Conceptualising Resource Discovery on the Internet

Sampling Search Services

Gisle Hannemyr
gisle@ifi.uio.no
Department of Informatics
University of Oslo, P.O.Box 1080, Blindern, N-0316 Oslo

Abstract

Information retrieval by means of Internet search portals has so far been treated as an extension of "classic" information retrieval technology adapted to search data captured from the Internet (as opposed to data captured from analogue media such as newsprint and periodicals). This paper argues that the classic information retrieval model is inappropriate for resource discovery on the Internet, and try to explain why this is so. Some resource discovery services on and off the Internet are studied, and resulting from this analysis, some new concepts that captures aspects of Internet resource discovery are proposed. It is conjectured that the present lack of awareness of such concepts results in impaired performance for Internet search portals.

Keywords: Internet, information retrieval, searching, metadata, resource discovery

1. Introduction

Lynch (Lynch, 1997) has described the Internet as "a chaotic repository for the collective output of the world's digital printing presses". Since the Internet was transformed from an engineering activity to a social phenomenon around 1994, millions of individuals and organisations have contributed text, images, audio and video recordings in a multitude of genres, formats and languages.

A number of services have been created to let users "search the Internet", but these often fail to satisfy. Common complaints from users are that Internet searches result in too many responses in no apparent order, that many of the responses are not relevant (i.e. outdated, or not helpful to the problem at hand), and that many of the responses yield data that are of low quality or are downright misleading. Quality resources, such as refereed academic papers or online collections maintained by reputable institutions, are difficult to discern among search results that are irrelevant, out of date or misleading. Sometimes, even extensive searches fail to reveal the existence and location of a particular resource that is known to exist on the Internet.

The author is currently working (in co-operation with the company *FAST Search and Transfer*, that develops technology to search the Internet) in a project to study and (we hope) improve Internet search technology.

In order to understand better the characteristics of Internet search portals, a study of information search systems focusing on data capture and management were conducted during 1999. The purpose of the study is to derive a tentative conceptual framework for

Proceedings of IRIS 23. Laboratorium for Interaction Technology, University of Trollhättan Uddevalla, 2000. L. Svensson, U. Snis, C. Sørensen, H. Fägerlind, T. Lindroth, M. Magnusson, C. Östlund (eds.)

resource discovery. The intention is to provide a better foundation to our own and others efforts to build computer systems for resource discovery on the Internet. This paper summarises the findings and conclusions from this study.

Please note that the present study does not cover issues related to graphical design and user interface. These issues are, of course, important to determine the overall characteristic of an Internet search portal, but will be dealt with in a separate study. For an overview on current research on search system usability and user interface design, see ch. 8 of (Nielsen, 1995), ch.4 of (Nielsen, 2000), ch. 6 of (Rosenfeld and Morville, 1998), and ch. 4 of (Spool et al., 1999).

1.1. Methodological Approach

The field study has been conducted by observing users engaged in resource discovery. The users have been observed in normal situations at work or at home. Usage situation involving a number of different tools and strategies (ranging from looking up books in a library card index to using state of the art Internet search portals) has been studied. Users have sometimes been asked to briefly explain what they are doing and why, but intervention has been kept at a minimum.

I have also conducted a series of longer semi-structured interviews with facilitators of resource discovery (i.e. librarians, cataloguers, metadata and classification experts, search engine programmers). These interviews have focused on methods and concepts, and on the type of problems they come across in their every day work.

The study has also included examinations of various artefacts commonly used to facilitate resource discovery (e.g. index cards, classification systems, relevant technical standards, and several information search systems).

Also, a survey of relevant literature on resource discovery has been conducted. Interestingly, most of the papers surveyed reported only on purely technological or quantitative aspects of different systems to search the Internet (Lesk, 1989, Lynch, 1997, Silverstein et al., 1998, Page et al., 1998), or on quantitative aspects of the Internet itself (Lawrence and Giles, 1998) (Lawrence and Giles, 1999). The single exception I have found so far (Dehn and van Mulken, 2000) reports on qualitative research, but only on subjects studied in laboratory environments.

The point of departure for my research is “grounded theory” as developed by Barney G. Glaser and Anshelm L. Strauss (Glaser and Strauss, 1967). The emphasis of grounded theory is on exploration and investigation. Findings are inductively and empirically derived directly from the data, rather than from interpreting the data in light of some existing theory. Hence, the analysis resulting from grounded theory is intimately and directly linked to the data, “to tease out themes, patterns and categories” (Easterby-Smith et al., 1991, p. 108). Creation of logically deduced theory that “explains” the data fall outside this methodological approach.

For in-field data collection, I’ve relied on a method developed by Karen Holtzblatt that is known as “contextual inquiry” (Beyer and Holtzblatt, 1998). They neatly summarise their approach as follows:

> «… go where the [user] works, observe the [user] as he or she works, and talk to the [user] about the work …» (ibid. 41)

One of the distinguishing characteristics of “contextual inquiry” compared to other types of usability studies is that nothing is “measured” or “tested”. The purpose of the study is

to acquire a rich empirical basis for the ensuing analysis. Four key principles (partnership, context, focus and interpretation) characterises the method of "contextual inquiry" (ibid, p. 46-64). They are briefly summarised below:

- **partnership**: both the traditional researcher/research-object and the consultant/client relationship are abandoned in a favour of a more equal-footed relationship where the researcher works with the user to establish a partnership where mutual learning about the subject matter is the objective. Observation, semi-structured interviews and walkthroughs are used to uncover various aspects of the activity.

- **context**: the activity is studied in the context it usually takes place, in the form it usually takes place and carrying out the task and solving the problems that arise through the activity. This in contrast to traditional usability studies, where the activity is studied in a usability lab, and where the tasks to be solved are often set up by the researcher.

- **focus**: though "contextual inquiry" is consciously managed to serve as an open approach to data acquisition, it is unavoidable that the study focuses on certain aspects of the activity at the cost of others. Therefore, it is important to be aware of the current focus and also be capable off, during the inquiry, to shift focus towards different areas to acquire as rich as image as possible. An important role is played by surprises and contradictions (i.e. the user is doing something "wrong", unexpected or idiosyncratic). Such events usually signify that something is happening which is not completely understood, and which merits increased focus.

- **interpretation**: the observations need to be interpreted. The researcher does the interpretation in concert with the user. It is the responsibility of the researcher to produce results and interpretations, but these must be tested out and validated by the users.

In addition, my analysis makes to some extent use of a branch semiotics known as actor-network theory (Bijker et al., 1987). According to actor-network theory, technological artefacts and human actors are interlinked in a socio-technical web, and in this mesh, the actors' interests are expressed in technical and social arrangements. In addition to the traditional semiotic view that these interests can be expressed in texts, signs and symbols, actor-network theory also look for the signified in artefacts and in the social rules surrounding the use of these artefacts. Two important concepts in this context are *translation* and *inscription*. Translation is the process by which actors impose interpretation upon the network they are entangled in. On the outset, there may be many competing translations (i.e. interpretations), but through negotiations, mutual adjustments, clever adaptations and enrolling of allies, one particular translation becomes dominant and a certain degree of stability (known as an "aligned network") is reached. Such a stable translation is called an inscription. Another way of viewing an inscription is that it is the result of a particular translation affixed to a particular medium or material (Callon, 1991, p. 143).

2. Basic Concepts

The first Internet search engine was probably *archie* (Emtage and Deutsch, 1992). Archie first became operational in 1990 and was created to provide a searchable index to all the files that could be accessed using anonymous ftp over the Internet. In their original paper on *archie*, Emtage and Deutsch describe the "*resource discovery problem*":

> The huge size and continued rapid growth of the Internet offers a particular challenge to systems designers and service providers in this new environment. Before a user can effectively exploit any of the services offered by the Internet community or access information provided by such services, that user must be aware of both the existence of the service and the host or hosts on which it is available. Adequately addressing this "resource discovery problem" is a central challenge for both service providers and users wishing to capitalize on the possibilities of the Internet. (Emtage and Deutsch, 1992)

This quote still stand as a formulation of the basic research question to confront if one want to make useful the enormous (and growing) aggregate of information resources that exist on the Internet.

Leaving aside, for the time being, what a resource actually is, we move on to a concept known as *metadata*. Metadata is "data about data". Real life examples of metadata include such things as a library catalogue card (the "data" on the card describes the data contained in the books in the library) or a TV guide (the "data" in it describes the data in the programmes about to be broadcast).

Hence, metadata describes and qualifies other data. Typical examples of metadata are important properties of the data (e.g. the name of the creator and the publisher, the year of publication), information required to locate the resource (e.g. the Dewey-code for a library book, and the time and channel for a television program), and data that is helpful when searching for the resource (e.g. a free-text description or a summary of the data, or a list of searchable subject keywords appropriate for the data), but there are no hard and fast rules about what constitutes metadata.

The actual data described by the metadata is called a *resource*. Again, there are no hard and fast rules about what constitutes a resource. On the Internet, anything interesting that has an identity is considered a resource.

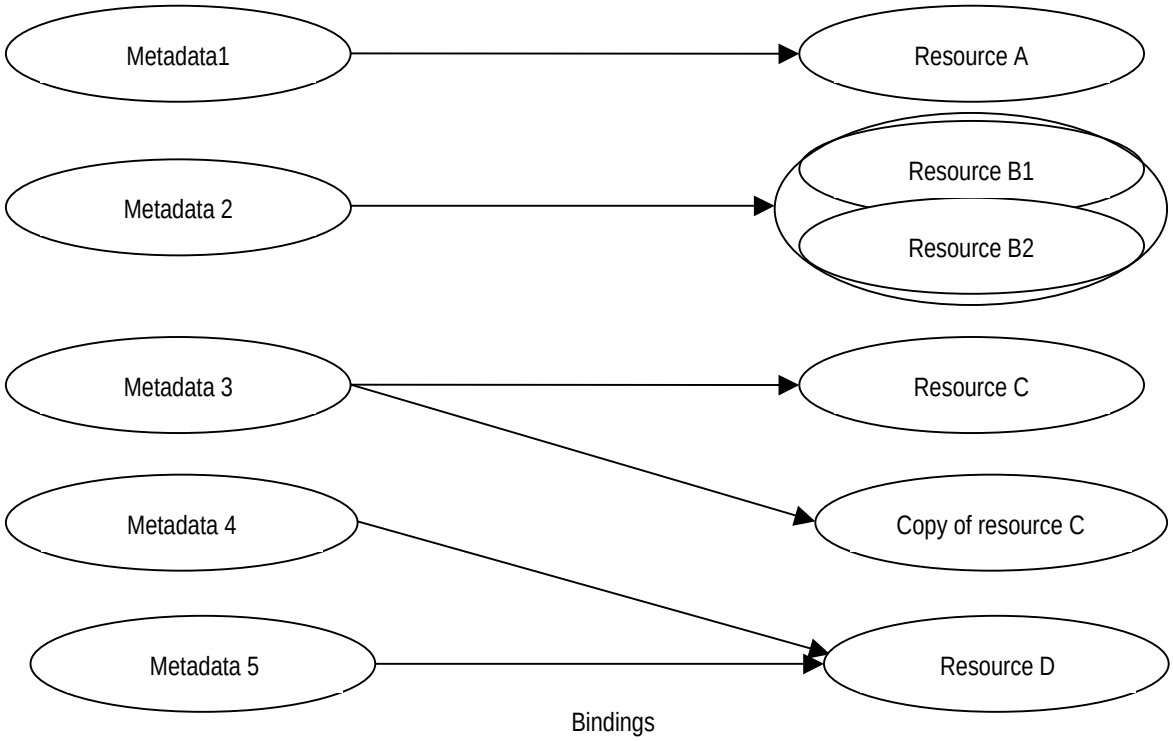
A resource should always be considered a separate entity from the its description. The two entities may be physically separated (e.g. a library catalogue card and the book it describes) or the metadata entity may be embedded in the resource itself (e.g. the information that usually appears on the title page verso in a printed book). The mechanism linking the two entities is called *binding*. In the physical world of a library, the binding may be a Dewey-number that points the person browsing the library catalogue to the location on the bookshelves where the actual book described by the library card may be physically located. On the Internet, binding may be facilitated by means of a Universal Resource Locator (URL) that identifies the location of the resource on the Internet.

This gives us the following basic definitions:

Resource: An identifiable object of interest that is accessible and available to the public.

Metadata: A machine-understandable set of properties that aggregates to some description of a resource.

Binding: The association between metadata and the resource it describes.



In a library, practical and physical constraints dictate that there is a single, directional relationship going from the card catalogue to the book, so there is a one-to-one relationship where a library card catalogue is bound to a single instance of a book.

While the binding model implied by all known Internet search portals seem to rely on a similar constraint, it is not difficult to imagine other arrangements, viz.:

1. There is a one-to-one relationship between the metadata and the resource. I.e. one metadata set is bound to one, and only one resource. This is how libraries classify their collections, and this is also the underlying model for metadata management in current Internet search portals.
2. A single metadata set is bound to a collection of closely related resources. This may be the case if we wanted to use a single set of metadata to describe an entire site or a closed collection of web pages such as a link farm.
3. A single metadata set may be bound to multiple instances of a replicated resource. A common practice on the Internet is “mirroring”, whereby popular resources are copied to several alternate locations. The reason for this redundancy is to provide multiple access points, and to provide backup access in case of network failures.
4. Finally, we may also want to use two or more metadata sets to describe a resource. This is the situation when the resource may be viewed from two different perspectives (e.g. a Java applet that can be studied by a programming student to learn about some specific feature of the Java language, and *also* is a fun game that may be viewed as a nice gaming resource by someone completely uninterested in Java).

   Like replication, this is a common occurrence in the Internet. This particular relation may be a property of the resource deliberately designed by the creator of the resource, but it may also be due to some serendipitous qualities of the resource.

3. Two Types of Information Search Services

Since the 1960s, *classic online information search services* such as Dialog, ESRIN and Lexis-Nexis have facilitated computerised resource discovery. These services let users locate resources published on traditional media (e.g. newspapers, periodicals and books).

In the present decade, the surge of popularity of the Internet, and the vast quantity of information resources presumed to exist on the Internet, has caused a new type of search service to emerge, directed towards resource discovery on the Internet. This new type of search service is often referred to as an *Internet search portal*. Examples of Internet search portals include such familiar brands to most Internet users as AltaVista, Deja, HotBot, InfoSeek and Yahoo.

One major (and obvious) difference between the former and the latter is that the classic online information search services are sold as a commercial service to professional users, while the Internet search portals are available at no cost to the general public.

3.1. Searches in Context

The classic online information search services were typically created to cater for the research related needs of professional researchers (i.e. people whose job is to discover information relevant for their own work, or relevant for the company they work for).

Dialog started its life as an internal service for the Lockheed aerospace corporation's library in 1965. In the early 1980ies Lockheed decided to make this service, Dialog available to external, paying clients.

Not only were this type of search system explicitly designed to function in a work related context, this was also how they are used. Since there is a charge for using these systems, access is controlled. Interviews with frequent users of this type of search system confirm that the dominant usage context is work.

While the majority of the persons I interviewed also reported that they used Internet search portals in the context of work, to discover information resources that they believed would be beneficial for them in their work, not all (or even the majority) of usage of Internet search portals appear to be work related. An analysis of the log of the popular AltaVista search engine conducted in the fall of 1998 yields the following top ten terms: *sex, applet, porno, mp3, chat, warez, yahoo, playboy, xxx, hotmail* (Silverstein et al., 1998). An alternative sampling in March 1999 (Blast Interactive Inc, 1999) yields an almost identical list.

4. Sampling Specific Search Services

I wanted to see how the characteristics of a classic information search service compared to a similar service created to specifically search the Internet. I did this by studying two services named *Atekst* and *Kvasir*.

Both services let users search for information resources through a (at least for search services) conventional user interface involving boolean operators. Both services presents the results to the user through a similar menu showing a ranked list of items matching the search criteria with title, date and (roughly) two paragraphs of descriptive

text. After being presented with the results, users can immediately retrieve the full text of any found item judged relevant by the user by typing in the corresponding item in the menu (Atekst) or just by clicking on it (Kvasir). What distinguishes these two services from each other, is how searchable data is captured, entered into the searchable data set, managed and maintained.

Also note that while classic search services mainly are used for work, and Internet search portals apparently are not, the present study focuses on work related searches for both types of system.

4.1. Atekst

Atekst is a classic online information search service. It started operating in 1987, originally contrived as an in-house information search service, but to offset costs it is also sold to outside clients on a subscription basis.

The Atekst data set consists of the complete text extracted from the daily editions of two major Norwegian newspapers, a full wire service feed, and the tables and summaries from an annual factbook named "*Hvem Hva Hvor*".

In March 1999, Atekst contained slightly more than 2 million articles. The internal format is plain text (i.e. no images or rich text).

All editorial text originating from the first three sources are entered into the Atekst database once daily by a semi-automatic process converting the typesetter files used for the production of the newspapers into a cleaner format (called TRIP) suitable for storage in the Atekst database. Data from the yearbook is entered once a year. The basic storage unit of the system is an "article", corresponding to an article in a newspaper or a yearbook article or table. A special function allows for extracting a smaller textual unit (fragment) if required. The system also recognises a larger unit that is called a "case". A "case" is defined as major news event on which a number of reports are likely to be filed over a period of time (e.g. the «Monica Lewinsky affair»). Currently, roughly 200 events have a «case» name assigned to them.

Special staff called «archivist» mediates the process of entering text into the database. All archivists have formal training as librarians, and most of them have worked in a library before joining Atekst. Their main responsibility is adding metatags to the articles labelling such properties as creator, title, date, and subject matter. The subject matter properties may be chosen from of controlled vocabulary of around 1000 keywords organised in two levels. Additional metatags delimits the item to a specific geographical area. Articles considered being part of a "case" are tagged with the appropriate "case" name.

The first two paragraphs of any article filed are by the default flagged as a searchable abstract. This happens automatically, but if the first two paragraphs do not contain material particularly well suited for this purpose, the archivist responsible for filing the article is expected to edit the text of the article to remedy this.

Other minor editorial changes to the material are also done at the archivists' discretion. For instance: Archivists sometimes split an article in two separate articles if it deals with two separate issues. Archivists may also catenate a number of very short notices dealing with the same subject matter into a single long article.

4.2. Kvasir

Kvasir is an Internet search portal that has been online and in use since 1995. It is part of the set of services provided to the public by a major public Internet website in Scandinavia (Sol).

Kvasirs primary purpose is to attract Internet users to the Sol website (Sol's primary income is from so-called page-views, which are random advertising exposed to a user every time a user visits the Sol website. Sol also extracts a secondary income selling focused advertising related to specific search terms. For example, a user searching for "books" would typically be exposed to banner advertising for online bookstores. This is because Sol sells advertising spots related to certain search words at a premium. In these, and in most other aspects, Kvasir operates in a manner similar to better known Internet search portals, such as AltaVista and HotBot.

A special program known as a «robot» (alternatively as a "scooter", "drone", "spider" or "crawler") facilitates data capture. A "round" starts by the robot going through its present collection of hyperlinks, visiting each site in the link collection turn and recursively following hyperlinks using a depth first algorithm. Link traversal terminates at nodes already visited within the present "round", and at hyperlinks pointing to domains outside the top level DNS domain called ".no" (this corresponds roughly to the set of Internet hosts physically located within Norway). The robot spends roughly three weeks on a "round". After completing a "round" it removes references to resources not visited at any point during the present or previous round, resets its internal state, and then immediately starts on a new "round".

All textual data (i.e. page contents excluding framed and inline elements) located by the robot on the World Wide Web are copied across the Internet and stored in the searchable data set on the search engine host computer. The process for cataloguing the data is automatic (i.e. without human intervention). The following basic properties and inline elements of a web page: title, URLs, inline image names, inline applet names, anchor text and visible text (i.e. all text data *except* the preceding), and the *Last Modified* timestamp, are recognised and tagged as part of the cataloguing process. Kvasir does not store the documents per se, but tagged properties and inverted plain text.

The basic searchable unit is a "document". A "document" corresponds to an individual web page and is identified with an URL (Uniform Resource Locator).

In March 1999, Kvasir contained searchable data from more than 9 million documents.

4.3. Concepts and Categories

Below is a discussion of the concepts and categories explored. The findings are mostly derived from working with staff and users of Atekst and Kvasir, as well as using and studying the actual services. Some findings are also extracted from relevant literature, and from a separate study of library work practices.

4.3.1. Hosted and Non-hosted Resources

The resources managed by Atekst are stored on the same computer system as is used to provide the information search service. When the user instruct the system to retrieve an

information resource after completing a search, it can be retrieved from the local system. This is what characterises a *hosted* system,

The resources managed by Kvasir are not, per se, stored on the Kvasir search portal host. Only the data set used to facilitate searches are. The actual resources remain on the Internet under the regime of their original publishers. This is called a *non-hosted* system.

This, in fact, seems to be the general pattern: Classic search systems are implemented as hosted systems; Internet search portals are implemented as non-hosted systems.

The cause of this dichotomy cannot be technical. It would be just as simple to create a hosted system as a non-hosted system for searching for resources published on the Internet – it would simply entail copying and normalising the actual resource to the search system host computer along with the searchable data set. The cause may be legal (copying the actual resource would violate the creator's copyright) or economical (as the resource can be presumed to exist on the Internet, there is no need to consume host computer storage space for it). It is beyond the scope of the present study to elucidate the actual cause (if any) behind this dichotomy, but, as we shall see, a number of other properties, in particular those related to persistence and synchronisation as discussed below, follows from this dichotomy.

4.3.2. Persistence and synchronisation

A major study of seven major Internet search portals conducted by Lawrence and Giles (Lawrence and Giles, 1998) in December 1997 found that these yielded between 1.6% and 5.3% invalid links (i.e. link requests that returned an HTTP error response code). Lawrence and Giles conjecture that the correlation between collection size and "*link rot*" incidence may be explained by the fact that search portals managing larger collections need longer times to complete one "round" resulting in less frequent updates.

A study of Kvasir conducted in March 1999 using the same approach as Lawrence and Giles yields 1.4% invalid links. While this is better than any of search portals studied by Lawrence and Giles, Kvasir also manage a smaller collection of documents. Kvasir manage 9 million documents, while the "best" performer in Lawrence and Giles study (ibid.), Lycos, yielded 1.6% invalid links and managed 10 million documents. The "worst" performer in this respect in Lawrence and Giles study (HotBot) yielded 5.3% invalid links and managed 109 million documents.

A related consequence of the non-hosted approach is *transients*. A transient is a web page that has had changes made to its content after the content has been copied into the searchable data set managed on the search portal host. Such changes may be minor incremental updates made to the page, or may constitute the replacement of one page with another, different, page on the same location (i.e. the URL is the same as the URL of the previous page). Transients may be due to version superimposition (i.e. a new and better version of replaces a previous version), or they may bear no relation or similarity to the previous page.

An increasing number of web pages are by design perpetually transient. This means that they are created on the fly, containing content extracted on demand from some up-to-date database. Such web pages carry, at least in principle, a different content each and every time they are viewed. Sampling the online editions of newspapers they all appear to use a web publishing system that works like this. This means that the online editions of these newspapers are not searchable in a meaningful the way through Kvasir

or any other search model relying on the non-hosted approach. Kvasir catalogues such ephemeral resources, but when searching for them, users end up with transients or invalid links.

Data entered into Atekst are never changed after they have been added to the system, and are never erased. This means that "link rot» as well as transients does not occur in Atekst. The same is the case with other hosted systems.

4.3.3. Genre

A genre has traditionally been used to classify literary works based on substance and form (Yates and Orlikowski, 1992). Substance refers to the themes and topics expressed in the work (the reporting of current affairs in "newsprint", or the staple mission statement of a "company profile"). Form refers to representational features of the work (the tabular format of a "price list" or the contact information of a company "home page").

Atekst only contains two genres: Newsprint and tables. This homogeneity makes it simple for the archivists to maintain consistency and apply keywords in a uniform manner.

A preliminary sampling of Kvasir revealed a very diverse range of genres, including raw data, software, personal CVs, company and product presentations, confessionals, fiction, poetry, pamphlets, advertising, as well scientific and pseudoscientific reports. Publications of lasting importance such as online novels, short lived data such as daily tv-guides, serious reports on major scientific developments, and the toilet humour resulting from some student fraternity's recent beer bash.

4.3.4. Data Types

In Atekst, all resources are stored as text. This simplifies management and retrieval, as there is no need for special software to display exotic data types. On the downside, the lack of images and typography also reduces the quality of the hosted resource compared to the original publication in the newspaper or in the yearly factbook.

In Kvasir, most resources are HTML pages or aggregates of HTML pages. There are also individual resources that either are, or inline one or more of the following: rich text, pdf, images, audio, video, flash, java, computer software and just raw data on a multitude of formats. Most of these are considered non-standard data types and are left out of the searchable data set. The *names* (but not the actual content) of inline images and applets are captured and are therefore searchable.

In Atekst, there is no compound data type. The "case" concept is a metadata property, not a data type.

Most web sites consist of more than one individual web page. While many web pages are self-contained and can be viewed in isolation, others do not. For instance, an online narrative constructed by means of a hyperlinked mesh where the individual scenes in the narrative are each on a separate web page is really a compound data type (narrative) aggregated from individual web pages.

4.3.5. Replication

All articles in Atekst are unique.

Kvasir contain a number of duplicates (or near duplicates) of a small number of resources. The duplicates fall in two broad categories: individual copies of most

requested resources (such as RFCs) put on the web by individuals, and so-called "mirrors" of popular and/or controversial sites where the full navigational structure and content of a web site is replicated.

As non-hosted resources are spread out to the literally millions of computers on the Internet, what appears to be the *same* resource is available from a number of *different* computers. A given resource may be "the original", an exact copy of the original, or even a slightly altered version – where the alterations may be of a beneficial, cosmetic or malicious nature.

Computer software is typically issued in versions. In the case of multiple copies of a software resource, users indicated that they are interested in knowing which versions (and which one is the newest).

4.3.6. Classification and vocabulary

In Atekst archivists tags each article with a set of subject keywords chosen from a controlled vocabulary of around 1000 keywords organised in two levels. Additional tags tie an article to a specific geographical area and/or a specific "case".

The Atekst system enforces consistent use of this vocabulary. The data entry software monitor the keywords entered by the archivists and flag as an error any keyword not contained in the approved list.

As part of the present study, keyword usage in a traditional library was also surveyed. When reviewing the history of keyword usage in the University of Oslo Library OPAC (Online Public Access Catalogue), these were found to mutate over time. For example: The keyword "*computer assisted*" that was used predominantly in entries created in the early 1990ies had in later years given way to the synonym keyword "*computer supported*" – probably reflecting changes in jargon in the relevant research community. Such shifts, however, creates problems for users of the search system. Users who are using keywords from one period will not locate relevant resources that are tagged with keywords specific to a different period. It is interesting to note that this particular OPAC had during the entire period been maintained by a single person, a skilled reference librarian well aware of the problems that entails from inconsistent keywords usage. When a single professional is unable to maintain consistency over a period of time, is it realistic to believe that keywords supplied by a heterogeneous group of Internet "publishers" will be of much use?

Most librarians/archivists interviewed expressed that while formal classification and controlled vocabularies played a major part in making the collections they themselves maintained manageable, they doubted that similar schemes would "work on the Internet". One librarian responded as follows when asked to imagine an Internet search service for medical resources making use of a controlled vocabulary such as MeSH (Medical Subject Headings):

> That would be a disaster. Imagine MeSH in the free for all chaos of the web! Every charlatan would use it in attempt to rub off some credibility for his or her methods or products, snowing out any legitimate medical resources on the web.

4.3.7. Agency

Agency is embodied capacity for action. While agency previously was considered to be an *attribute* of a free individual, actor-network theory maintains that agency is not an attribute but an *effect* of networks consisting of humans and artefacts.

Also following from actor-network theory is that as networks mature, they become more aligned, as their embodied agency wields it effect and makes their structure durable. This effect is clearly visible in the Atekst system, where the archivists, the data, the computer system and the conventions and work practices that has been negotiated over years of usage has created a robust system aligned towards the goal of quickly finding old newsprint articles in the Atekst repository.

The same level of alignment is not evident in the case of Internet search engines. The human actors in this particular network are the web publishers, the staff of the web search engines, and the users, the artefacts are the search engines and the robots performing data capture on the web on behalf of the search engines. What we can observe is an interesting set of translations, where some web publishers use an amazing range of tactics to enrol the artefacts (in particular the data capture robots) into acting in their interest:

> Louis Monier, AltaVista's technical director, estimates that half of the 20000 pages added to the search engine each day are schemes to boost Web site rankings. When Princess Diana was killed, Monier says, "people were adding the words 'Diana,' 'death,' and 'car crash' to their pages, even if they were totally unrelated." (Judge, 1997)

This finding is corroborated by the chief maintainer of Kvasir, who in March 1999 stated that around 50% of the URLs submitted to the service through the «add a link» facility made use of some sort of presumed hit-increasing deception.

Echoing the hapless hotel manager who in Latours classic paper on inscription puts up a sign with the grammatical imperative "Please leave your room keys at the front desk before you go out" (Latour, 1991, p. 104), the manager of the AltaVista search engine has posted the following notices on its web page:

> AltaVista is an index, not a repository for pages of low or misleading informational value. Attempts to fill it with misleading or promotional pages lower the value of the index for everyone. Left unchecked, this behavior would make Web indexes worthless. We will disallow URL submissions from those who spam the index. In extreme cases, we will exclude all their pages from the index. (Alta Vista, 1999)

Needless to say, this has about the same preventive effect on index spam as the sign put up by the hotel manager had on preventing hotel keys from being lost.

The next phase is of course technological innovation. Metatags was originally created to improve Internet search engine accuracy by letting web publishers label their pages more accurately. For this precise reason, they also permit more accurate mis-labeling of web pages. This has resulted in most search-engine sites discontinuing use of metatags, as explained by the managers of the Excite search engine:

> Unfortunately, metatag information is not always reliable. It may or may not accurately reflect the content of the site. In general, our [robot] does not honor metatags. This means that we do not index the content of the metatag. We will still index the body text of a page even if metatags are present. (Excite, 1999)

But index spammers have already come up with countermeasures to Excite's policy of not honouring metatags. Instead they embed misleading information directly into the text of web pages themselves. This is usually done with "invisible" text (e.g. white or very near to white text set on a white background).

In addition to index spam, some concerns have been raised about the integrity of information resources located through the Internet. In addition to such extreme pheno-

menon as Holocaust revisionists and other extremist groups using the Internet as a medium to disseminate alternative “facts”, medical and legal professionals have referred to uncritical use of the Internet by the public as a source for medical and legal resources as a problem (Eysenbach and Diepgen, 1998).

Users of Internet search portals observed and interviewed did, in general, not share this concern. Most users responded that they were aware of the problem, but felt that they were capable of making the appropriate judgements when confronted with the actual resource. One user noted, however, that in questions where the integrity of the source was vital (in his case: the official stance of the Roman Catholic Church on certain issues), he did not rely on resources located through Internet search portals, but used only the navigation structure provided on the official pages of the Roman Catholic Church.

4.3.8. Cardinality

Both Atekst and Kvasir currently associate each resource with one and only one set of metadata.

This seems to be reasonable for Atekst, as Atekst contains a number of alternate provisions that compensate for lack of flexibility in this particular area. I.e.: There are no duplicates, the “case” property provides an alternate mechanism for creating collections, and archivists splitting or catenating articles removes the need for creating one-to-many or many-to-one relationships.

On the Internet, the situation is more complicated; because copies or near copies of resources may exist at different locations. When performing a search with *archie* (Emtage and Deutsch 1992), a search for a particular resource will often yield dozens of copies or near copies of the same resource and a result set will therefore contain little information.

On the Internet, this sort of redundancy is not “planned” or “designed” in the traditional sense. It simply arises through uncoordinated actions from different users and groups. Therefore, there is no point in having the means to express this particular relationship in the external schema available to the creator. Instead, we should have the means to *discover* replication, and an internal schema that lets us express what we have discovered.

The inverse situation: that a single resource may merit more than one metadata description because it is a container for several orthogonal entities that have different, but equal merit depending upon perspective also arises sometimes.

From this analysis follows that systems to search the Internet should be capable of handling bindings between metadata and resources of higher levels of cardinality

5. Conclusion

Comparing and contrasting capture, management and maintenance of data in a classic hosted information search system with an Internet search portal yields a number of important differences, summarised in the table below.

#	**Concept**	**Atekst**	**Kvasir**
1.	Main Usage	Work related	Non-work related
2.	Resource	Hosted	Non-hosted
3.	Persistence	Persistent	Ephemeral
4.	Genre	Newsprint	Diverse
5.	Data Type	Text, Tables	Diverse
6.	Replication	None	Some
7.	Vocabulary	Controlled	Chaotic
8.	Agency	Aligned network	Non-aligned network
9.	Cardinality	One-to-one	Higher

Table 1: *Atekst and Kvasir, summary of characteristics*

The conceptual differences between searching a hosted resource and locating a non-hosted resource are profound. Current Internet search portal designs have, at least to some extent, failed to recognise these differences, which may partly explain their less than satisfactory performance.

This, however, do only lead to further research questions:

For instance: How can Internet search engines deal with index spam and with malicious actors who plant resources that have some undocumented side effect (ranging from pranks, via virus spreading, to sabotage).

Also, as demonstrated, Internet resource data types are often complex and poorly labelled. No current Internet search portal attempt to deal with this complexity. This results in a failure to capture content represented by non-standard data-types. It also results in failure to represent and present compound data types. How, beside the obvious, but un-practical, remedy of requiring standard, explicit and truthful labelling of datatypes, can Internet search engines capture this concept?

For designers of search systems to adequately address the challenges to arising from this study is probably a large and non-trivial task. I have resisted the temptation to make specific design recommendations in the present paper, as such recommendations need to implemented and tested out in real usage to become interesting.

I hope, however that the concepts for Internet research discovery outlined in the present paper is a useful point of departure for designing more useful Internet search services.

6. References

Alta Vista (1999) *No SPAM Please*, viewed: *1999-03-01,* <http://www.altavista.com/av/content/addurl.htm>.

Beyer, H. and Holtzblatt, K. (1998) *Contextual Design: Defining Customer-Centered Systems,* Morgan Kaufmann Publishers, Inc., San Francisco.

Bijker, W. E., Hughes, T. P. and Pinch, T. J. (Eds.) (1987) *The Social Construction of Technological Systems. New Directions in the Sociology and History of Technology,* MIT Press, Cambridge, Massachusetts.

Blast Interactive Inc (1999) *Most Popular 100 Internet Search Words*, viewed: *1999-03-05,* <http://www.searchwords.com/>.

Callon, M. (1991) Techno-Economic Networks and Irreversibility, In: *A Sociology of Monsters. Essays on Power, Technology and Domination.*, (Ed, Law, J.), Routledge, London, pp. 132-161.

Dehn, D. M. and van Mulken, Susanne (2000) The impact of animated interface agents: a review of empirical research, *International Journal of Human-Computer Studies*, Vol. 52, pp. 1-22.
Easterby-Smith, M., Thorpe, R. and Lowe, A. (1991) *Management Research. An Introduction*, SAGE Publications, London.
Emtage, A. and Deutsch, P. (1992) Archie - an Electronic Directory Service for the Internet, In: *USENIX Winter 1992 Technical Conference*, USENIX Association, San Francisco, pp. 93-110.
Excite (1999) *Understand Meta Tags*, viewed: *1999-03-01*, <http://www.excite.com/Info/listing8.html>.
Eysenbach, G. and Diepgen, T. L. (1998) Towards quality management of medical information on the internet: evaluation, labelling, and filtering of information, *BMJ*, Vol. 317, Nov. 28, pp. 1496-1502, URL: <http://www.bmj.com/cgi/content/full/317/7171/1496>.
Glaser, B. G. and Strauss, A. L. (1967) *The discovery of grounded theory; strategies for qualitative research*, Aldine de Gruyter, New York.
Judge, P. C. (1997) Revenge of the Search Engine, *Business Week*, November 17.
Latour, B. (1991) Technology is society made durable, In: *A Sociology of Monsters. Essays on Power, Technology and Domination.*, (Ed, Law, J.), Routledge, London, pp. 103-131.
Lawrence, S. and Giles, C. L. (1998) Searching the World Wide Web, *Science*, Vol. 280, April 3, pp. 98-100.
Lawrence, S. and Giles, C. L. (1999) Accessibility of information on the web, *Nature*, Vol. 400, July 8, pp. 107-109.
Lesk, M. (1989) What To Dp When There's Too Much Information, In: *Hypertext '89*, ACM, pp. 305-318.
Lynch, C. (1997) Searching the Internet, *Scientific American*, Vol. 276:3, March, pp. 44-48.
Nielsen, J. (1995) *Multimedia and Hypertext : the Internet and beyond*, AP Professional, Boston.
Nielsen, J. (2000) *Designing Web Usability: The Practice of Simplicity*, New Riders Publishing, Indianapolis.
Page, L., et al. (1998) *The PageRank Citation Ranking: Bringing Order to the Web*, Draft , January 29, Stanford University, Palo Alto.
Rosenfeld, L. and Morville, P. (1998) *Information architecture for the World Wide Web*, O'Reilly, Sebastopol, CA.
Silverstein, C., et al. (1998) *Analysis of a Very Large AltaVista Query Log*, SRC Technical Note 1998-014, October 26, Digital Systems Research Center, Palo Alto, CA.
Spool, J. M., et al. (1999) *Web Site Usability: A Designer's Guide*, Morgan Kaufmann Publishers, San Francisco.
Yates, J. and Orlikowski, W. J. (1992) Genres of Organizational Communication: A Structural Approach to Studying Communication and Media, *Academy of Management Review*, Vol. 17:2, April, pp. 299-326.