useful conceptual framework through reviewing and re-evaluating an important contribution to the field (Neill, 1987, 1992) in light of more recent developments.

Design/methodology/approach - This paper reviews the discussion of central dilemmas within information science (IS) and information retrieval (IR), through literature review and conceptual analysis. It assesses the extent to which they remain intractable problems or whether improved solutions have been developed and discusses the implications of these ongoing challenges. The main problem addressed is, in Neill's (1987, 1992) terminology "the dilemma of the subjective in information organisation and retrieval" which is understood as the problem of how the meaning of documents can be represented to meet the needs of the user.

Findings - Many of the dilemmas discussed within IS and IR remain fairly intractable primarily because information and meaning have both subjective and objective qualities which often have a complex relationship. Recent technological developments have, however, altered the nature of some of these dilemmas and also created some new dilemmas for the subject.

Research implications/limitations - Historical perspectives within IR and IS should be used when discussing theoretical and technological developments in the subject. The conceptual framework of dilemmas remains a useful theoretical tool for IS and IR in terms of examining the nature of problems in research and practice.

Originality/value - This paper re-visits an important theme in IS and IR and provides an updated perspective on some central issues.

Keywords Information theory, History of information science, Conceptual dilemmas, Information retrieval

Paper type Conceptual paper

Motivations

The motivation for this paper came from the planning and delivering of an updated course on "information retrieval and organisation" at the School of Information and Library Studies, University College Dublin. This course is taught to a wide range of undergraduate students and a resource was needed which would encapsulate some of the central issues of the subject to a non-specialist audience. The author re-discovered some of the work of Neill from the late 1980s and early 1990s and found his dilemma framework a useful way to introduce some of the important problems in IR and IS. It is fascinating to consider the extent to which these dilemmas remain and the ways in which they evolve as technology and our use of it changes. This paper then is both an appeal for a wider recognition of Neill's contribution and an exploration of how these dilemmas present themselves in modern IS and IR.

Structure of paper

The paper begins with a review of Neill's (1987, 1992) thesis that the central problems of IS and IR can usefully be interpreted as dilemmas. It then proceeds to discuss a selection of these dilemmas and assess how and why they have developed over the last two decades. The aim is to show that IS and IR are indeed characterised by persistent and intractable dilemmas. New technologies have alleviated some of the conflict within these dilemmas, however, in most cases the central nature of the problem remains resistant to solution and, in some cases, new dilemmas have arisen. Thus in terms of research and practice in IS and IR we need to remain alert that complete solutions to our problems will remain elusive and that progress will normally come from the balancing and re-adjusting of competing forces and priorities.

Introduction

Firstly the claim that IS and IR are, in fact, disciplines characterised by dilemmas will be assessed and reviewed. Neill (1992) begins his book with the claim:

It is also my contention that information use problems are, in certain instances, insoluble dilemmas. At least I depict them as insoluble. I do so because they are information problems grounded in human nature and can be solved only by altering that nature (Neill, 1992, p. xii).

His central contention is that information in documents, as it is represented knowledge, will always have an uneasy relationship with the kind of information we use in the immediate present:

The history of the development of intellectual access to the store of knowledge is the history of the tension between the fluid uniqueness of the individual enquirer and the essential stability and concreteness of the store of knowledge itself (Neill, 1992, p. 19).

Some research contemporaries of Neill had ideas which were also concerned with dilemmas and conflicts within the representation of information. Buckland (1991, p. 1) analyses the concept of information into three different types: information as process which is the act of informing; information as knowledge which is what the information is about; information as thing which is data, documents, objects, etc. He identifies as a central problem the fact that information as process and knowledge are subjective and intangible yet, within an information system, they must be expressed as an objective entity, i.e., information as thing. According to Buckland an individual's knowledge is a subjective mental state which, in order to be communicated, has to be represented in some physical way (information as thing). Thus information has a number of different, conflicting yet related properties and the difficulties of how these properties interact is one of the central problems of information science.

This problem of contradiction and conflict is not only seen in discussion of the implementation of techniques and methods of meaning representations but also in debate regarding the theoretical framework and development of information science. Saracevic, over a decade after Neill's paper on dilemmas, in his review of information science (1999) discusses the difficulties and lack of communication and coherence between different traditions within IS:

Scientist-poets wanted: I see the field of library and information science (L&IS) as highly centrifugal and greatly in need of high quality synthesisers. Library and information science has always been easy to enter by persons trained in other disciplines, particularly if they bring quantitative skills. The pattern has been many fresh starts by new entrants rather than strong accumulation. Nor is there full agreement as to which work is paradigmatic (Saracevic, 1999, p. 1052).

So how does the claim that research and practice in information is characterised by dilemmas stand up to developments over the last couple of decades?

Meaning representation

The problem of representing the meaning of documents for effective retrieval by users is one of the central concerns of IS and particularly IR. The difficulties identified by Neill concern the differences between the static nature of stored information and the fluid and changeable needs of the enquirer. Buckland (1991, p. 53) also highlights that representation, by its very nature, is “always incomplete in some regard” and that “representation is (probably) a description that was accurate at one time” (p. 59). Thus one dilemma arises from the lack of completeness of representation and another one from the gap in time between the stored document and the user’s current needs. So firstly a document has been reduced or summarised in some way to facilitate both effective storage and retrieval, yet paradoxically, this very reduction may result in the loss of information which some people may have found useful. Secondly, once a document has been represented and stored it, in one sense, inhabits a static past yet a user will come to it with new and perhaps unpredictable (at the time of representation and storage) information needs. Buckland (1991, p. 59) emphasises that “the historical nature of information is a central feature of information services, even though it is not always recognised as such”.

Is this lack of completeness still a problem or dilemma for IS and IR? I argue that it does remain a dilemma because, regardless of technological advances, a representation of a document will normally result in a simplification of its meaning. A document, in itself, is also just a representation of an event or the thoughts of an author so it is also only a partial perspective. The limits of words in expressing content are not only a problem for IS and IR but are also considered in literature. In his poem *In Memoriam* (1850, section 5), an elegy for his friend, Tennyson acknowledges his disquiet about writing the poem:

I sometimes hold it half a sin
To put in words the grief I feel;
For words, like Nature, half reveal
And half conceal the soul within.

Meaning is by its very nature not identical to that which it represents and, when this is further condensed or reduced within an information retrieval system, this disjunction often becomes more pronounced. Has research in information, however, made any great progress in the last 20 years in at least mitigating the difficulties of this dilemma? The next three sections discuss particular issues in the problem of meaning representation and analyses progress in resolving these dilemmas.

The partial nature of meaning representation

This dilemma is not possible to resolve totally as a more exhaustive approach to meaning representation which recognises ‘all the concepts and notions included in the document’ (Salton and McGill, 1983, p. 55) will generally improve recall (the number of the relevant documents in the collection which are retrieved). It will also, however, reduce precision (the percentage of retrieved documents which are in fact relevant) as it will increase the number of non-relevant documents which are retrieved. A ‘perfect’ IR system would retrieve every single relevant document in the collection and no irrelevant ones i.e. all and only all the relevant documents in the collection. In practice as more relevant documents are retrieved the number of irrelevant documents retrieved also rises. This has improved in IR but it has not been totally resolved. Thus representing the meanings of documents, whether through human indexing or automated techniques, will normally involve some kind of trade off between the requirements of representation (representing the meaning of the document) and discrimination (discriminating the document from other documents which do not share its meaning). It is still very difficult to represent a document in a way which will guarantee it would be retrieved if it was relevant to a user and also guarantee that it wouldn’t be retrieved if it wasn’t relevant.

The historical nature of meaning representation

This “gap in time” between a document stored, its representation and the user has been changed by recent technology though it may not always result in more effective or accurate retrieval. Perhaps the most well known example of this is Wikipedia (www.wikipedia.com) which allows any user to update or edit content of its web encyclopaedia. This should reduce the historical nature of IR because the content is being constantly updated and corrected (or at least edited) on a continuous basis. Thus although there will still be some kind of historical gap between the content of the web page and the exact time of the user’s query this should be considerably reduced compared to a more traditional IR system.

Does this solve the dilemma of meaning representation, at least in terms of its historical distance from the user, or does it just provide new ones to consider? It does mean that mistakes and inaccuracies in documents can be corrected rather than remain indefinitely; however, it also means that new mistakes can be added to documents that were previously correct. It also means that we lose a record of how people once perceived an entry in the encyclopaedia (e.g., an incident or historical figure) with which we can compare current perspectives as the entries are a continuous present rather than a “snapshot” of the past. Thus the ability to update documentation does reduce to an extent the dilemma of the gap between the historical nature of documentation and the present needs of a particular user. It also, however, presents new dilemmas in terms of currency versus accuracy (as anyone can update entries) and it reveals that, in some cases, the historical nature of documentation is itself a provider of information in terms of how people at a certain time perceived a certain event.

The nature of communication between IR system and user

The historical gap between the information stored and the present needs of the user is one factor in the communication difficulties between an IR system and a user. Another related factor is identified by Buckland as “indirectness”, i.e., the difference in context (e.g., time, space, background knowledge) between the user and the context

in which the documentation was originally represented and stored. Thus there is a “lack of direct link between the source of the message and its recipient” (Buckland, 1991, p. 61). This can make it difficult for the user to think of the correct terms to use when searching for documents as they do not have access to the document’s social and temporal context which may have provided them with clues about appropriate search terms. This problem is also identified by Blair (1990, 2006). In his view there are many different ways to describe a subject and many different ways to represent the meaning of a document, in everyday conversation context often gives us clues to help identify the “where the other person is coming from” in how they describe things. In IR this is rarely the case and thus there is a tension between the user’s interpretation of the meaning of documents and the authors and indexers, described by Neill (1992, p. 14) as the “confrontation between the individual and the text”.

Has this dilemma or “confrontation” between the user’s perception of the text and the indexer or system’s perception of the text changed or reduced as a result of new technologies. Social bookmarking sites such “del.icio.us” (<http://del.icio.us>) allow individual users to tag (pick terms that they think best describe documents) as they wish and these can be structured in a more personalised information management system. In one sense this means users are not constrained by cataloguing or indexing systems which may not suit their own interpretation of the meaning of documents so it could be said to alleviate this dilemma. It doesn’t, however, remove the historical problem as users may forget over time how they classified certain documents and their interests and priorities also may change making some tags irrelevant. Web 2.0 technologies (O’Reilly, 2005) in general also allow access to other user’s searching and tagging behaviour to allow sharing of “effective” search strategies. Collaborative filtering, as used, for example, by Amazon (<http://www.amazon.com/>) allows IR systems to pool user information in ways that may assist the user (and Amazon of course) by suggesting books which users with similar buying histories also bought.

These technologies then both facilitate increased individualism or subjectivism in terms of personalised tagging of documents which reduces the extent to which a particular user is bound by the indexer’s interpretation of a text and they also reduce the isolation and subjective perspective of a particular user by allowing the sharing of search information between multiple users. The latter facility raises new dilemmas for information privacy in terms of the tension between effective searching and privacy. The sharing of search strategies requires the storing of individual’s search histories which could be seen as (at least a potential) invasion of privacy. Google (www.google.com) have been criticised for this (see, for example, <http://www.google-watch.org/bigbro.html>) and have, as a defence, highlighted their duty to provide effective searching technology for their users. This kind of collaborative tagging and rating of documents can also be argued to reduce the extent to which readers take personal responsibility for assessing the quality of documents and also make it less likely that unusual or “blue sky” documents will get read. Thus it could be argued that it brings the pros and cons of the peer review process, until recently confined to academic publishing, to more widely available documents. The process of assessing and reviewing documents, either by experts or others, is itself a source of dilemmas. Who is really qualified to judge, what factors apart from academic merit may influence them and how much weight should we give to their opinion? Extending this process to the wider public does not necessarily remove or mitigate these dilemmas. A recent experiment by the “Nature” journal (Jackson, 2006; *Nature*, 2006) in which

any web user could comment on papers submitted for publication had a very limited uptake and contributions (partly because they had to be made public) were found to be of little value in actually assessing the papers, suggesting that user comments may not always be as pertinent as we might hope or assume.

Technological developments over the last 20 years have clearly changed the relationship between the user and indexer (or IR system) and reduced some of the “indirectness” identified by Buckland and Neill as a source of tension between an individual and a text both in terms of allowing more personalised methods of meaning representation (through tagging) and in reducing the isolation of the user by providing some kind of social context through the sharing of other user’s search strategies and histories. They have also, however, raised new dilemmas about the effectiveness of one person’s tagging system (such as whether it will still make sense in a few years and how can they share it with others who may have different interpretations) and the tension between the desire to harness collective search information to enhance the effectiveness of IR systems, whilst maintaining a critical distance about the reliability of their judgements as well as a desire to keep an individual’s search strategies and history a strictly private matter.

Research, evaluation and development in IS and IR

The dilemma of how to test the effectiveness of IR systems has been a long running problem within IR and IS since the early tests comparing different indexing techniques done in the 1960s at Cranfield college (Cleverdon *et al.*, 1966). The crux of the problem is how to measure (which implies some kind of objective and quantitative procedure) the effectiveness of an IR system in retrieving relevant documents and discarding irrelevant ones when the criteria for judgement is the concept of relevance which is, at least in some ways, a subjective experience of a particular user at a particular time of a particular document. Neill (1992, p. 141) outlines the nature of the problem.

The dilemma in the area of research methods in IS is that information creation and use is complex, imprecise and subjective. Quantitative measurement just cannot address the major problems which are human-human, human-machine and human-language.

In IR testing and evaluation there are broadly two main research or methodological traditions (Ellis, 1996; Ingwersen and Jarvelin, 2005): the quantitative tradition tests (Sparck Jones, 2000), based on large scale test collections such as the annual TREC experiments (Text Retrieval Conference <http://trec.nist.gov/>); and the qualitative tradition, which focuses more on small scale user studies. In the quantitative approach, relevance judgements, which determine which documents are relevant to which queries, are defined in advance within the IR system and these judgements are then used to evaluate how effective particular IR systems are at retrieval. Within the qualitative tradition there is much more of an emphasis on how individuals perceive and use the information, including many factors “outside the text” which may influence relevance such as work task, etc. (Borlund, 2000). There is then a conflict or a dilemma between the desire to provide reliable results based on large-scale quantitative measures and the desire to provide data on the use of IR systems which accurately reflect the experience of users in “real life” contexts. This has been a difficult problem for IR to resolve. Ford (1999, p. 1151), for example, claims that the

quantitative approach produces highly reliable answers to highly meaningless questions and the qualitative approach produces highly meaningful questions with highly unreliable answers. The gap or division between data collection and a meaningful theoretical framework has been hard to close. Ellis (1992) argues that the physical (or quantitative) tradition has data but no paradigm and that the cognitive (or qualitative) tradition has a paradigm but no data. There is then a history within IR or a conflict or dilemma between these two different research and evaluation traditions, but has much progress been made in more recent times to resolve this problem?

Essentially the dilemma has not been effectively resolved; TREC has expanded to cover some newer aspects of IR systems such as Blogs (see recent information on new research themes at <http://trec.nist.gov/tracks.html>) and numerous user studies have been carried out. The complex problem of using relevance as a measurement tool has, however, remained. Sparck Jones, whose career spanned some decades of IR testing observed in 2005 (p. 581), when reflecting on the continued difficulties of the field, that “IR testing was like marching into a bog with no boots on”. There is also limited evidence of improved synergy between the two research traditions. Ingwersen and Jarvelin (2005, p. 108) argue that improvements in our understanding of user behaviour have not necessarily translated into better IR systems:

While the understanding of the task effects on information seeking has advanced, the understanding of how to derive and apply design criteria for IR systems has not advanced correspondingly.

Conclusions

A number of different dilemmas in IR and IS have been discussed and reviewed in terms of new developments to assess whether any progress has been made towards a resolution. In the cases discussed, technology has decreased the severity of some of the dilemmas but then also normally raised new dilemmas (such as Google’s improved performance but the new dilemmas of privacy of search records). As such the author would argue that IR and IS remain disciplines characterised by intractable dilemmas, even if the exact nature or manifestation of these dilemmas may have evolved since Neill’s original work on this problem. In the author’s view (Thornley, 2005, 2007) this perspective on understanding IR and IS as a discipline characterised by dilemmas remains, and indeed is perhaps even more pertinent, than it was at Neill’s original time of writing. It provides a way of articulating problems in the field without necessarily having to argue for a particular solution or position which, in the case of information, will almost always result in over-simplification and dogmatic assertions. Many of these dilemmas remain perennial and intractable as befits their close relationship with the nature of meaning and information, both of which depend upon the complex relationship between the objective world and our subjective experience of it. New levels of understanding and new technologies will undoubtedly change our experience of how these dilemmas operate in IR and IS but they are very unlikely to make them go away. This presents an ongoing and fascinating challenge for those researching and working in information.

References

- Blair, D.C. (1990), *Language and Representation in Information Retrieval*, Elsevier Science Publishers, Amsterdam.
- Blair, D.C (2006), *Wittgenstein, Language and Information: 'Back to the Rough Ground!'*, Springer, Dordrecht.
- Borlund, P. (2000), "Experimental components for the evaluation of interactive information retrieval systems" *Journal of Documentation*, Vol. 56 No. 1, pp. 71-90.
- Buckland, M.K (1991), *Information and Information systems*, Greenwood, New York.
- Cleverdon, C.W., Milles, J. and Keen, E.M. (1966), *Factors Determining the Performance of Indexing Systems, Vol. I, Design; Vol. II, Test Results*, Aslib, Cranfield Project, College of Aeronautics, Cranfield.
- Ellis, D. (1992), "Domain, approach and paradigm in IR research", *Journal of Documentation*, Vol. 48 No. 3, pp. 328-331. [Letter to editor]
- Ellis, D. (1996), *Progress and Problems in Information Retrieval*, Library Association Publishing, London.
- Ford, N. (1999), "The growth of understanding in information science: towards a developmental model", *Journal of the American Society for Information Science*, Vol. 50 No.12, pp. A1141-1152.SI.
- Ingwersen, P. and Jarvelin, K. (2005), *The Turn: Integration of Information Seeking*
- Jackson, J. (2006), "Collaborative filtering no peer review", *Government Computer News*, Tech Blog, 22 December 2006, available at: <http://www.gcn.com/blogs/tech/42839.html> (accessed 10 June 2008).
- Nature* (2006), "Peer review and fraud", Editorial, Vol. 444 No. 7122, pp. 971-972, available at: <http://www.nature.com/nature/journal/v444/n7122/full/444971b.html> (accessed 10 June 2008).
- Neill, S.D. (1987), "The dilemma of the subjective in information organisation and retrieval", *Journal of Documentation*, Vol. 43 No. 3 pp. 193-211.
- Neill, S.D. (1992), *Dilemmas in the Study of Information: Exploring the Boundaries of Information Science*, Greenwood Press, Westport, Connecticut.
- O'Reilly, (2005), "What Is Web 2.0? Design Patterns and Business Models for the Next Generation of Software", available at: www.oreilly.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html (accessed 10 June 2008).
- Salton, G. & McGill, M.G. 1983. *Introduction to Modern Information Retrieval*, McGraw-Hill, USA.
- Saracevic, T. (1999), "Information science", *Journal of the American Society for Information Science*, Vol. 50 No.12, pp.1051-1063.
- Sparck Jones, K (2005) "Some thoughts on classification for retrieval" *Journal of Documentation*, Vol. 61 No. 5, pp. 571-581.
- Sparck Jones, K. (2000), "Further reflections on TREC", *Information Processing and Management*, Vol. 36 No. 1, pp. 37-85.
- Thornley, C. (2005), *A dialectical model of information retrieval: exploring a contradiction in terms*, PhD thesis, University of Strathclyde, Glasgow. Available at: http://www.cis.strath.ac.uk/~ir/research_students/digital%20library/clarethornleyphd.pdf (accessed 20 February 2009)
- Thornley, C. and Gibb, F. (2007), "A dialectical approach to information retrieval", *Journal of Documentation*, Vol. 63 No. 5, pp. 755-764.

Received
Revised
Accepted