Information-Centred Research for Large-Scale Analyses of New Information Sources[1]

Mike Thelwall
School of Computing and Information Technology, University of Wolverhampton, Wulfruna Street, Wolverhampton WV1 1SB, UK. E-mail: m.thelwall@wlv.ac.uk
Tel: +44 1902 321470 Fax: +44 1902 321478
Paul Wouters
Virtual Knowledge Studio for the Humanities and Social Sciences, Royal Netherlands Academy of Arts and Sciences, Amsterdam, The Netherlands. Email: paul.wouters@vks.knaw.nl
Jenny Fry
Department of Information Science, Loughborough University, Loughborough, UK. Email: J.Fry@lboro.ac.uk

New mass publishing genres, such as blogs and personal home pages provide a rich source of social data that is yet to be fully exploited by the social sciences and humanities. We claim that information-centred research (ICR) not only provides a genuinely new and useful information science research model for this type of data, but can also contribute to the emerging e-research infrastructure. Nevertheless, ICR should not be conducted on a purely abstract level, but should relate to potentially relevant problems.

Introduction

Information-centred research (ICR) is an e-research methodology that focuses on a new information source by 1) developing generic research tools that can be applied across a number of problem areas and 2) identifying relevant research problems (Thelwall & Wouters, 2005). ICR is deliberately open-ended in the knowledge domains that the relevant problems may derive from. ICR stemmed from two issues:

a. The proliferation of different online information sources in the sense of large numbers of documents that could potentially be categorised in various meaningful ways (e.g., "Spanish university webs", "teenagers' MySpaces", "all blogs"). This is not the e-science "data deluge" (Hey & Trefethen, 2003) but is more like an e-research "document deluge".
b. The difficulty in understanding the potential research relevance of new online information sources because of their often informal and innovative nature. Thus there are many problems for which the new information sources seem *a priori* to be relevant, but for which they are later found to be inappropriate.

ICR researchers may contribute directly to knowledge in the form of publications or reports, or may attempt to deliver the information and associated processing techniques to appropriate knowledge domain experts to use for collaborative or solo research.

In contrast to ICR, the more standard problem-centred research (here: PCR) approach is to investigate whether an information source would aid in a specific research problem. For example, blogs might be analysed to see whether blog discussion volume was a good indicator of public interest in political issues, news stories, or new book releases. Each of these three could form one traditional (PCR)

[1] Thelwall, M., Wouters, P., & Fry, J. (2008). Information-Centred Research for large-scale analysis of new information sources, Journal of the American Society for Information Science and Technology, 59(9), 1523-1527.

investigation. In contrast, the ICR approach would be to investigate blogs to see what kinds of topics were discussed in them and then to direct blogs to those research issues that they could usefully address. More concretely, ICR can result in two different kinds of scholarly output: an ICR article or a PCR article. The ICR article would be an exploratory analysis of a new information source authored by the ICR researchers and not containing domain-specific research hypotheses or detailed theoretical frameworks. One ICR article, for example, explains how blogs can be a useful information source for any social sciences research touching on public opinion and describes a simple investigative method (Thelwall, 2007). This article does not contain research hypotheses or theoretical frameworks, however – at least one of which would normally be expected from social sciences research. In contrast, a PCR article derived from initial ICR research would contain domain-specific research hypotheses and would be authored by domain experts on their own, by domain experts in collaboration with ICR researchers, or by ICR researchers incorporating appropriate domain expertise. The PCR article would hence be a "normal" research article *triggered by* an ICR investigation that had identified the relevance of the information source to the domain.

ICR is not only more open-ended than PCR but has a different goal: directing information to appropriate problems rather than solving a given problem with the information. Claims 1 and 2 below were made in a previous article (Thelwall & Wouters, 2005). Here we expand the social sciences research claims in 2 to include cultural research that might be described as humanities-oriented, and introduce claims 3 and 4.

1. ICR is more effective than PCR at identifying useful applications of new information sources, especially if conducted by information scientists.
2. Teams of information scientists should provide ICR hubs to assess new information sources for potential use in social sciences research.
3. ICR is not covered by existing information science theories.
4. ICR should not be conducted on a purely abstract level, separated completely from potentially relevant problems.

Finally, we discuss the future prospects for ICR in order to arm information scientists for seeking funding. The goal is not to create a new discipline from fields with a shared interest, as was achieved with communication sciences (Paisley, 1984), but for *some* information scientists to consciously provide an ICR service to the social sciences. The extent to which 1 and 2 above are desirable varies for different stakeholders in the system, and hence the likelihood of them being accepted will be contingent on incentive systems, economic factors and institutional structures.

ICR and information science research theories

ICR is not covered by existing information science theories. The three most prominent information science theories are domain analysis, ASK (Anomalous States of Knowledge) and cognitive theories. These are all problem or "use" based, and do not provide generic tools for analyzing new information sources across problems.

Domain analysis (Hjørland & Albrechtsen, 1995) is useful to position ICR theoretically in respect to other Library and Information Science (LIS) activities. This theory posits that information should be best understood and analysed through the users, not as individuals, but as part of specialist knowledge domains. In particular, Hjørland (2002) proposes 11 specific competencies of information specialists that can apply to specific knowledge domains (e.g., indexing domain-specific document collections). In contrast, ICR is a cross-domain activity because the information

source analysed could be relevant to many different domains, and hence is not a form of domain analysis. A second point of differentiation is that ICR is not proposed as a core activity for general information scientists but as a specialist activity. It is not unique in this because a large part of some other LIS fields, such as information retrieval and bibliometrics, is general purpose rather than domain specific. For example, there are many bibliometricians in national research evaluation centres around the world.

One of the most influential information theories is *ASK*, which tries to move away from users explicitly formulating information needs. Instead, it focuses on the users' problem statements, from which representations can be built of underlying anomalies in knowledge that could be resolved by information, for example by an information retrieval system delivering appropriate documents (Belkin, 1980). There are many other *cognitive theories*, including those taking into account the context rather than just the individual's cognitive state (Ingwersen & Järvelin, 2005). More widely, there are many theories of information seeking and behaviour that engage holistically with information-related behaviour (e.g., Bates, 1989; e.g., Case, 2002; Foster & Ford, 2003). ICR is not covered by cognitive theories, however, because the end recipients of the information are not seeking a solution to a specific problem, instead ICR researchers pro-actively offer them an information source to explore a problem that is relevant to them. Hence the end user of ICR is not an actor in the process of solving the 'problem' of the new information source.

ICR could be seen as *information filtering* (Belkin & Croft, 1992), since the goal is to get information sources to appropriate users, except that information filtering systems typically deal with similar kinds of documents, routing them individually to appropriate recipients. Information science *channels of information seeking* research (e.g., Spink & Cole, 2001) has similarities with ICR in the sense that ICR is channelling data to potential users. Channelling research is the opposite of ICR, however, in the sense that it investigates the channels that users select to get the information that they need (e.g., formal vs. informal) whereas ICR provides a new channel (i.e., a research methodology) through which end users (i.e., non-ICR researchers) can be given access to data and methods relevant to their research field.

ICR is similar to *data mining*, which focuses on extracting previously unknown patterns from databases, except that it is knowledge-domain independent, deals with unstructured collections of documents, and identifies the types of patterns that can be extracted rather than the actual patterns. The data mining literature does not seem to have developed ICR-relevant theories, however, supplying instead prescriptive procedural models as aids to practice (e.g., Hand, Mannila, & Smyth, 2001, p. 11-15; Pyle, 1999, p. 29).

The previous LIS research style most closely related to ICR is *literature-based discovery*, which develops algorithms to extract research hypotheses or connections from literature databases (e.g., Swanson & Smalheiser, 1997; Swanson, Smalheiser, & Bookstein, 2001). This is radically different from most information science research in that there is no necessary involvement of end users: in literature-based discovery, the research hypotheses are generated from the investigation and are its end product rather than the starting point (i.e., the opposite of standard problem-centred research). This is similar to ICR in the sense that the end user is not pre-determined, but would presumably be a scientist working in a knowledge domain that could assess the hypothesis suggested by the literature connections discovered by the system. ICR is broader in focus than literature-based discovery, however. ICR delivers data sources rather than individual facts or specific hypotheses, although general hypotheses may

well arise out of the ICR data exploration. More importantly, ICR is proposed as a generic methodology for approaching emerging new information sources, rather than a methodology for more effectively employing of a set of databases.

Finally, note that ICR has some parallels with Paisley's (1984) idea of "variable fields" that cut across other fields by focussing on a particular issue. ICR, although not proposed as a field in its own right, can cut through a range of fields for which Internet-derived data is potentially useful.

Problem abstraction in ICR

Problem abstraction is a key theoretical issue for the validity of ICR as a research methodology. *How can the potential of new information sources as research objects be revealed if they are abstracted from the kind of use context brought to bear by problem-centred research*? In reality a "pure problem" unrelated to any information source does not exist, neither does "pure information" exist prior to interpretation. The act of distinguishing data is already loaded with interpretative frameworks and implicit assumptions. ICR cannot, therefore, be built upon the idea of interpretation-free information. Rather, ICR emphasizes that the act of interpretation should occur at a higher level of abstraction, which takes the set of conceivable research questions from a variety of fields into account.

For PCR, note that in reality researchers normally have the flexibility to adapt their initial research questions to cope with unexpected results and it is arguably an important research skill to be able to report results whilst hiding preliminary wrong steps and unsuccessful suppositions (e.g., Latour & Woolgar, 1979). Nevertheless, a problem-based perspective on a new data source is at least constraining on the range of problem types that may be considered.

In terms of Stokes' (1997) types of research, ICR is a step away from the "use-inspired basic science" that a problem-oriented approach may take, towards the pure basic research quadrant because of its additional remove from a specific problem. As a consequence of this shift, there is a related risk that ICR will reify information sources, especially if it becomes institutionalised as a recognised field activity, and so this problem will need to be guarded against by researchers. This is most relevant to ICR articles, as described in the introduction. Knowledge of a range of social contexts is an important safeguard, perhaps through contact with a range of types of social issues or researchers from a variety of different fields.

Research funding objectives: A wider perspective

Even if the above argument for ICR is accepted within information science, it does not mean that it will be acted upon internationally. This depends upon various factors including the availability, capability and willingness of sufficient information scientists in each country to play an information hub role, perhaps extending the library paradigm to active ICR. Some additional important factors will be considered here, both predominantly external to the information science discipline.

The importance of ICR should be considered in the broader context of the informational turn in scientific knowledge creation (Wouters, 2006). The Human Genome Project is an exemplar of the increasing importance of large-scale distributed data sets in scientific research practice. The development of 'in-silico' experimentation and data production that the Human Genome Project innovated has been one of the factors leading to an exponential growth in the production of data in the life and biomedical sciences. Developments in the digitisation of scientific outputs

have resonated to other areas of scientific knowledge creation leading to what has been referred to as the 'data deluge' (Hey & Trefethen, 2003).

E-research has been a programmatic response by governments and funding bodies to address the challenges and opportunities presented by the development and application of advanced distributed computing resources and infrastructure in the sciences, social sciences and humanities. For example, in 2001 the United Kingdom initiated a £250 million, 5-year e-Science programme to develop tools, technologies, and infrastructure to support multi-disciplinary and distributed collaborations. Like the Cyberinfrastructure Program (Atkins, 2003) in the United States, e-Science in the UK embodies a vision that responds to the increasing needs of scientists for computationally intensive simulations, the management of ever-larger stores of data, and for shared access to expensive instruments.

The goals of e-Science have entailed a greater involvement of computer scientists in domain-specific problems, so it seems that e-Science is changing the roles and boundaries of computer science. This could indicate a Kuhnian (Kuhn, 1962) paradigm shift in computer science with the discipline fragmenting into interstitial applied areas between other disciplinary boundaries. The recruitment and retention of programmers in e-Science, however, has been problematic largely because domain-specific problems do not constitute interesting computer science problems for developers.

The notion of e-Science is also being taken up in the humanities and social sciences. In 2004 the National Centre for e-Social Science (NCeSS) was set-up in the UK. The same year, the Dutch Royal Academy of Arts and Sciences decided to fund the Virtual Knowledge Studio for the Humanities and Social Sciences (www.virtualknowledgestudio.nl). The aim of NCeSS is to investigate how digital tools and infrastructure developed during the five year UK e-Science programme can benefit the social science research community. In e-Science, infrastructure has been conceived as a generic resource that can support domain specific applications of e-Science through the development of tools and middleware. The central metaphor for this infrastructure is an Internet-based 'Grid'-like structure (Berman, Fox, & Hey, 2003), akin to a national utility system, such as the electricity grid.

One of the key issues for developing Internet-based services to support social and cultural researchers through the e-Science infrastructure has been the provision of access to data, combining data sets and developing tools for extracting information from them. Current projects include the development of tools for the distributed annotation of video-based data, Internet-based tools for visualizing geo-spatial data, the development of visual corpora and modelling and simulation tools. Although these data sets are not on the same scale as those being produced in the life and biomedical sciences, they do present access, processing and curatorial issues.

ICR is, therefore, of growing importance within the social sciences and to some extent the humanities, but thus far information science is not a discipline represented by activities and initiatives under the programmatic umbrella of e-research. The aims of ICR, e.g., the development of generic tools and the channelling of data and tools to relevant problem areas, fits well with the aims and direction of e-Social Science or e-research more broadly. Given the current climate in social science and the expertise of information science, benefit could be gained through the establishment of an ICR collaboratory that would provide tools, resources, training and foster collaboration amongst information scientists and support their interaction with social science and humanities research more broadly. The establishment of such a collaboratory would also contribute to an increase in the degree of technical-

certainty within information science through the diffusion of techniques and tools both within the discipline and across related fields.

Conclusion

The emergence of new information sources associated with new technologies seems set to continue and expand. Hence, there will probably be many initially plausible social science and humanities applications for data collected on a large scale from the new sources. This has created the need for a new information-centred style of research that seeks to identify for which research problems the data sources may be useful rather than assessing the data for a given research problem. Information science as a discipline is most suited to this role because ICR can be most easily be conceptualised as being within its disciplinary boundaries. Moreover, its combination of computing skills and contact with a range of social science and humanities fields gives many information scientists the necessary skills. Although ICR can be conducted by individual researchers and research groups, information scientists should organise and apply for funding for ICR because it is beneficial to the social sciences and humanities in general. Funded ICR hubs should probably take the form of national centres or, more informally, of individual research programmes designed to interface with existing national networks and centres, such as the UK's National Centre for e-Social Science.

The extent to which information-centred research is adopted on an international scale will be determined by factors including the success of early initiatives and the support of existing senior information scientists. The extent to which information scientists can play a role at the national level of brokering new information sources will be affected by political and practical considerations, determined perhaps by the influence of information science amongst senior national social scientists and the skills of national information scientists to persuade funders that this will support wider social science goals. This article has the primary purpose of arming information scientists for this task.

Acknowledgements

Thank you to the referees and to Anne Beaulieu, Iina Hellsten, Nick Jankowski, Jan Kok, Matt Ratto, Andrea Scharnhorst, Ernst Thoutenhoofd, Katie Vann, Charles van den Heuvel, and Sally Wyatt for comments on an earlier draft.

References

Atkins, D. (2003). *Revolutionizing science and engineering through cyberinfrastructure. Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure*. Arlington, VA: Directorate for Computer and Information Science and Engineering, National Science Foundation.

Bates, M. J. (1989). The design of browsing and berry picking techniques for the on-line search interface. *Online Review, 13*(5), 407-431.

Belkin, N. J. (1980). Anomalous states of knowledge as a basis for information retrieval. *Canadian Journal of Information Science, 5*(1), 133-143.

Belkin, N. J., & Croft, W. B. (1992). Information filtering and information retrieval: Two sides of the same coin? *Communications of the ACM, 35*(12), 29-38.

Berman, F., Fox, G., & Hey, T. (2003). The Grid: past, present, future. In F. Berman, G. Fox & T. Hey (Eds.), *Grid computing: Making the global infrastructure a reality* (pp. 9-50). Chichester: John Willey & Sons Inc.

Case, D. O. (2002). *Looking for information: A survey of research on information seeking, needs, and behavior*. San Diego, CA: Academic Press.

Foster, A., & Ford, N. (2003). Serendipity and information seeking: an empirical study. *Journal of Documentation, 59*(3), 321-340.

Hand, D., Mannila, H., & Smyth, P. (2001). *Principles of data mining*. Cambridge, MA: MIT Press.

Hey, A., & Trefethen, A. (2003). The data deluge: an e-Science perspective. In F. Berman, G. C. Fox & A. Hey (Eds.), *Grid computing: Making the global infrastructure a reality* (pp. 809-824). Chichester: John Wiley.

Hjørland, B. (2002). Domain analysis in information science. Eleven approaches - traditional as well as innovative. *Journal of Documentation, 58*(4), 422-462.

Hjørland, B., & Albrechtsen, H. (1995). Toward a new horizon in information science: Domain-analysis. *Journal of the American Society for Information Science, 46*(6), 400-425.

Ingwersen, P., & Järvelin, K. (2005). *The turn. Integration of information seeking and retrieval in context*. Berlin: Springer.

Kuhn, T. (1962). *The Structure of Scientific Revolutions*. Chicago, IL: University of Chicago Press.

Latour, B., & Woolgar, S. (1979). *Laboratory life: The social construction of scientific facts*.Los Angeles: Sage.

Paisley, W. (1984). Communication in the communication sciences. In B. Dervin & M. J. Voigt (Eds.), *Progress in Communication Sciences* (Vol. V).

Pyle, D. (1999). *Data preparation for data mining*. San Francisco, CA: Morgan Kaufmann.

Spink, A., & Cole, C. (2001). Information and poverty: The information environment of low-income African American families. *Library and Information Science Research, 23*(1), 45-65.

Stokes, D. E. (1997). *Pasteur's quadrant: Basic science and technological innovation*. Washington, D.C.: Brookings Institution.

Swanson, D. R., & Smalheiser, N. R. (1997). An interactive system for finding complementary literatures: A stimulus to scientific discovery. *Artificial Intelligence, 91*(2), 183-203.

Swanson, D. R., Smalheiser, N. R., & Bookstein, A. (2001). Information discovery from complementary literatures: Categorizing viruses as potential weapons. *Journal of the American Society for Information Science and Technology, 52*(10), 797-812.

Thelwall, M., & Wouters, P. (2005). What's the deal with the web/Blogs/the next big technology: A key role for information science in e-social science research? *Lecture Notes in Computer Science, 3507*, 187-199.

Thelwall, M. (2007). Blog searching: The first general-purpose source of retrospective public opinion in the social sciences? Online Information Review, 31(3), 277-289.

Wouters, P. (2006). What is the matter with e-science: Thinking aloud about informatisation in knowledge creation. *Pantaneto Forum, 23*, Retrieved February 20, 2007 from: http://www.pantaneto.co.uk/issue2023/wouters.htm.