An Overview Of Catalogue Design Problems In Resource Discovery

Andrew Goodchild

Department of Computer Science,
The University of Queensland.
QLD 4072, Australia

email: andrewg@cs.uq.oz.au
phone: +61 7 3365 4209
fax: +61 7 3365 1999

Abstract

This paper discusses some of the problems designers face in building catalogues in large networks and relates them back to the resource discovery problem. Currently many catalogues tend to be built in an adhoc fashion – which leads to a great variety in the quality of publically accessible network catalogs. Furthermore, the research surrounding these catalogues tends to focus on narrow technical issues – resulting in difficult to use catalogues. To address this problem, this paper provides a usability framework based upon the library science and human computer interaction literature, and demonstrates some of those principles via an example of a prototype. Our results are interesting to resource discovery tool developers in that we provide a framework for understanding the general resource discovery problem and present some techniques for dealing with those problems.

1 Introduction

The popularity of placing resources on electronic networks for large communities to use is rapidly increasing. An example of this phenomenon is the World Wide Web. Currently the design strategy for catalogues on the internet appears to be one of mimicking the more popular catalogues on and off the internet. The problem with such an approach is that what is popular dictates how most catalogues are built – and this can result in either re-invention of existing knowledge, or popularizing poor practices in catalogue design. Furthermore, the catalogues which are the most widely promoted inevitably become the catalogues that are the most popular, even though their content may be of a poorer objective quality than some lesser known harder to use catalogues. The focus of this paper is to consider usability issues in catalogues independently of current trends on the internet. Promotion of catalogues and assurance of the quality of catalogues' content are only considered here in as much as they are linked to usability.

A seemingly unrelated problem is occurring in the research field for resource discovery. Resource discovery research studies tools which allow users to search for and employ resources in large networks. A criticism of the resource discovery field is that it is driven by computing researchers who are focussed on narrow technical issues and tend to ignore existing knowledge about user search behavior.

Rather than mirroring existing design strategies; this paper surveys the library science and human computer interaction literature in order to identify what the design features of a well designed public access network based catalogues are. Some of the ideas from the survey are illustrated in a laboratory scale prototype of an advertisement broker.

Section 2 introduces the background topic of resource discovery, and discusses the motivating issues in that area. The central theme about the issues that make catalogue design difficult are discussed Section 3. Throughout this paper we illustrate potential approaches to these design problems via our prototype of an advertisement broker and section 4 gives an overview of the technology used to build the advertisement broker. Resource discovery as a research field and future work are discussed in section 5. Conclusions are presented in section 6.

2 Resource Discovery

The topics discussed in this paper largely belong to the emerging field of resource discovery. Resource discovery tools are systems which enable users to search for and employ resources in very large networks. Well known examples of resource

discovery tools include things like: gopher and its catalogue of resources "veronica", FTP and its catalogue of resources "archie", and the world wide web and its associated catalogues of resources like "infoseek" or "lycos".

Resource discovery tools typically support two main features: a network based protocol which allows access to remote services and presentation of information, and some general repository performing the role of a catalogue. From this, it is easy to see that the fields with the biggest influences on resource discovery are protocol engineering and information retrieval. To a smaller extent other fields like: artificial intelligence, human computer interaction, federated databases, multimedia, and security also have an effect on the nature of resource discovery research efforts.

The most popularly addressed issue is the impact of *scale* on resource discovery tools. Bowman, Danzig, and Schwartz (1993) , describe the issue of scale in terms of three dimensions: information system diversity, growing user base, and data volume. *Information system diversity* makes connection to the huge number of available remote services on networks difficult because they all use different access protocols and have heterogeneous representations for their data. The *growth in the user base* results in demand for services that can quickly outstrip the capacity a service is currently capable of. Finally the nature of public access networks is that there are few barriers to "publishing" information on the network and the growth in the user base has resulted in an explosion in *data volume* currently available. Even larger increases in data volume are expected as sound, image and video data are added to the network as well. Related to the increase in data volume is the "literature scatter" phenomenon – the more a field grows the more dispersed literature is in terms of both content and quality. Locating relevant information becomes increasingly more difficult as the size of the information space grows.

Development of resource discovery tools has primarily evolved to cope with the problems of scale. The diversity in information systems has been largely dealt with by standardization of remote access protocols and the construction of gateways between protocols. Examples of this includes the development of protocols like ftp, telnet, wais, gopher, and http for the web. Problems in increase in demand for services are dealt with by increasing the performance of the underlying hardware; like increasing the bandwidth of the network or the speed of the computer underlying the service. Given the public nature of some networks, this is not always a viable option, so other alternatives like caching and replication of services are sometimes attempted.

The growth of data volume on the network has led to a change in the way resources are cataloged. Early attempts at cataloging network resources mainly consisted of hand crafted catalogues distributed simply as files. Eventually the growth of the internet made such catalogues obsolete because they could not keep

up with frequent changes. The next generation of tools where designed primarily to cope with the problem of a frequently changing network; hence where focussed on automated information gathering. Examples of data collection systems are archie and veronica which automatically build indexes of of FTP sites and gopher sites respectively. Experience with these tools has led to more efficient methods of data collection and usage of more sophisticated indexing methods.

Recently the focus has shifted to providing an underlying architecture that provides facilities for distribution of services and support for gateways to services which rely on different remote access protocols. Such environments also provide mechanisms for indexing of services in many different formats and support for the development of new indexing methods. Examples of such tools include Harvest [Bowman, 1994], and OIL [Iannella et al., 1996].

It can be seen that research into resource discovery is heavily focussed on dealing with issues of scale, hence the bias towards protocol engineering and information retrieval. For a more detailed coverage of the area refer to Clifford Lynch's (1995) overview of current issues . This paper intends to explore other issues, mostly related to the common understanding of communities who use a catalogue, and how they relate back to the problems of scale endemic in large networks like the internet.

3 Design problems in network based catalogues

3.1 Surrogates

The key feature of catalogues is that they provide and organize a *surrogate* for each information source. A *surrogate* is a brief representation of a resource that is designed to convey an indication of the informations source's intent. Over time popular ways have emerged for providing resources with surrogates:

- *Titles, Abstracts, and Documentation* provide varying degrees of depth in explanation of a resources purpose.
- *Pictures* can describe visually what a resource is for.
- *Topics, genres, classifications, categories, types, etc* are all ways of grouping resources together by what the resources are generally about.
- *Lists of keywords* which are not necessarily organized into a sentence can be used to give the user some idea of a resource is about.

- *Other simple properties* like location of the resource, cost to use the resource, who the resource provider is, etc.

3.1.1 The "Perfect Surrogate"

Each of these surrogates have two main properties: firstly, the surrogate is based upon a common understanding in the community who uses the catalogue and secondly, the surrogates themselves can be organized into a way that is searchable. Therefore a "perfect surrogate" is one which strongly has both of these properties.

The first property of the surrogate is to describe the intent of a resource in terms of the public's common understanding. A resource in a tightly knit community is well understood, but once it is made available in a much larger more loosely knit community it is out of the context that helped users distinguish this resource from other resources. For example, in a publishing house, the concept behind a book's manuscript is understood by its author and publisher; but the community of potential readers would have difficulty telling the difference between it and other manuscripts. Therefore, one of the roles of the surrogate therefore is to convey the intent of the resource to its potential user. In this case the surrogate is performing the role of a what Brown and Duguid (1994) describe as a "portable context". A portable context is like the cover on a book which tells potential readers about the genre of the book it binds. Hence a resource in a large community must acquire more widely recognized, public indications of its potential uses – its surrogate needs to convey the resource's portable context.

The second property of a surrogate is to be organizable into such a way that resources they represent can be found without relying on serendipity (*i.e.* searching without relying on luck). Thomas Mann (1993) describes this feature as "predictability" – meaning that user can count on moving from whatever search strategy they think of to the strategy that will produce the best results. For example, predictability is a feature of the way the vocabulary is structured in the Library of Congress Subject Heading (LCSH) system. Links between subject headings in LCSH allow users to navigate to subjects which better describe their information need.

A "perfect surrogate" would be a surrogate which defines a portable context that all members of the catalogue user community readily understand, and which conveys the purpose of each resource. Furthermore, the style of surrogate lends itself to being part of a structure that can be searched in a predictive way.

3.1.2 Difficulties in finding the "Perfect Surrogate"

Finding a style of surrogate that performs the role of a portable context and can be structured to allow searching isn't easy. An example of a surrogate is a glossy advertisement for a product, which informs potential buyers of the benefits of a product. In this case the surrogate is effectively performing the role of Brown and Duguid's "portable context"; but it is very difficult to organize images into a way that they can be searched. This point illustrates that some surrogates make excellent portable contexts; but are difficult to organize. The converse situation also exists. The library of congress subject heading system is well organized and has many cross links to allow Mann's "predictability" concept. The subject heading system is used by less than half of library patrons [Hancock, 1987], indicating that there isn't sufficient overlap between the subject headings and the communities' common understanding; hence, although LCSH is well structured, it performs weakly as a portable context.

An example of a more effective surrogate are abstracts. Users can readily understand what an abstract is trying to convey (the portable context) and they can be easily searched using keyword matching (structurable to allow searching). But keyword searching is not entirely "predictable" as Mann would describe, since it is subject to problems associated with recall and precision. When performing a query one can rarely expect good recall (*i.e.* retrieving all relevant documents) without poor precision (*i.e.* not all the documents retrieved are relevant) and vice versa [Rowley, 1992, p171].

The choice of a surrogate must be based on the public's common understanding, but it seems that the more closely tied to common understanding a surrogate is the less suitable it is for building a searchable catalogue. The problem appears to be in that the public's common understanding is not easily usable as a scheme for structuring a catalogue. Support for this comes from Marcia Bates (1986) . Bates observed that common understanding of library users is subject to variety and uncertainty. Put simply, there are many ways to describe a concept, and the variety of ways increases the more specific the concept is and the variety of ways decreases the more general the concept is. For example, most people understand the concept of a "chair", but there are many ways to describe specific chairs: "the brown chair on wheels" vs "the cloth covered recliner". The variety and uncertainty inherent in common understanding seems to be related to the size of the community. It is easy to get a large group of people to agree on a common way of describing general concepts, but it is quite difficult to get them to to agree on a way to describe specific concepts.

3.1.3 Dealing with variety and uncertainty

One way of dealing with uncertainty and variety in searching is to narrow the focus of the catalog to specific topics in an attempt to attract mainly users who are already familiar with a topic. Such a user community is likely to share a common vocabulary, relieving some of the problems associated with variety, and the depth of common understanding alleviates some of the problems associated with uncertainty. This tactic works well for established fields, but poorly for newly formed fields in which the community has yet to settle on a common way of describing concepts.

Another way of dealing with uncertainty and variety in searching is to integrate both querying and browsing techniques. Queries can be used to bring users closer to their desired information source and browsing can be used in locating specific items. Searchers frequently rely on an integration between querying and browsing, because first attempts at queries often retrieve unsuitable results; but these results give users ideas on how to reformulate their queries to improve retrieval – this is commonly known as "query reformulation". Another behavior users frequently use is "known item searching" [Hancock, 1987], which involves searching for a resource that is already known and then browsing around that resource for other interesting resources.

An important aspect related to "query reformulation" is that users of network based catalogues frequently rely upon the ability to quickly inspect the resource described by the surrogate. Thomas Mann describes this feature of catalogues as "depth of access" [Mann, 1993]. No surrogate can ever match the information provided in the service itself, and in some searches only systematic evaluation of the service itself will provide users with enough clues on where to search next.

In practice, users conduct searches by using query reformulation, known item searching, depth of access, and many other techniques in endless variation. As Bates (1989) pointed out, from the standpoint of effectiveness in searching, the searcher with the widest range of search strategies is the searcher with the greatest retrieval potential. This means that information retrieval systems typically provide multiple ways of exploiting their file structure in searching; for example, this could be done through providing best match queries integrated with other browsing methods.

A method that incorporates ideas from querying and browsing is relevance ranking. Exact-match models, like boolean queries, have a problem in that they do not allow for any form of relevance ranking of the retrieved descriptions (*i.e.* it may be clear that some services are more related to the user's information need than others), and they exclude possibly relevant service descriptions that do not pre-

cisely match the query. Best-match retrieval models, like the vector space model [Salton, 1968], rank service descriptions according to their relevance to the query, thus introducing some browsing elements to help users deal with uncertainty and variety.

The variety of descriptions in surrogates themselves can be reduced. In keyword based systems this traditionally has been achieved by vocabulary control, *i.e.* reduce the variation by eliminating variety in words by morphological analysis to reduce different word forms to common "stems", for example: removing plurals, verb conjugations, synonyms, etc. More sophisticated techniques based upon natural language understanding have yet to be shown to be cost effective [Belkin and Croft, 1992]. The simpler techniques have had considerable success in improving search performance – but have their limits. We can decrease the variety in language used to describe information, but we cannot reduce the variety of information itself without defeating the purpose of the information retrieval system.

The variety in user's queries can be increased. The principal mechanism for achieving this is to use cross referencing. Users can increase the variety in their search by following links between terms. Another technique is to have users expand queries by selecting additional terms from lists suggested by the system. However, Belkin and Croft (1992) noted that this technique was not effective. The reasons for these differences are not obvious, although it appears that using only system suggestions is too restrictive and does not make full use of the user's domain knowledge. Another way of increasing variety was suggested by Salton (1968) . Salton enhanced the notion of query reformulation by proposing a method where the effectiveness of best-match queries can be improved by using elements from the result-set of one query in the next query – this method was called "relevance feedback".

3.1.4 Design of Surrogates in the Advertisement Broker

The advertisement broker itself is essentially a catalog of two things:

- *Advertisements* which are aimed at helping users find out what a service can do for them, *i.e.* the portable context for a service.
- *Service Offers* which describe details about how users (or their software) can make a connection to a service.

The surrogate for a resource in our case is an advertisement. Given the difficulty of finding a perfect surrogate, we chose to use many different styles of surrogates

for a resource in an advertisement to help users search. Each advertisement incorporates a number of features: title, the body of an advertisement which is either an abstract or an image, list of keywords, list of topics, and a list of simple properties like cost and location.

To help minimize problems associated with variety and uncertainty in searching the advertisement broker should focus its content on a particular topic. In our prototype we focussed on advertising stockbroking services. The advertisements are written by the service providers and can be entered into the broker via a standard form.

Advertisements can be associated with many service offers and vice versa. For example, an advertisement for an airline could be associated with three service offers: list flight arrival times, reserve a ticket, and cancel a reservation. Similarly, a service offer for listing flight arrival times could be associated with four advertisements tailored for different categories of customers. Advertisements can stand alone and just tell users the service provider's address and phone number (albeit this is not desirable from the "depth of access" perspective). Service offers should not stand alone, since without advertisements users will only discover the service by serendipity. The relationship between topics, advertisements, and service offers is illustrated in figure 1. Topics form a Rooted Directed Acyclic Graph (RDAG) with the root node as a fixed point for starting browsing operations. A RDAG of topics was chosen for its generality. To avoid the problems of the user being lost in the RDAG of topics, it was necessary to provide users with a means of determining where they are. We did this through the notion of "context" (for an example see figure 3 at the end of this paper). A context tells the user what are all the possible paths from the root of the RDAG to the current topic they are browsing.

Users have a number of strategies open to them. They can browse around topics, use keyword searches and search by properties to locate advertisements. Both the keyword search and the search by property techniques use best-match methods to achieve a semi-integration between querying and browsing. In keyword searching, advertisements are matched and ranked according to the number of matching keywords. Similarly, results of search by property are formulated using a best-match method, where advertisements are ranked according to the number of predicate atoms that match (for example, the search by property query in figure 4 results in the result set displayed in figure 5).

When a user decides that they want to know more about a particular advertisement they can click on a "more-info" button to retrieve a page (see figure 6) containing everything the advertisement broker knows about the advertisement. From this page the user can connect to the service the advertisement describes. At this stage the "back" button on the web browser becomes important, particularly if the

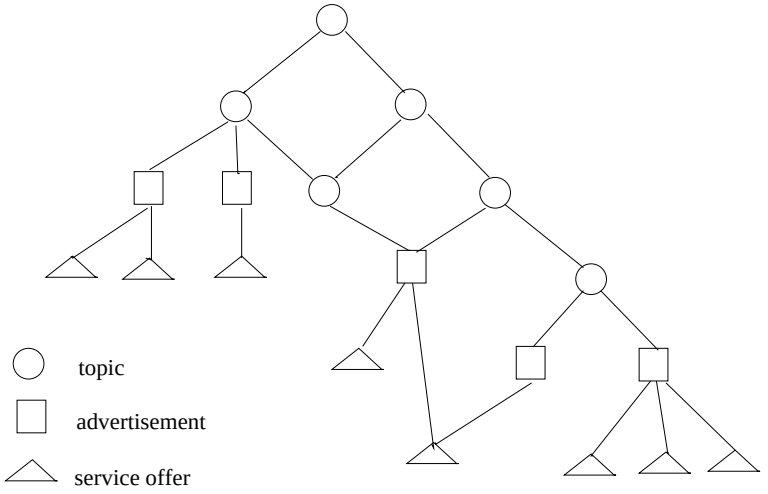

Figure 1: Relationship between topics, advertisements, and service offers

user is returning from the service after using a search strategy that relies on "depth of access".

Another design goal was to allow the user move between querying and browsing style of searching *seamlessly* so that we could facilitate "known item searching". We have partially achieved this goal by following the same principle by which a library works. When you find a particular book in a library, you often look at the books nearby to see if there is anything else related. To simulate this we provided a button next to each item on the full-info page (see figure 6), which the user can click on to browse advertisements of a similar nature. For example, when a user is reading what topics an advertisement can be found under they can start browsing those actual topic from that page. Similarly, when a user reads who the service provider is, they can locate other advertisements by the same service provider.

Presently the keyword matching facility does case insensitive substring matching. The effectiveness of the keyword matching facility can be improved by removing stop words (like "is", "the", etc) and reducing words to their stems (*i.e.* removing plurals and conjugates). Another extension includes adding relevance feedback to both the keyword searching and search by property facilities.

Response time and availability of the advertisement broker server is a critical issue as users are frequently reformulating their queries and new users are wanting to make requests. At present the advertisement broker server is a centralized database. Replication or distribution of the server maybe necessary to improve availability and response time. Since the server has been built upon Oracle v7.0,

it seems that this could be approached using Oracle's parallel server or distributed database technology with little migration effort.

3.2 Resource Discovery as a Pragmatic Activity

Frequently resource discovery episodes occur in conjunction with some other activity, like writing a report, performing a review, or using information gained in one resource discovery episode to aid in another episode. Resource discovery has a *pragmatic* aspect to it. In other words, the user is prepared to act upon the information they find, as opposed to just finding it to satisfy curiosity. This pragmatic nature has two implications for the catalog: authority needs to be managed and support must be made available for integrating resource discovery into work activities.

3.2.1 Managing the authority of a catalogue

Because people are prepared to act upon information they gain from the service, they expect to know how "truthful" or how "authoritative" it is. Failure to pick up on the elements that contribute to a service's authority will lead to a service being perceived as untrustworthy and mainly used to satisfy curiosity.

Increasing the authority of the catalog itself essentially requires a catalog provider to ask themselves the question: "if I perform a search on this catalog will I find all the relevant services, or will I have to continue the search elsewhere?". To deal with this the catalog provider not only has to provide some mechanisms for dealing with variety and uncertainty, but they also have to provide some way of ensuring that the catalog includes most of the relevant services and maintains the truthfulness of existing catalog entries. To ensure that the catalog includes most of the relevant services usually involves either having a high profile catalog that every service wants to advertise in, or having scouts locate new services to include in the catalog. The authority of the services included in the catalog requires the catalog provider to authenticate things like:

- *The location of services.* As services on networks frequently move, information on the location of services needs to be maintained.
- *The authority and identity of the author of the service descriptions*. The authority of the author describing services directly determines the authority of the catalogue. In addition many catalogues contain information on competitive services, so there needs to be some guarantees that other services entries are not falsified by competitors.

- *Claims made in service descriptions*. Not all claims made in entries in a catalogue are truthful. There are a few reasons for this: the nature of the service has changed since the description was written, the author of the description is deceiving users to gain a competitive advantage, or the author of the description misunderstands the nature of the service.

Authentication is a time consuming exercise, and large catalogs like the Yellow Pages Phone Book often do little more than annually ask existing customers if their advertisements are still correct. Such catalogs cannot vouch for the truthfulness of claims made in catalog entries, and their user communities understand this. Some of the things that contribute to a catalog's authority can be made available by the catalog provider when the catalog is created, but others are implicit and discovered only through a community's growing familiarity with a catalog [O'Day, 1994].

3.2.2 Integrating resource discovery into other activities

Librarians [Hancock, 1987, Bates, 1989, O'Day and Jefferies, 1993] frequently observe that searching is an adaptive process. Each new piece of information a user encounters gives them new ideas and directions to follow and, consequently, a new conception of the search. Furthermore, at each stage, with each new conception of the search, the user may identify useful information and references. In other words, the search is not satisfied by a single retrieved set, but by a series of selection of bits of information at each stage of the search. Bates (1989) used the term "berry-picking" to describe this bit-at-a-time information retrieval.

Resource discovery, or searching in general, is more than this "bit-at-a-time" retrieval process. These "bits" are frequently accumulated in places like reports that users are writing. In this sense resource discovery starts to more closely resemble research and is now integrated into some of the more creative aspects of user's work activities.

This style of integration requires the user interface to the resource discovery tool that is compatible with the other tools a user uses. For example, it ought to be a simple task to locate a document via a resource discovery tool and then save a reference to it in a local bibliography tool on the user's workstation. Another example, would be to locate some statistical information and then save it into a local spreadsheet. This concept of compatibility between the user interface and local software goes further than suggested here. Users become accustomed to a particular set of tools, and their resource discovery client is one example of this. Perhaps there should be some way of connecting the user's preferred resource discovery interface to different resource discovery services. Both of these things are achievable

though standardization of data exchange formats and interaction methods between user interfaces and resource discovery services.

Research into personal client software that can take advantage of user's history or preferences in searching is starting to emerge. While it is important for user's to be able to specify their preferences, having a system that guess at a user's preferences based on their history of searching can be detrimental. Users usually seek predictable systems and shy away from complex unpredictable behavior [Shneiderman, 1992]; thus a personal client that keeps on unexplainably adapting could end up frustrating users.

3.2.3 Incorporating the Advertisement Broker into Pragmatic Activities

The advertisement broker is seen as a tool that goes beyond the Yellow Pages phone book; where users have many different ways of locating advertisements and can obtain connections to underlying services. This means that the expectations of the broker user community with respect to authority are similar to the Yellow Pages. The similarities are that they expect the broker to contain most of the relevant advertisements and do not expect the broker to authenticate the claims of advertisements.

The broker is different from the Yellow Pages phone book in that the users expect the broker to be updated more regularly than annually and the broker needs to have fast response times and be rarely down. Maintaining up-to date advertisements and service offers is not managed at present, but maintenance of these descriptions could be facilitated using transponder technology proposed by the IETF [Weider, 1994]. In this scheme each service provider maintains a local copy of their advertisements, service offers and a transponder. When the service provider updates their local copy, the transponder informs the advertisement broker of the changes.

At present the advertisement broker is poorly incorporated into user's work activities. Currently integration is achieved by using "cutting and pasting" of results between windows allowed in the X-windows system.

4 Architecture

The construction of the advertisement broker was relatively straight-forward. The division of components was done in a standard fashion: a user interface, a server which contained the catalogue itself, and a client to mediate communication between the user interface and the server. The components where built using well

understood technology: web browser for a user interface, http and rpc for communication between components, and a relational database as a repository for the catalogue. More details are provided below:

Advertisement Broker Server The server contains two components: an Advertisement Broker Database which is a database of advertisement and service offer records and an Advertisement Broker Agent which manages client connections to the server. The Advertisement Broker Database was implemented using relational database technology. Oracle v7.0 was selected as a database management system because of its availability and our familiarity with it. The Agent uses Oracle function calls to execute the Advertisement Broker Client's SQL queries on the database.

Advertisement Broker Client The Advertisement Broker Client provides the user's connection to the Advertisement Broker Server. Users interact with the Advertisement Broker using a WWW interface. We chose to implement the Client using Web technology for a few reasons: the Web is a simple environment to prototype in, it is easy to demonstrate the broker at remote sites, and of course its popularity as a resource discovery tool. The Advertisement Broker Client is a cgi-script written in C and submits requests to the Advertisement Broker Agent in the form of SQL queries and processes the results of SQL queries to generate the next Web page that the user sees.

Communication Layer In the current architecture (see figure 2), the position of the Advertisement Broker Server in the network is assumed to be known to advertisement broker Clients, therefore UNIX RPC technology has been sufficient to implement connectivity in the prototype. The prototype recognizes two forms of service offers: SQL databases and Web pages. SQL databases in general provide read access to data stored in remote relational databases. Connection to Web pages is managed by passing the page's URL from the server to the client and then letting the client load the document into the user's WWW interface. Connection to SQL databases are currently managed by hard-coding the location of the SQL database into the Client and then letting the Client make the connection to the SQL database.

5 Resource Discovery as a Research Field and Future Work

One of the more difficult questions frequently asked about the resource discovery field is: "where is the computer science?". Depending upon how you interpret

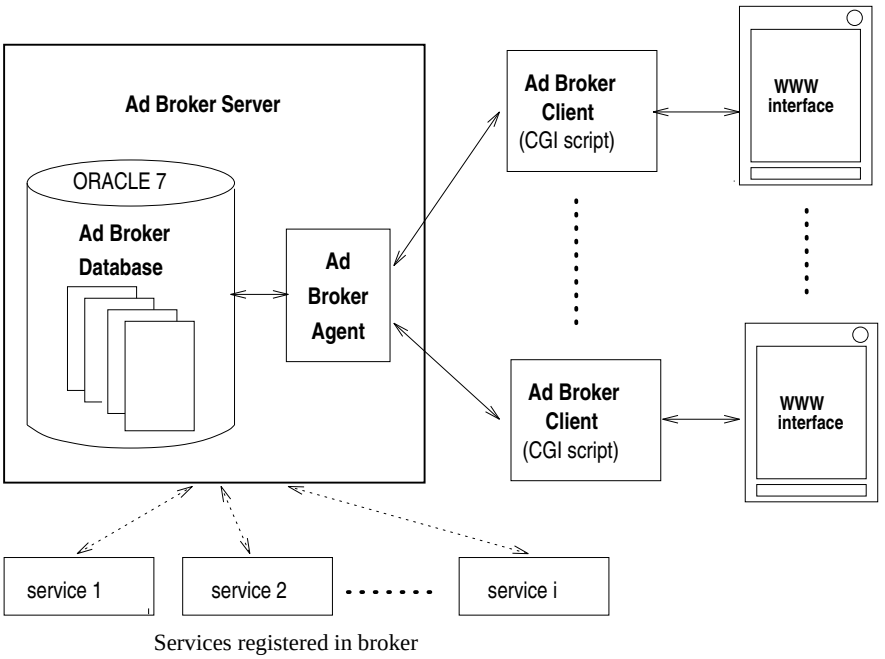

Figure 2: Architecture

this question, this can be viewed as one of the bigger problems that resource discovery is facing in establishing itself as a respectable research field. Computer science roughly consists of two major disciplines: systems and technology. Systems researchers consider a problem as a whole, or a system, and try to construct an architecture to deal with the problem. Technologists tend to focus on narrow issues which can be solved by applying rationalistic methods like deductive mathematical laws. Resource discovery is field that tends to favor more systems style research rather than research of a technical nature.

The core of the problem lies in that resource discovery is inevitably dependent on the common understanding of its user community; meaning that anything too novel is unlikely to be used. As a result there do not seem to be many technical difficult problems to be solved in resource discovery – meaning that from a technologist view point resource discovery is not a respectable field of research. This paper, for example, is suspect because it provides high level overview of the resource discovery problem, and the only technical thing it describes (the advertisement broker) does not show significant technical merit.

The value of this paper comes in providing a overview of the issues related

to catalogue design and then using those highlighted issues to guide future work. Hence from a systems point of view, this paper provides some insights into how a resource tool should be constructed. Much of the current work in resource discovery is suspect, from a systems point of view, of focusing on narrow technical issues like efficient algorithms for data collection or architectures for distributing tools around the network. Such a focus can lead to the development of tools that miss some important points with respect to usability. The advertisement broker was built as a hands on demonstration of some of the points about catalogue design made in this paper to our local research community. It demonstrates the baseline services a catalogue should offer, and gives researchers an idea as to where novel extensions can be added.

5.1 Future Research in Resource Discovery

Current work in resource discovery stresses deriving surrogates from the services themselves. In terms of protocol spaces like gopher or the world wide web, this is fine because these protocol spaces are quite rich in text. In fact this focus may even be preferable given the constantly increase size and diversity of the internet. Deriving surrogates is not always possible. For example, sounds, images, and non-document databases are very difficult to derive descriptions from. Such services must rely on hand crafted surrogates. The combination of derived surrogates with hand crafted surrogates in the one catalogue is a poorly studied topic in resource discovery. Furthermore, the inclusion of evaluative information in resource descriptions to help users select suitable resources is also poorly understood.

An example of mixing derived surrogates with hand crafted surrogates in the one catalogue is trying to advertise databases. Our motivation for building the advertisement broker was to explore problems in advertising database services; but we never got fully into that since we first needed to understand the resource discovery problem from a systems point of view. Now that work has been completed to satisfaction we can now look more deeply at the problem of advertising structured database services. Database services are a prime candidate for exploring surrogates which mix hand crafted and algorithmically derived data. The data in the database and the schema provide much information that can be used in deriving surrogates, but is difficult to interpret without knowing what the intent of the database is. For example, an employee database provides many useless names and address; unless of course you know what organization/department it is from. This additional information can only be assigned in a hand crafted surrogate. Furthermore, some databases are more reliable than others, and this knowledge of reliability is only gained through experience and cannot be derived. We intend on exploring the rela-

tionship between assigned, derived, and qualitative data in surrogates for databases.

Personal client software offered by current resource discovery services are fairly primitive, and there seems to be a focus on the server side of the resource discovery tool. Since resource discovery is usually linked to other creative activities like writing a document or research, users can benefit from resource discovery tools which fit nicely into the user's other work activities. Suggestions for improvement to this include the development of personal client software, as suggested by Robert Colomb (1993) , that supports seamless interchange of data and operations between local tools like word processors or spreadsheets and the resource discovery tools. Another aspect for improvement is the development of personal clients that can connect to a number of different resource discovery services while maintaining the same user interface – thus allowing user to continue using their preferred mode of interaction.

6 Conclusions

In this paper we presented a framework for understanding the principles behind catalogue design, illustrated them with a prototype of an advertisement broker, and examined current research in resource discovery. Unlike most other resource discovery tools which are typically built by mimicking existing catalogues or focussed on narrow technical issues, we took a user centric approach to the design of the advertisement broker. A survey of the library science and human computer interaction literature was conducted, and issues that make resource discovery difficult for users where identified. We then proposed some approaches to dealing with these difficulties and included them in the design of the advertisement broker. The main thrust of future work will be to use advertisement broker as a platform which we can extend to explore the problem of how to discover and use database services.

References

[Bates, 1986] Bates, M. (1986). Subject Access in Online Catalogs: A Design Model. *Journal of the American Society for Information Science (JASIS)*, 37(6):357–376.

[Bates, 1989] Bates, M. (1989). The design of browsing and berrypicking techniques for the online search interface. *Online Review*, 13(5):407–424.

[Belkin and Croft, 1992] Belkin, N. and Croft, B. (1992). Information Filtering and Information Retrieval: Two sides of the Same Coin? *Communications to the ACM*, 35(12):29–38.

[Bowman et al., 1993] Bowman, Mic, C., Danzig, Peter, B., and Schwartz, Michael, F. (1993). Research problems for scalable internet resource discovery. Technical Report CU-CS-643-93, University of Colorado at Boulder.

[Bowman, 1994] Bowman, M. (1994). Harvest: A Scalable, Customizable Discovery and Access System. Technical Report CU-CS-732-94, University of Colorado, Boulder, USA.

[Brown and Duguid, 1994] Brown, J. and Duguid, P. (1994). Borderline issues: social and material aspects of design. *Human Computer Interaction*, 9(1):3–36.

[Colomb, 1993] Colomb, R. (1993). Use of a personal workstation to access open network services. *Australian Computer Journal*, 25(1):7 – 13.

[Hancock, 1987] Hancock, M. (1987). Subject Searching Behaviour at the Library Catalogue and at the Shelves: Implications for Online Interactive Catalogues. *Journal of Documentation*, 43(4):303–321.

[Iannella et al., 1996] Iannella, R., Ward, N., Wood, A., Sue, H., and Bruza, P. (1996). The open information locator project. Submitted to ADL'96.

[Lynch, 1995] Lynch, C. (1995). Networked information resource discovery: an overview of current issues. *IEEE Journal on Selected Areas of Communications*, 13(8):1505–1522.

[Mann, 1993] Mann, T. (1993). *Library Research Models*. New York: Oxford University Press.

[O'Day, 1994] O'Day, V. (1994). Intermediares Sustain Borders. *Human-Computer Interaction*, 9(1):105–108.

[O'Day and Jefferies, 1993] O'Day, V. and Jefferies, R. (1993). Orienteering in an Information Landscape: How Information Seekers Get From Here to There. In *Human Factors in Computing Systems (INTERCHI'93)*, pages 438–445.

[Rowley, 1992] Rowley, J. (1992). *Organizing Knowledge*. Aldershot, Hants, England: Ashgate, second edition.

[Salton, 1968] Salton, G. (1968). *Automatic Information Organization and Retrieval*. New York: McGraw-Hill.

[Shneiderman, 1992] Shneiderman, B. (1992). *Designing the User Interface*. Reading, Mass: Addison Wesley, second edition.

[Weider, 1994] Weider, C. (1994). Resource transponders. Internet Draft: ietf-iiir-transponders-02.

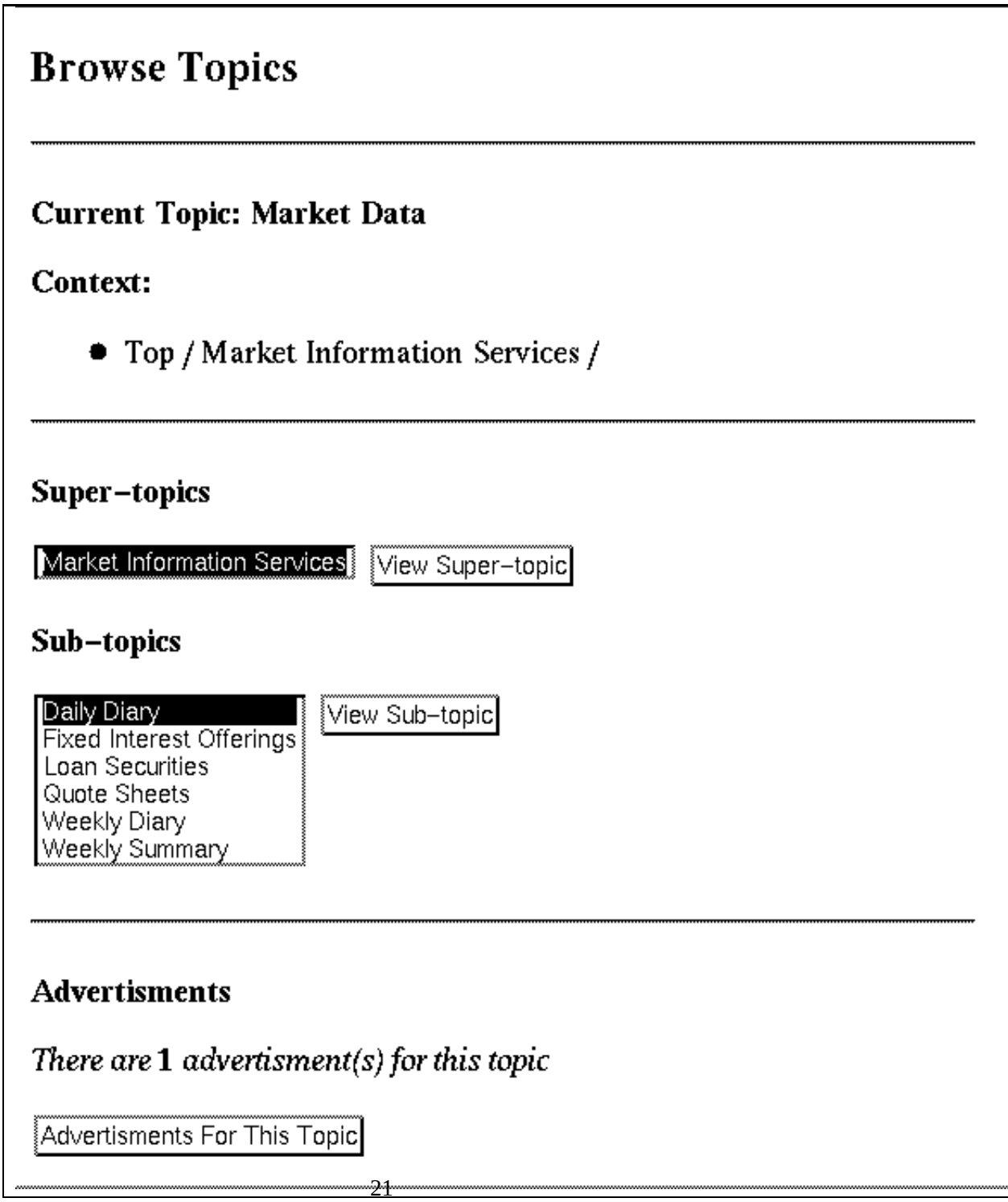

Figure 3: Browsing a Topic

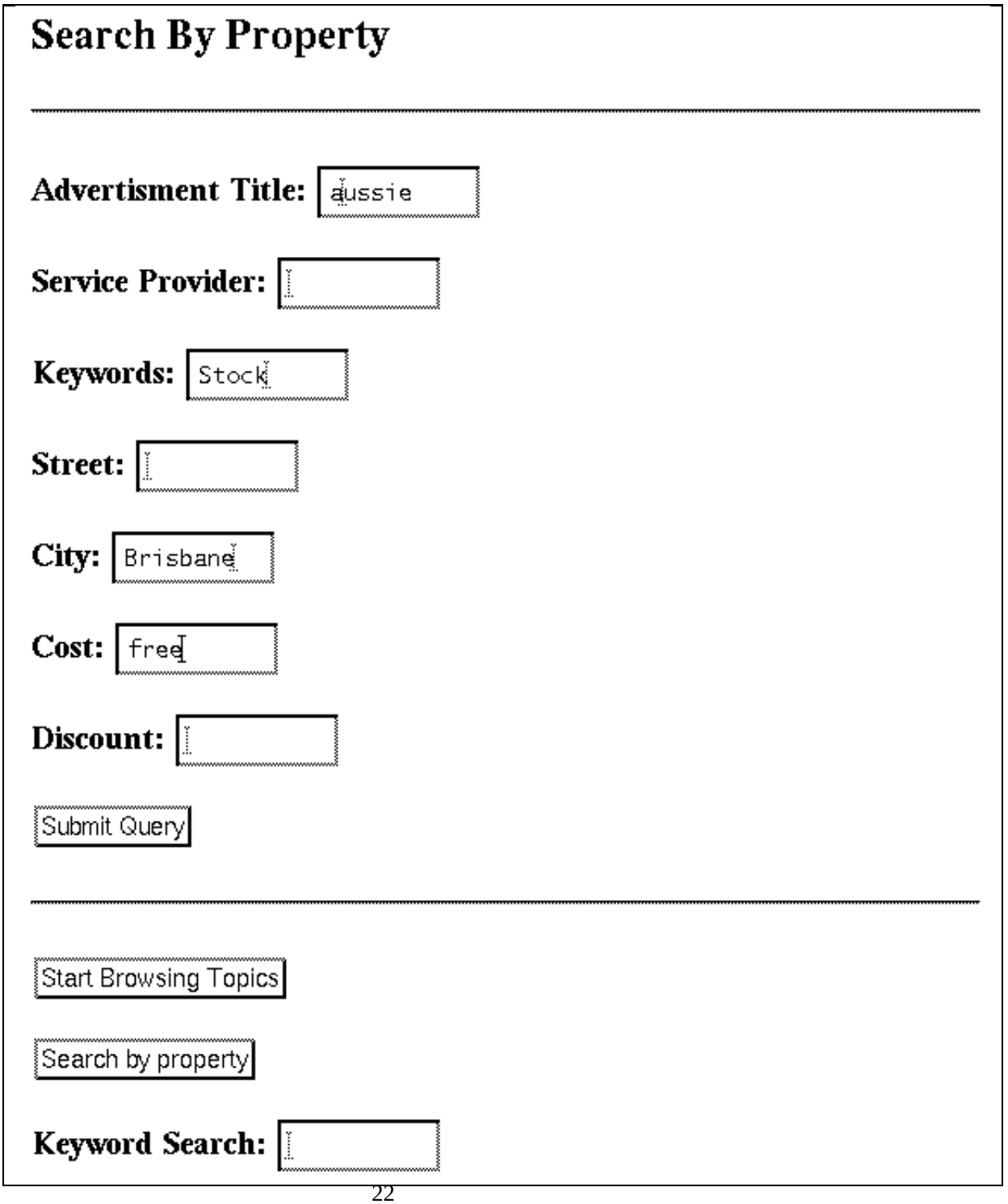

Figure 4: Entering a search by property query

Result Of Search By Property

Query:

- title = aussie
- keyword = stock
- city = brisbane
- cost = free

Advertisment Title: Murphys Stockbroking

3 properties matching: city = brisbane, cost = free, keyword = stock,

More Info

Advertisment Title: Aussie Stock Exchange Ltd

3 properties matching: city = brisbane, keyword = stock, title = aussie,

AUSSIE STOCK EXCHANGE

Figure 5: Result of a search by property query

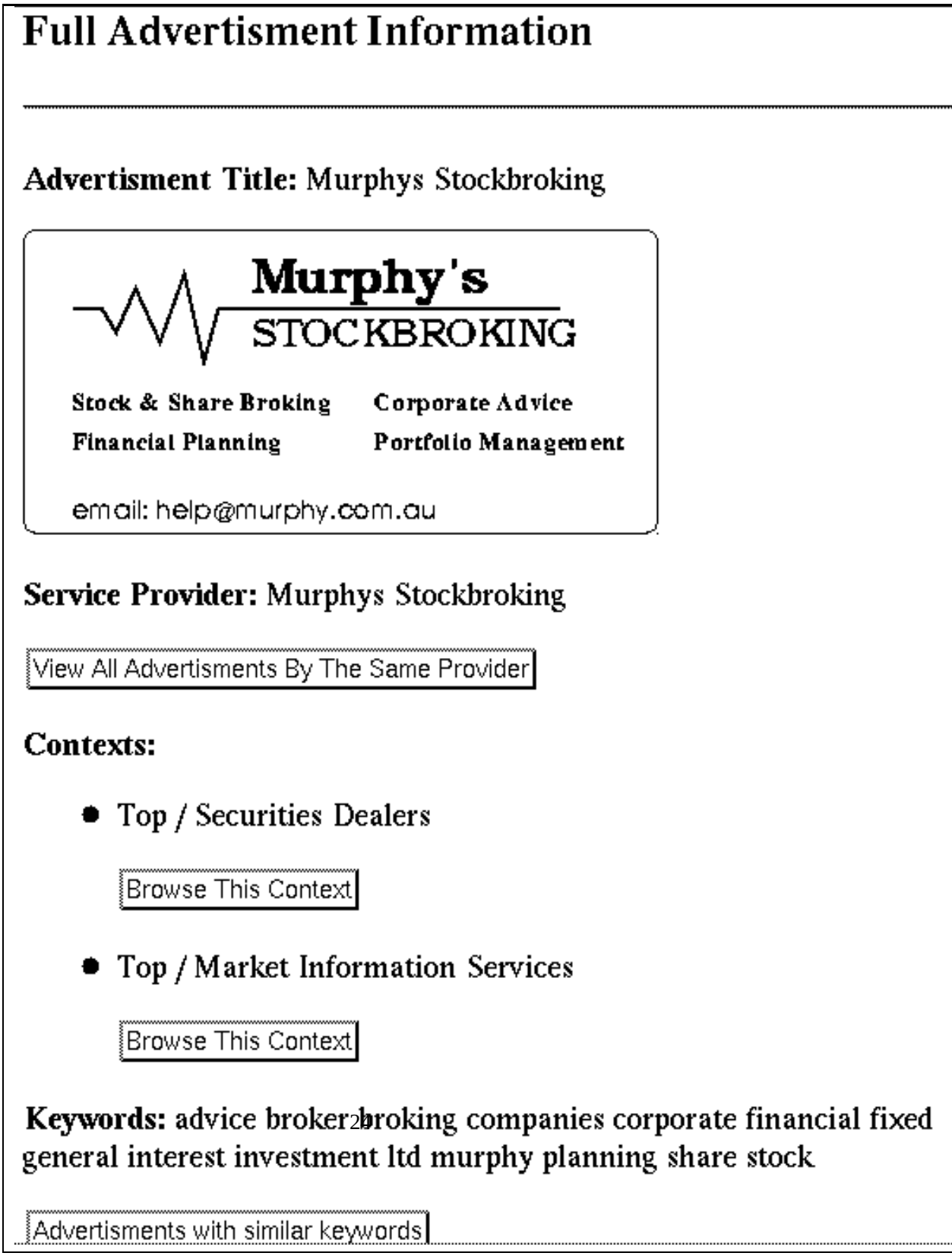

Figure 6: Full information page on a service offer