General Help Systems for Network Resource Discovery

Clarisse Sieckenius de Souza

Departamento de Informática — PUC-Rio
Rua Marquês de São Vicente 225
Rio de Janeiro, RJ — Brazil
clarisse@inf.puc-rio.br

Center for the Study of Language and Information
Stanford University
Stanford, Ca. — USA
desouza@csli.stanford.edu

Monografias em Ciência da Computação

Departamento de Informática — PUC-Rio

MCC 24/93

Abstract

Wide area computer networks offer users a large volume and variety of resources that can help them solve many kinds of daily problems in their work environment. However, the existence, let alone the contribution, of such resources is generally ignored by networkers. This represents a gap in the usability of items available in the network. This work discusses usability issues as perceived in the INTERNET, and presents a research project underway, relative to General Help Systems (GHS's) — an intended means to narrow the distance between users needs and network resources.

Keywords: network resource discovery systems, indexing strategies, resource usability, collective knowledge.

As grandes redes de computadores oferecem a seus usuários considerável variedade e volume de recursos, capazes de auxiliá-los a resolver muitos tipos de problemas cotidianos encontrados no trabalho. Contudo, a existência, e muito mais a contribuição, destes recursos é geralmente ignorada pelos participantes das redes. Isto representa uma lacuna quanto à utilidade dos itens disponíveis. O presente trabalho discute questões de utilidade observadas na INTERNET e introduz um projeto de pesquisa em andamento sobre os *General Help Systems* (GHS's) — um pretendido meio de estreitar a distância entre as necessidades dos usuários e os recursos da rede.

Information Usability

Empirical research about user acceptance of information technology [Davis'93] has shown that actual system use is much more influenced by its **perceived usefulness** than by the user's attitude towards it or the system's perceived ease of use. Wide area computer networks such as the Internet offer a **huge volume and variety of resources** to users. One of the major features of Internet Resources is that they are constantly changing — new items are added, existing items are changed or removed. Moreover, new items are not necessarily of the same kind as previous ones. Anything that can be expressed in computable form is a candidate item. Thus, the limits of item variety in the Internet are the limits of computable matter itself.

One of the greatest challenges faced by Network Resource Discovery Systems (NRDS's) designers is to make users perceive the **usefulness of resources**. The most popular NRDS's — Archie, WAIS, Gopher and WWW [LaQuey & Ryer'93] — allow users to navigate throughout an incredible wealth of resources. Their goal is to support access to and retrieval of all available items, which range from people to software, from on-line databases to publications, from music to weather reports, and much more. However, there is a major gap between the current installed technological capacity and its actual use by people in daily breakdown situation. This is the **network resource usability gap**.

As stated in a recent introductory book to the Internet [Lane & Summerhill'93], "what keeps many potential network users from 'plugging in' is not necessarily a lack of opportunities, but rather a lack of *knowledge* of those opportunities." The challenge designers must face regarding these potential users is to **show them how useful Internet resources are**.

Many networking guides have been written to bridge this detected knowledge gap [LaQuey & Ryer'93, Kehoe'92, Lane & Summerhill'93]. However, these are "off-line" information. When a breakdown occurs, these guides are not likely to provide users with timely and contextualized helpful advice. This embedded information retrieval facility is of the sort commonly available in on-line help systems [Horton'90]. If existing NRDS's included an **on-line contextualized help facility** the resource usability gap could be more easily bridged.

The Issue

Text editors, programming language compilers and interpreters, hypertext authoring systems and many other types of applications currently provide on-line contextualized help facilities. They do so by matching the conditions in which users ask for help and the embedded model of the software functionality [Horton'90]. It must be emphasized that help information is provided about the application program, not about data. It increases the user's perception of the program's usefulness.

The situation is not quite the same with network resources. Here it is not only a matter of increasing NRDS's perceived usefulness (which can be done with an on-line help facility they all include), but rather of increasing the perceived usefulness of data! As Lane & Summerhill put it [op.cit.], this is a matter of increasing the user's knowledge about resources.

Can on-line contextualized help facilities handle information about information?

The issue is one of building a **model** of the network: not of its physical connections, paths and gateways, but of the types of items it contains in relation to the types of situations in which people are likely to need them. In short, a taxonomy of resources and a taxonomy of needs are apparently called for. Contextualized help could be derived from **mapping needs onto resources,** and this should help users perceive the positive impact of the available items on their daily activities.

The Barriers

The problem with taxonomies is that they are static — structures freeze relevant abstractions derived from reality into a set of elements and a set of relationships between them. The dynamic aspects of reality, in computer models, are captured by rules that manipulate structures and transform them according to certain conditions and principles. A change in the way abstractions are made cannot be supported from within a computer model. For this, a meta-model is needed, where the conditions and principles to create abstractions from reality are clearly stated.

Thus, in the case of network resources, if the current state of affairs is captured by a model M, **the only kind of growth supported by this model is one in which the same abstractions are used**. For instance, if the structure of resources reflects abstractions having to do with computer sites, file types, email addresses and the like, all new items must be classified according to these. An item whose real life abstraction had to do with hardware, basically, like the type of terminal the user has, could not be accommodated within the existing model of the network.

Since resources can be anything representable in **digital form,** some existing taxonomies used by NRDS's (e.g. Archie [Emtage & Deutsch'92]) choose wide abstraction levels like file names (and extensions) and FTP sites. The informativeness of such abstractions, however, is relatively low to support mapping from needs onto resources. For example, if a user needed to find a recent document about help systems for NRDS's, how would s/he proceed? Search for documents with the name "help"?

Another approach is to narrow down the width of the above schema and to represent **content**. There are two ways of dealing with content. One is to create a profile of items based on the rate of occurrence of certain words in its external (name and extension) and internal (text or code) form. An automatic indexing of items generates profiles that can support "content navigation" with relevance feedback [Kahle'88]. This is the strategy adopted by WAIS [Kahle & Medlar'91]. Another approach is to created semantic links between items, following the hypertext paradigm. This is the strategy followed by WWW [Berners-Lee et al.'92].

Form-based taxonomies — either external (like in Archie) or internal (like in WAIS) — **accommodate the evolutionary aspect of network resources** in different ways. The Archie approach sees growth as a structurally stable process: any new item is bound to conform to certain standards of digital representation. The WAIS approach handles growth by means of a stable screening process: the identification of relevant recurring string patterns in the item's content. Again, since all items are bound to conform to existing standards of naming (external form) and codification (internal form), new items are likely to be successfully classified.

Content-based taxonomies relying on semantic links between items face more obstacles as far as evolution is concerned. First, if any new item is to be classified according to an existing taxonomy of semantic links, these cannot be but some sort of conceptual primitive link. And as the network can contain anything human intelligence is able to conceive of and put into digital form, building such a taxonomy of links amounts to building a model that approximates that of human intelligence itself. Second, if the links do not have to belong to set of conceptual primitive relationships between resource items, and can be freely assigned by users as a result of an interpretation process, then there actually isn't any taxonomy in the pure (structural) sense, but only an open set of interpretations about the potential associative links the various resources maintain with each other.

Consequently, an on-line contextualized help system to support searches for information in wide computer networks would have to include a model of resources based on one (or more) of three taxonomies: **(a) one based on external form only, (b) one based on internal and external form, and (c) one based on primitive conceptual relationships holding between items.**

However, the above taxonomies solve only part of the original problem. The entire model supporting the help system envisaged for NRDS's still needs a taxonomy of needs, or breakdown situations, and a mapping from needs onto resources.

In that which concerns a **taxonomy of needs**, the underlying structure is derived from social activities involved in problem-solving situations. Recent work [Rogers'92] exploring the concept of distributed problem-solving for purposes of Computer Supported Collaborative Work (CSCW) reports that there are five major theoretical frameworks from which structures are drawn: **activity theory, conversational analysis, distributed cognition, anthropological theories of culture, and organization/community memory**. The author points out that each time a breakdown situation occurs, however, a very complex pattern of social interaction emerges to solve the problem at hand, which resists "being transformed into a crystallized set of working procedures" [Rogers'92, p.353].

In the same line as the comments made about content-based resource structuring, what can be said here is that either some of the existing theoretical frameworks claim to have achieved a set of primitive activities performed by people in a social context, or this set is open to human interpretation and can freely incorporate new needs as new situations arise. Recent work about the usability of technological tools [Adler & Winograd'92] suggests the latter to be the case. According to the authors, "the effectiveness of a system depends on how well it supports higher-order [interpretation] cognitive activities and social interaction" [op.cit.,p.5].

So, a form- or primitive-based taxonomy of needs, analogous to the one suggested for resources, seems to override an important factor having to do with the usability of the resource discovery system itself. From the moment problem situations are frozen within structural patterns, design choices concerning NRDS's seem to fall back into a plan-based model of human activity. This, however, has been extensively rebutted by evidences from ethnomethodological studies [Suchman'87].

<u>A Solution</u>

In actual problem-solving situations, people are seen as the prime source of task-relevant information [Berlin & Jeffries'92]. This is certainly due to the fact that people can interpret the global context in which problems occur and make pertinent contributions to the needed solution. Moreover, relevant contributions are made possible by people's awareness of what the others are doing [Dourish & Bellotti'92].

The perceived usefulness of resources in a wide computer network is directly related to the **timeliness and context-sensitiveness of retrieval conditions**. These two features are what makes people's help preferable to on-line help facilities. However, people cannot be found or interrupted whenever somebody needs their advice. **Asynchronous communication** is a way around physical or social barriers in getting contextualized information to solve problems. Nevertheless, timeliness is often a critical factor. The storage of collective knowledge in a community database is an alternative for accessing what people know. The complexity of this knowledge collection can vary widely, ranging from a set of annotations about items to a full fledged knowledge base.

annotation to: (<site>) /path(/file.name) or email address

made by: <user's email address>

Item Use:

Talk	Read
Interlocutor: ( ) human ( ) computer Communication Language: ( ) NL ( ) pseudo-NL ( ) specialized ( ) icons/menus	Message Coding: ( ) NL text ( ) Computer Code

Comments: [free text]

Figure 1: The GHS Annotation Structure

In order to get around the tension raised from the contrast between the intrinsic difficulties to accommodate novelty presented by abstract system models, on one side, and the evolving nature of the amount and kind of resources available in wide computer networks, on the other, we propose the incorporation of **Generalized Help Systems (GHS's)** to what LaQuey & Ryer have named "the Internet Toolbox" — email, remote login and file transfer protocol facilities [LaQuey & Ryer'93, p.23].

A GHS is an **annotations database** whose records present both **structured and unstructured information**. The structured part of a GHS record (Figure 1) incorporates a taxonomy of resources, reflecting both form-based and content-primitive-based structuring principles. **Data, programs and people** are the basic categories of available network resources. What they can do is closely related to whether they are active (**agent)** objects, passive (**file**) objects, or both. Moreover, agents are essentially message senders and interpreters (i.e. they communicate), whereas files are essentially messages themselves (i.e. they are a codified text to be read by human or computer processors). The way resources can be interpreted, then, is determined by the domain in computing where they belong: **communication**, **codification**, or both. [See Table 1]

The above considerations are crucial for the GHS approach, because the annotations database is organized according to the notion of **communicative indexing**, an indexing technique based on the **collective discourse structure** users can make about how resources may meet ends, in a totally contingent way [de Souza'93]. The principle that operates in communicative indexing is that resources — people, programs and data — may be used to an open variety of purposes, according to their intrinsic ontological nature — agents, files, or both — and their inherent interpretive possibilities — message senders and receivers, or messages themselves. Agents are objects one can **talk to**, whereas files are objects one (people or computers) can **read** (or view). If one is communicating with resources, it may be important to discriminate whether resources are human or computer entities, and in which language such communication takes place. Similarly, if resources are to be read (or viewed), it may be important to discriminate the coding of the message they represent, for interpretive purposes.

Resource Class	DP Ontological Nature	Domain in Computing
people	agents	comunication
programs	agents/files	communication/codification
data	files	codification

Table 1: A Taxonomy of Network Resources

The unstructured part of a GHS database record (See "Comments" field in Figure 1, against grey background) does not derive from a taxonomy, but is **a statement about the conditions of need** encountered or anticipated by users with respect to the resource. The mapping between resource type and conditions of need is performed by human beings, and not by systems, when they generate or retrieve annotations about resources. The retrieval process is a traditional database query as far as the structured part is concerned. In the unstructured part, queries are performed by keyword search. The combination of the two parts provides the basis for a collection of contextualized information about resources in a wide computer network (**meta-information**), inaccessible to the users of most NRDS's.

An example of a possible GHS query can be seen in Figure 2. There, a user is searching the GHS database for annotations to any file or agent (uspecified "annotation to" field), made by anyone (unspecified "made by" field), provided that the resource can be "read" and consists of "computer code" (i.e. a piece of software). The most relevant part of the query as far as usability goes is that, in the "comments" field where users state how resources connect to need situations, the following words can be found: "design documentation".

Possible annotations retrieved by this search schema are what is pictured in Figures 3 and 4. By reading the "comments" field, we can see the wealth of information retrieved based on **communicative indexinf** of a collectively built network usability model.

SEARCH	annotation to: (<site>) /path(/file.name) or email address
	made by: <user's email address>

SELECT Item Use:

Talk	Read
Interlocutor: ( ) human ( ) computer Communication Language: ( ) NL ( ) pseudo-NL ( ) specialized ( ) icons/menus	Message Coding: ( ) NL text (*) Computer Code

SEARCH FOR THE FOLLOWING KEYWORDS IN COMMENTS:
[DESIGN DOCUMENTATION]

Figure 2: An example of a GHS Search

annotation to: this.site /pub/design/soft/DesignDoc

made by: <user's email address>

Item Use:

Talk	Read
Interlocutor: ( ) human ( ) computer Communication Language: ( ) NL ( ) pseudo-NL ( ) specialized ( ) icons/menus	Message Coding: ( ) NL text (*) Computer Code

Comments:
[DesignDoc is a good design documentation tool for CAD applications. The spectrum of things you can document, though, is somewhat narrow.]

Figure 3 Annotation retrieved by a search about Design Documentation

annotation to: other.site /pub/design/WriteDesign.hqx	
made by: <user's email address>	
Item Use:	
Talk Interlocutor: ( ) human ( ) computer Communication Language: ( ) NL ( ) pseudo-NL ( ) specialized ( ) icons/menus	Read Message Coding: ( ) NL text (*) Computer Code
Comments: [WriteDesign is a new documentation tool developed by a research team of Newstuff,Inc. Developers want it to be tested by as many users as possible and welcome suggestions and bug notifications. WriteDesign works well for parametric design models, but doesn't help with creative design documentation.]	

Figure 4: Annotation about a another Design Documentation tool

The GHS Project has its roots in CSLI's *People, Computers and Design* Project, at Stanford University. There, the fundamental ideas about **General Help Systems** [de Souza'92] and **communicative indexing** [de Souza'93] have been discussed with participants of the PCD Project. Currently, most tasks in the GHS Project are being carried out at the Computer Science Department of PUC-Rio. At this site, the author coordinates a small team of collaborators who are building a prototype GHS environment to demonstrate the fundamental issues arising in the research project. Clearly, two types of goals are involved in the investigation. There are theoretical goals and practical goals. For purposes of instantiation, the INTERNET will be the prototypical network for the research.

3.1. Theoretical Goals

The first theoretical goal in the project is to investigate mechanisms of **meaning assignment** for the whole variety of resources in the INTERNET: application programs, files, database systems, discussion groups, electronic mail and others. Such mechanisms must be constrained by two relevant factors: (1) the fact that the INTERNET can accommodate any item in computer format, and item contents may be as varied as human imagination allows; (2) meanings must support retrieval, introduction and updating of items; and (3) meanings must support some sort of reference to specific items in natural language structured, semi-structured, or non-structured, text (i.e. one person must be able to refer to a particular item of the INTERNET when writing a note about it).

The second theoretical goal in the project is to investigate how **collective knowledge** is built and used in computer network environments. This goal explains the reason for constraint number (3), above. We assume that the most common means of communicating knowledge is language. Therefore, unless items can be spoken or written about, the sharing of knowledge among individual users is not possible in our scenario.

The expected outcomes of the above are:

a) a structure of meaning for all INTERNET items

b) a structure of knowledge exchanges among INTERNET users

3.2. Practical Goals

The first practical goal in the project is to build a prototype version of a GHS for a small community of users. This prototype will function as a testbed for the theoretical results concerning meaning assignment and collective knowledge. The prototype should focus on data structures and procedures of the GHS, without further commitments with the form of the interface: contents are the goal.

The second entailed practical goal is to build the complete interface system, with form and content. Here, issues related to interface layout and expressiveness, communication language and the like are in focus. As for the first prototype, the complete system version will serve a small community of users.

The expected outcomes of the above are:

a) an adequate architecture for GHS

b) a useful application system for the targeted community

c) an appropriate environment to support the design of GHS's for larger communities

The GHS Project tries to find one solution for a critical problem faced by large computer network systems. The assumptions made in the project, especially those concerning the use of natural language text in notes users will write about items in order to share their knowledge with other users, have a high chance of leading to community-dependent results. Such dependence may be of a linguistic nature or of a cultural one. Only comparative work can reveal the extent of this phenomenon. Thus, one parallel project can be carried out with a different community of users, preferrably **users with a different native language**.

There are a number of INTERNET exploration systems available at this time, like WAIS, GOPHER (McCahill'92], ARCHIE and WWW, for example. These systems do not have an embedded GHS approach and users experiment many and varied breakdown situations. Another interesting parallel project is to find out an appropriate **analytical framework** in which WAIS, GOPHER, ARCHIE and WWW could be compared in terms of the **kinds of breakdowns** users can face. The same analysis could be then made with the prototype GHS to explicit what relative merits each system has over the others.

Acknowledgements:

The author would like to thank CSLI, at Stanford University, for continuing support of this research initiative. Special thanks go to Terry Winograd, coordinator of the PCD Project, at CSLI, and to Noemi Rodriguez, collaborator of GHS, at PUC-Rio.

6. References

Adler,P. and Winograd,T. *Usability — Turning Technologies into Tools*. New York. Oxford University Press. 1922

Benners-Lee,T.; Cailliau,R.; Groff,J.; and Pollermann,B. World-Wide Web: The Information Universe. **Electronic Networking: Research, Applications and Policy, 2**(1), pp. 52-58.

Berlin,L.M. and Jeffries,R. Consultants and Apprentices: Observations about Learning and Collaborative Problem-Solving. *Proceedings of CSCW'92*. New York. ACM Press. 1992

Davis,F.F. User Acceptance of Information Technology: System Characteristics, User Perceptions and Behavioral Impacts. **International Journal of Man-Machine Studies (1993) 38**, pp. 475-487

de Souza,C.S. *Communicative Indexing fin Network Resource Discovery Systems*. To appear as a CSLI-PCD Technical Report. CSLI - Stanford University.

de Souza,C.S. *Augmenting Informativeness and Learnability of Items in Large Computer Networks*. Report No. CSLI-92-164-PCD-4. ISSN 1042-1173. May 1992. CSLI - Stanford University.

Dourish,P. and Bellotti,V. Awareness and Coordination in Shared Workspaces. *Proceedings of CSCW'92*. New York. ACM Press. 1992

Emtage,A. and Deutsch,P. Archie — An Electronic Directory Service for the Internet. *Proceeding of USENIX Winter Conference*, pp. 93-110, January 1992.

Horton,W.K. *Designing and writing online documentation* . New York. Jonh Wiley & Sons.1990.

Kahle,B. *Wide Area Information Servers Concepts*. Thinking Machines Technical Report TCM 202. November, 1988.

Kahle,B. and Medlar,A. An Information System for Corporate Users: Wide Area Information Servers. **ConneXions — The Interoperability Report, 5**(11), pp. 2-9, Interop,Inc. November, 1991.

Kehoe,B. *Zen and the Art of the Internet: a Beginners Guide to the Internet*. Chester. Widener University. January 1992.

Lane,E. and Summerhill,C. *Internet Primer for Information Professionals*. London. Meckler. 1993

LaQuey,T. and Ryer,J.C. *The Internet Companion*. Reading,Ma. Addison-Wesley Publishing Co. 1993

McCahill,M. The Internet Gopher: A Distributed Server Information System. **ConneXions — The Interoperability Report, 6**(7), pp. 10-14, Interop,Inc. July,1992.

Rogers,Y. Ghosts in the Network: Distributed Troubleshooting in a Shared Working Environment. *Proceedings of CSCW'92*. New York. ACM Press. 1992

Suchman,L.A. *Plans and Situated Actions*. Cambridge. Cambridge University Press. 1987