VISUAL AND SONIC COMMUNICATION REPRESENTATIONS TO SUPPORT DESIGN INTERACTION

ANDREW BURROW, JANE BURRY, GREGORY MORE, YAMIN TENGONO, MARK BURRY

RMIT University, GPO Box 2476V, Melbourne VIC 3000, Australia

{andrew.burrow,jane.burry}@rmit.edu.au,
{gregory.more,yamin.tengon, mburry}@rmit.edu.au

Abstract. Interaction between collaborating parties exerts strong and surprising forces on the design process and thus its outcomes. Hence, there is much to be gained by investigating the nature of these communications. By definition, an organisation establishes clear protocols to realise effective communication, and does so in response to business needs. In contrast, creative endeavours must often harness loose teams of specialised individuals, working in different contexts, disciplines, and language bases, and relying on non-formalised yet highly selective exchanges of information. It is therefore a considerable challenge to comprehend and improve these processes. One approach is to exploit the discrete information contained within digital communications to produce models of these complex conditions. This paper reports preliminary work to identify the components of an email visualisation system. The goal is to apprehend developing structures, providing the collective with the capacity for reflection.

1. Introduction

Design is an inherently non-linear process that draws on many and often unexpected sources in the process of invention and iteration. In particular, the nature of the interaction between collaborating parties exerts strong and surprising forces on both the process and its outcomes. Hence, there is much to be gained by investigating the nature of communication in a design process, and good cause to represent and exploit this knowledge within the design process itself. By definition, an organisation establishes clear protocols to realise effective communication. These structures embody processes designed to respond to business needs. In contrast, creative endeavours must often harness loose teams of specialised individuals, working in different contexts, disciplines, and language bases, and relying on non-formalised yet highly selective exchanges of information. It is therefore a considerable challenge to comprehend and improve these processes. One approach is to exploit the discrete information contained within digital communications to produce models of these complex conditions.

This paper reports on work in progress that investigates the potential for systems to harness audio and visual spatialisation to depict and reify the information structures arising within large scale design collaborations. Experiments in representing real time information in navigable digital spaces, combining both visual and sonofication of data are as yet relatively untested, but the limitations of one dimensional lists and two dimensional matrices are all too apparent in locating, prioritising growing and changing information repositories fostering non-linear relationships. The nonlinearity of such relationships promotes the need to develop equally nonlinear manners of representation. By utilising 3D real-time graphic and audio synthesis technologies, this research seeks to understand the potentials non-traditional and nonlinear methods of informational representation. The ability to integrate spatial sound as an inherent dimension of digital Cartesian space offers the potential to expand the sensorial experience of a dataset beyond the field of view. In this work, sound is examined for its thematic and navigational potentials. This work focuses on navigable digital spaces that respond to interactive reading and listening, over and above the notion of static and canonical depictions of the information.

In the case of large-scale design collaborations, digital communication broadens the geographical and temporal opportunity for collaboration in real or near real time, but can also spin a denser web of fragmentary exchanges. Visualisation of the embedded information structures responds to this problem by enabling reflection on abstract relations in the communication. This feedback loop has the potential to improve communication: firstly, by enabling higher order analysis and problem solving on the communication itself; and secondly, by providing new types of access to the content of the communication. The first stage of this research involves the in depth examination of a substantial email archive for a collaborative design project that involved expert individuals in widely disparate fields distributed right around the world, working together on the highly innovative interactive, responsive, dynamic wall, Aegis Hyposurface©. Visual representation of information is an important approach in the process of, first, mining this data. In this respect it is analogous to the design communication process itself, which relies heavily on a combination of animated visualisations, video and photographs of prototypical experiments and the use of highly visually evocative written language to communicate about the design's conception and throughout its realisation.

2. Difficulties Encountered with Existing Email Archive Analysis

The proposed methodology for our current case-study based research into the potential for ontology-based tools for long distance cross disciplinary collaboration, called for the detailed analysis of an existing email archive of an appropriate [completed] collaborative design project. In theory, this should be a straightforward exercise that could be tackled by searching the archive database in obvious and less obvious ways. The first objective would be simply to establish the graph of the communication network, volumes in particular bilateral or multilateral directions, the changing form of the graph over time. It would then progress to more focused keyword, key phrase or key character searches of the email subjects and contents to try and establish individual and shared ontologies and again map where these were to be found and the changes that might occur in these over time.

In practice, this has proved a much more challenging and time consuming exercise than anticipated for three principal reasons: the quality and form of the raw data; privacy sensitivity; cognitive limitations in gaining real information from the data.

Dealing with these in order, there were 14 principal participants in the design project, whose email aliases totalled 25 in the course of the project. There were, in addition many more minor contacts. The raw data for the archive was sourced from only two of these, albeit principal participants and coordinators to whom the vast majority of communications would have been copied, if not directed. However it seems a reasonable supposition that the sourcing does influence the shape of the communication graph, in particular as you start to distinguish direct communication from third party copies. Still on the subject of the raw data, this was delivered in a multitude of very large overlapping archives. The email for the project that we had access to had not been filed separately from the general inbox in most cases. This could be overcome by searching the archive by participant names and aliases but this still demanded that the contents of every email be individually scanned as the subject lines rarely yielded useful information about the content and some of the participants communicated regularly on many subjects rather than being specifically linked by this project. This still left the problem of identifying and eliminating duplicates. These were generated in many ways and not always immediately identifiable. Many emails had gone to several accounts through automatic forwarding for instance, the raw archives overlapped in time span, and then there was the more fundamental issue of email strings that landed the same communication in the same place at different times via different routes. What should be included and what excluded? What should be regarded as a genuine duplicate? For instance where there were two otherwise identical emails, one of which was orphaned and one of which was flagged with a reference to related mail, the orphan was deleted.

On the second subject of privacy sensitivity, some of the email was of a highly focused, project specific, technical nature with very appropriate subject heading. However, much of it was contained in communications that ranged over a number of subjects not all project specific. Even the project specific content could be emotive, relating to emotionally tinged interpretations of other members of the team, strategic references to presentation plans, and attitudes to client as examples. This type of material is potentially rich as the basis of a thick description of the project [Ryle, 1971a,b, Geertz, 1973], a tool that could be used as a useful basis of comparison with the outcome of the database searches. However, it also limits the way in which the archive can be circulated and exposed in collaborative analysis work. This requires a set of neutral aliases and concealed content in any wider presentation to avoid privacy breaches.

Finally, on the subject that is most pertinent to this paper, it was very difficult to gain any meaningful overview of which communication channels were open and productive, which apparently silent or "neglected" at particular stages in the process. Whether communication activity one quarter gave rise to communication activity elsewhere. It was also unclear how heavily the map was biased by the source. The project was chosen for its exceptionally dispersed organisation. While the architect in Paris acted as the communication hub and client contact, the other participants were distributed around the whole world, never all met at one place at one time, and in many cases never met at all or even met the coordinator. Each participant, or in some cases partnership, came from a different organisation, and there was a diverse range of disciplines represented with limited knowledge base overlap in many cases. Reading through the email suggests there were some communication break downs or faltering between particular participants that lead to communication re-routings and offshoot communication eddies. But this type of information is difficult to confirm without lengthy numerical analysis. The overall complexity and diversity of this situation seemed to call for some form of semantic visual "at a glance" representation in the tradition of Geospatial Information Systems with their rich potential to extract widely different information from the same dataset.

3. Information Visualisation

Information visualisation emerged from the confluence of scientific visualisation and knowledge engineering including software visualisation, and from a growing appreciation of the perceptual and cognitive factors impacting successful user interfaces. It distinguishes itself from scientific visualisation in the nature of its data sets and the type of sense making that it fosters. Whereas scientific visualisation typically deals with the induction of observation from very large scale numeric data sets, information visualisation focuses on abstract relations and thus typically on discrete and symbolic structures. Card et al. [1991] provides an early account of the development of the field.

Another approach to understanding information visualisation is to trace developments in the field of human computer interaction. Shneiderman [1983] marks a pivotal point by introducing the notion of direct manipulation to describe systems enjoying the following essential properties.

> *visibility of the object of interest; rapid, reversible, incremental actions; and replacement of command language syntax by direct manipulation of the object of interest — hence the term "direct manipulation"*

These observations still resonate with current HCI efforts. Determining that the actions be expressed in the higher level domain, that the environment shift user tasks from cognitive to perceptual activities, and that we develop and rely on visual formalisms that bridge the formal and the manipulable.

Direct manipulation grows from these roots in two distinct directions. Card et al. [1991] and Shneiderman [1994] characterise an emphasis on the impact of quantitative improvements in computer graphics. Each attend to the dynamics of interactivity, making a connection between responsiveness and perceptual processing and expressing these facts in terms of software engineering requirements. Pineda [1988] and Harel [1988] consider the formalisms that contextualise user actions. Pineda attempts to formalise the concept of a meaningful drawing and provide a foundation for interactive graphics in AI. Harel proposes a role for visual formalisms that are both amenable to visual communication and machine manipulation and analysis. The two directions can be unified in the search for strengthening the role of positive affordances in user interfaces.

Gaver [1991] introduces the term affordance to the human computer interaction literature. Affordances are clues to the way in which an object may be manipulated or used, and thus to the functions it may fill. They are bound to the user's sensory experience of an object. Thus, an affordance is jointly determined by the user and the interface, namely, it is a relation between the properties of the interface and the cognitive models of the user [Thimbleby, 2001]. The provision of positive affordances in information visualisation is particularly challenging, because the sense making that might inform affordances is happening in situ, namely, the goal itself involves bringing affordances to abstract relations. Information visualisation can be characterised as the task of bringing positive affordances to abstract relations.

The search for visual metaphors for abstract models makes for a large space of possible designs for information visualisation systems. This space grows as new types of view are designed, as well as processes for mapping data to abstract relations. Furthermore, this search space may need to be realised in the user interface of an information visualisation system because quantitative changes in the data can result in the need for qualitatively different visualisations. For this reason, the information visualisation literature includes attempts to classify and taxonomise this design space: according to the data and visualisation transformations by Card and Mackinlay [1997]; according to cognitive aspects byWiss and Carr [1998]; and according to a sophisticated notion of the visualisation pipeline Chi [2000]. In the remainder of this paper, we analyse the problem of email visualisation in terms of the operator based taxonomy of Chi [2000].

4. Applying the Data State Model to Email Visualisation

The Data State Model [Chi and Riedl, 1998] factorises information visualisation systems into data operators. A particular information visualisation system is represented by a pipeline that proceeds from the raw data to the view. The pipeline is anchored to the source of data at one end, and to the views at the other end.

The left of Figure 1 depicts the four stage visualisation pipeline in the data state model. It can be understood as an amplification of the distinction, well established in user interfaces, between the state of an object and its presentation. For example, the subject and view in Unidraw [Vlissides and Linton, 1990]. The extra depth provides opportunities to further model the response of the data structures to user interaction, and introduces joints in the design at which the system can be reconfigured by replacing components.

The four stages can be characterised by an increasing focus on affordances. The value stage is characterised by data in its raw form, e.g., a collection of emails. The analytical abstraction stage corresponds to a model of the data that makes explicit the abstract relations in the data, e.g., a graph denoting the sender and recipient relations. In practical terms, it is distinguished by representing the visualised information in the whole, and before decomposition into a number of specific view which elide detail in favour of comprehension. The analytical abstraction thus lies on the boundary separating those parts of the pipeline under explicit system control from those responding directly to user interaction. On the other side of this boundary is the visualisation abstraction stage which is the context for algorithms to render views of the data into the final view stage. It is also the structure upon which user interactions are interpreted.

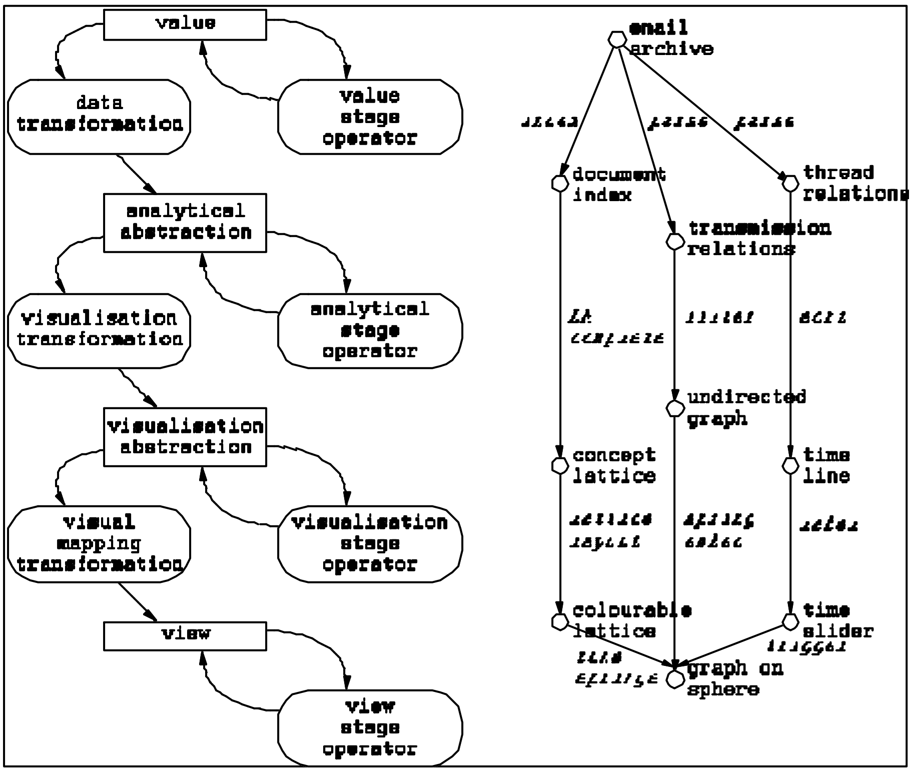

Figure 1: The information visualisation data state reference model of Chi [2000] and sketch of a visualisation pipeline.

The model includes further operations outside the data, visualisation, and visual mapping transformations directly implied by the four stages. These within stage operations allow us to consider a network of visualisation operators, rather than a simple one dimensional pipeline. Thus we are able to explore the interaction of a number of analysis products in the information visualisation. The right of Figure 1 depicts a possible visualisation pipeline. The data state model informs the engineering of an information visualisation system. By focusing on the data types and processing operations, it allows the space of possible operators to be classified and a taxonomy of operators and thus techniques to be constructed [Chi, 2000]. Thus, the design problem becomes one of search in a space of possible networks, where the value data stage is typically fixed and the view stage is constrained by the problem to be solved and the type of user. Another benefit of the data state model is that deepening affordances are characterised by user feedbacks to earlier stages in the pipeline. Drawing a particular pipeline as a network of operators makes explicit the operators effected by these feedbacks as they flow back down the pipeline. In the remainder of this section, we begin such an engineering analysis of our email visualisation system.

4.1. VALUE DATA STAGE

The value data is trivially described as an email archive. In order to instrument the collection we need configure the SMTP servers dispatching the participants' email. This avoids issues with reconstructing from the archives of individual participants. A well defined policy for this capture needs to be part of the work environment to prevent pollution of the archives with sensitive email.

4.2. ANALYTICAL ABSTRACTION STAGE

Since we aim to visualise the channels of communication operating within a design project, the sender and recipient relations are the basis of our model. Namely, we begin with a bipartite graph comprising the two classes **Emails** and **Participants** structured by the relation $\mathbf{sends} \subseteq \mathbf{Participants} \times \mathbf{Emails}$ and the relation $\mathbf{reaches} \subseteq \mathbf{Emails} \times \mathbf{Participants}$. An example is depicted in Figure 2.

The complex of relations here is real. Significant social messages are encoded in the inclusion and exclusion of participants during a single thread, e.g., certain emails may be carbon copied to a superior, or a subset of thread participants may switch to an exclusive conversation. At the other extreme email lists flatten the structure to a broadcast amongst participants, in which case the **reaches** relation becomes redundant.

Temporal structure arises from `Date` and `In-Reply-To` email headers. The time stamping of emails places the collection into a sequence, while the reference to identifiers groups the emails into a forest of trees commonly referred to as threads. A subtle distinction exists between the trees in the model and threads in an email reader, in that each participant's thread is a likely to be a subset of the entire tree. Clearly, the structure yielded by these timing relations is less reliable. In particular, we cannot accurately determine the time of reading, but must assume proximity to the time of sending. Likewise, a participant is not compelled to deliver their replies in any particular email.

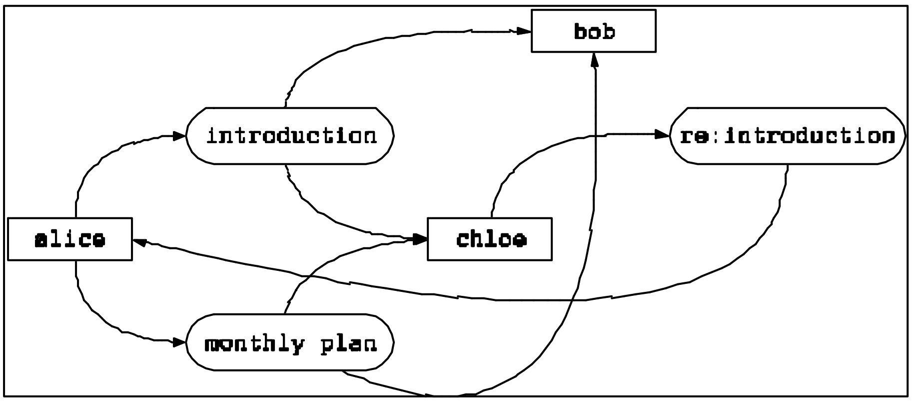

Figure 2: Alice introduces Bob to Chloe, who thanks Alice. Alice then sends work plans to both.

A distinct model in the analytical abstraction stage is the document collection. Emails are themselves documents, as well as being document containers via MIME. Thus an email archive enjoys the analytical abstractions of a document collection. In particular, it may be treated as an indexed collection [Witten et al., 1999], and supplemented by a relation denoting the containership.

In summary there are three models in the analytical abstraction stage. These are the transmission relations denoting the relationship between participants and emails, the thread relations denoting the temporal relationship between emails, and the document index denoting the collection of documents under full text indexing. The two models of transmission relations and document index interact in that either may be used to select subsets of the other, likewise for the thread relations.

4.3. VISUALISATION ABSTRACTION AND VIEW STAGES

In the remainder of this paper, we speculate on the contents of these stages and describe our future work. Following the suggestion at the start of this section, we traverse backward to meet the analytical abstraction stage. That is, we consider the questions we wish to answer in the visualisation, search the literature for views that might explicate the structures, and then triangulate to the analytical abstractions.

5. Graph Visualisation of Email Archives

Graph visualisation is a well established field, and plays a key role in information visualisation. Herman et al. [2000] suggest that graph visualisation is a valid strategy whenever the data contains inherent structure, as is obviously the case from our consideration of the analytical abstraction stage. The fie ld grew from graph drawing which considers the problems of two dimensional graph depictions [Di Battista et al., 1999], but has a distinct flavour in emphasising the handling of large graphs and the attendant need to provide abstraction, navigation, and interaction.

The most immediately obvious structures are the directed graphs depicting the transmission of messages between sender and recipient. If Alice sends Bob an email, this is depicted by an edge directed from a node representing Alice to one representing Bob. Such structures are innately susceptible to analysis based upon connectedness.

This directed messaging graph is one structure in a sequence that increasingly abstracts away detail. The initial structure is the the bipartite graph of Figure 2. It is a detailed view of the discourse explicitly representing participants and emails, and is appropriate where the archive is heavily filtered by a document index query. A structural elision collapses emails into single edges of a hypergraph, so that "introduction" and "monthly plan" are represented by a single hyperedge. A further elision recovers the messaging graph, in which directed edges represent the existence of email from sender to recipient. Finally, we can discard edges that do not occur in two-way pairs to construct an undirected graph depicting the existence of dialogue. In all but the first structure, an edge comprises a collection of documents.

Each of the structure discussed above is amenable to visualisation. This requires that the graph be spatialised, that components be allocated graphic and sonic marks, and that interaction and animation be considered [Card and Mackinlay, 1997]. Consider the semantic dimensions competing for attention. Immediately obvious are the topological and temporal dimensions of the transmission and thread relations. Additionally, for globally distributed design projects, the geographical dimension is worth preserving because non-overlapping work days accord with occlusion on the globe. Finally, we wish to include semantic dimensions extracted from the document index. Each of these dimension compete for spatial, sensory, and interactive allocations in the visualisation.

In graph visualisation, semantic dimensions are often mapped into spatial dimensions by clustering. Clustering algorithms partition, or hierarchically compose the elements according to distance measures. The literature considers the problem in terms of layout of clustered graphs [Eades and Huang, 2000] where the clusters are preexisting, or clustering from the layout [Sprenger et al., 2000] where the similarity measures are inputs to a spring-embedding system and the clusters are derived after layout from the spatialisation. The technique of Sprenger et al. [2000] allows competing similarity measures to be input as additional springs, for instance, to map the structure onto a globe [Gross et al., 1997]. We do not have space here to consider the graph theoretic approaches to clustering, such as cliques or producing skeleton structures such as spanning trees, but discuss conceptual clustering via the document index in Section 6.

The temporal dimension requires special handling. This is because temporal proximity is inadequate — the arrow of time. An information visualisation might devote a Cartesian dimension to time, animate changes over time, or depict the age of elements by some mark such as transparency. Chi et al. [1998] dedicate a spatial dimension to time in order to visualise structural changes in a document collection. Alternately, spring-embedding systems afford the opportunity to animate changes directly from the computation between successive states. This is because the node elements of each proposed graph structures are participants, and hence the set is relatively stable.

6. Conceptual Structures in an Email Archive

We close this paper by considering the role of formal concept analysis (FCA) in generating visual abstractions for email visualisation. FCA [Wille, 1992] is a mathematical approach to generating and visualising conceptual structures over data sets taken to represent the context of a discourse. Given a collection of objects, a collection of attributes, and the disposition of each object with respect to each attribute, FCA constructs a concept lattice. This structure clusters objects and dually attributes. Each point in the lattice denotes a term in the discourse, and each term is expressed equivalently by the indicated set of attributes or the indicated set of objects. This structure is represented and consulted as a diagram. Thus FCA generates a conceptual and spatial structure over collections.

Using FCA, the document index can be used to generate a visualisation of the terms which pick out key subsets of documents and attributes. Cole [2000] shows how FCA applies to the semantic structuring of an email archive, and provides algorithms to automatically layout the resulting concept lattices. Where the concept lattice is visualised, the transmission relations and document index each afford manipulations of the other: the concept lattice allows exploration and selection of email subsets to filter the graph visualisation, and dually the graph visualisation allows selection of contexts such as the emails associated with a cluster or transmission edge. The interaction of these two visualisation abstractions affords many interactions.

An additional avenue of visualisation is the temporal aspects of the FCA structures. Some work has been done on this subject by Neouchi et al. [2001]. It fits with the approach of Chi et al. [1998] but goes further in explicitly identifying the morphism from one concept lattice to the next.

7. Summary

There is a large body of existing work in information visualisation that pertains directly to email visualisation. In this paper, we have motivated the need to visualise evolving communications in large design projects, set out some of the difficulties in working with email archives, surveyed the field of information visualisation, and elaborated a systematic approach using the data state model to explore visualisation pipelines for email archives. Furthermore, we have identified key visualisation techniques, and made a connection to techniques for conceptual clustering of document collections.

The right of Figure 1 depicts a visualisation pipeline possible under the schemes described in this paper. A spring-embedding system maps the dialogue graph onto the sphere under the influence of weights supplied by colouring a concept lattice. The evolution of the system can be followed by altering the time interval for included emails. Future work will investigate this and other configurations.

Acknowledgements

Australian Research Council funding is facilitating research into Ontology-based Collaboration in Design

References

Stuart K. Card and Jock D. Mackinlay. The structure of the information visualization design space. In *Proceedings of IEEE Symposium on Information Visualization*, pages 92–99, Phoenix, USA, 1997. IEEE Press.

Stuart K. Card, Jock D. Mackinlay, and George. G. Robertson. The information visualizer: An information workspace. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Reaching Through Technology*, pages 181–188, New Orleans, USA, 1991. ACM Press.

Ed H. Chi. A taxonomy of visualization techniques using the data state reference model. In *Proceedings of IEEE Symposium on Information Visualization*, pages 69–75, Salt Lake City, USA, 2000. IEEE Press.

Ed H. Chi, James Pitkow, Jock Mackinlay, Peter Pirolli, Rich Gossweiler, and Stuart K. Card. Visualizing the evolution of web ecologies, 1998. Ed Huai-hsin Chi and John T. Riedl. An operator interaction framework for visualization systems. In *Proceedings of IEEE Symposium on Information Visualization*, pages 63–70, Research Triangle Park, NC, USA, 1998. IEEE Press.

Richard J. Cole. *The Management and Visualisation of Document Collections Using Formal Concept Analysis*. PhD thesis, Griffith University, Australia, December 2000.

Giuseppe Di Battista, Peter Eades, Roberto Tamassia, and Ionnis G. Tollis. *Graph Drawing: Algorithms for the Visualization of Graphs*. Prentice Hall, 1999.

Peter Eades and Mao Lin Huang. Navigating clustered graphs using forcedirected methods. *Journal of Graph Algorithms and Applications*, 4(3): 157–181, 2000.

William W. Gaver. Technology affordances. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Reaching Through Technology*, pages 79–84, New Orleans, USA, 1991. ACM Press.

Clifford Geertz. *The Interpretation of Cultures: Selected Essays*, chapter Thick Description: Toward an Interpretative Theory of Culture. Basic Books, 1973.

M. H. Gross, T. C. Sprenger, and J. Finger. Visualizing information on a sphere. In *Proceedings of IEEE Symposium on Information Visualization*, pages 11–16, Phoenix, USA, 1997. IEEE Press. David Harel. On visual formalisms. Communications of the ACM, 31(5):514–530, 1988.

Ivan Herman, Guy Melançon, and M. Scott Marshall. Graph visualization and navigation in information visualization: A survey. *IEEE Transactions on Visualization and Computer Graphics*, 6(1):24–43, 2000.

Rabih Neouchi, Ahmed Y. Tawilk, and Richard A. Frost. Towards a temporal extension of formal concept analysis, *Proceedings of the 14th Conference of the Canadian Society for Computational studies of Intelligence*, Ottawa, Canada, 2001. Springer Verlag.

Luis A. Pineda. A compositional semantics for graphics. In David A. Duce and Pierre Jancene, editors, *Proceedings of the European Computer Graphics Conference and Exhibition*, pages 155–169. North-Holland, September 1988.

Gilbert Ryle. *Collected Papers*, volume 2, chapter Thinking and Reflecting. Hutchinson, London, 1971a.

Gilbert Ryle. *Collected Papers*, volume 2, chapter 37, The Thinking of Thoughts: What is "Le Penseur" Doing? Hutchinson, London, 1971b.

Ben Shneiderman. Direct manipulation. *Computer*, 16(8):57–68, 1983.

Ben Shneiderman. Dynamic queries for visual information seeking. *IEEE Software*, 11(6):70–77, 1994.

T. C. Sprenger, R. Brunella, and M. H. Gross. H-blob: A hierarchical visual clustering method using implicit surfaces. In *Proceedings of IEEE Conference on Visualization*, pages 61–68, Salt Lake City, USA, 2000. IEEE Press.

Harold Thimbleby. Affordance and symmetry. In C. Johnson, editor, *Interactive Systems: Design, Specification, and Verification*, number 2220 in Lecture Notes in Computer Science, pages 199–217. Springer Verlag, 2001.

John M. Vlissides and Mark A. Linton. Unidraw: A framework for building domain-specific graphical editors. *ACM Transactions on Information Systems*, 8(3):237–268, 1990.

Rudolf Wille. Concept lattices and conceptual knowledge systems. *Computers Math. Applic.*, 23(6–9):493–515, 1992.

Ulrika Wiss and David Carr. A cognitive classification framework for 3-dimensional information visualization. Technical report, Luleå University of Technology, 1998. http://epubl.luth.se/1402-1536/1998/04/index-en.html.

Ian H. Witten, Alistair Moffat, and Timothy C. Bell. *Managing Gigabytes: Compressing and Indexing Documents and Images*. Morgan Kaufman, second edition, 1999.