Wild Beasts and Unapproachable Bogs

Chris Weider
clw@bunyip.com

Abstract

The current suite of Internet information tools only allow their users to reach specific subsets of the information available on the Internet, and require information providers and consumers to interact with data through specific paradigms. This has limited our ability to present and use information, and in fact seem to be little more than extensions of the text based world we have grown up with. There are many untapped possibilities in our interaction with data, but much work needs to be done to provide a framework on which new tools can be deployed. This paper examines the current tools, the infrastructure required to make the current tools work together, and suggests some new techniques for human interaction with data.

I. Surveying Tools

As geographers crowd into the edges of their maps parts of the world which they do not know about, adding notes in the margin to the effect that beyond this lies nothing but sandy deserts full of wild beasts, and unapproachable bogs.

Plutarch

Despite the recent increase in the number of information tools deployed on the Internet, they all use one of two techniques for disseminating their information. The 'Send it everywhere' technique, used by tools such as electronic mail and USENet News, allows the information provider to replicate her message, at her discretion, to other individuals. She in essence is 'pushing' the data out to the consumer. The 'Come get it' technique, used by most of the other major information tools, allows the information provider to create a single copy of the resource and have the consumer retrieve it at their need. The consumer 'pulls' the data from a repository somewhere else on the Internet.

These two techniques are really endpoints of a continuous spectrum, with a tool such as an MBone multicast representing the extreme of the 'Send it everywhere' point, and with perhaps FTP archives representing the other extreme. There are a number of factors which determine which tool or combination of tools are best for publishing a specific set of information.

The first factor is that the 'send it everywhere' tools carry with them an unstated assumption of immediate delivery. When someone posts a USENet News article, or broadcasts a CU-SeeMe video stream, they assume that the information consumer will be able to use the information (almost) immediately as they will have a copy 'close' to their host machine. The 'come get it' tools, on the other hand, carry an implicit assumption that the information is provided for some future audience. and that the consumer has control over when the information is copied to a local machine.

The second factor is that the 'send it everywhere' tools have an unfortunate side effect that the information provider usually has no indication that the information has been consumed. For example, although there have been several attempts in the last year or so to build 'notification of delivery' and 'notification of access' into the various electronic mail protocols, many people feel that these extensions are an invasion of privacy. 'Come get it', on the other hand, allows extensive logging of consumption.

'Send it everywhere' tools will continue to be developed, but will probably continue to be primarily used for human - human direct communication. 'Come get it' tools, however, will probably continue to get more attention, because they allow more control over the information, its presentation, and access to the information.

This taxonomy of information tools has a number of important consequences for the future design of tools. Long-haul bandwidth is always going to be a scarcer resource than local bandwidth, because a) the cost of the actual transmission media (fiber optic cable, or whatever) is length-sensitive, so a 1000 km fiber is more expensive than a 1000 m fiber, and because b) a larger number of individual data streams need to be combined on the long-haul connection. Thus, there is an interesting balance between the strict 'come get it' model, in which a single copy of a resource is placed on the Internet, and the 'send it everywhere' model, which attempts to replicate the resource as close to the eventual consumer as possible. Many groups of consumers, from universities with small connections to the Internet, to countries such as Australia or Japan which must pay a premium for bandwidth to the rest of the Internet because the fibers must traverse water, need some sort of local caching mechanism to allow their users rapid access to resources and to reduce the load on their congested links out. But the necessary engineering to accomplish this local caching is just starting to be developed. While a document cache (for example) is trivially easy to create, allowing other users to access the cache, and (most importantly) determine if a desired document is already in the cache, requires a much more complete and sophisticated Internet information architecture, which we will explore in the next section.

Thus, while current information tools seem to have been developed as though they were strictly 'send it everywhere' tools or strictly 'come get it'

tools, the tools of the future will need to incorporate features of both if they are to provide good service to the users.

II. Building the maps

One of the most frustrating things about using the various information tools is that each one allows access to only a small subset of the available information on the Internet. Although clients like Mosaic allow the user the illusion that they can reach everything, search and navigation tools for the information bases available through each protocol are strictly protocol specific, which tends to shatter the illusion pretty rapidly.

In [1], Peter Deutsch and I outlined a preliminary architecture for building a map of the Internet. One of the major components of this preliminary architecture was to provide a consistent set of landmarks so that information tools could consistently refer to the same points on the Internet landscape. These landmarks will consist of the Uniform Resource Locators (URLs) [2], already familiar to many Internet users because of their extensive deployment in the WorldWideWeb, and of the Uniform Resource Names (URNs) [3]. Once this set of landmarks is in place, any information tool can refer to any resource, anywhere on the Internet.

The basic layout of this map is quite simple. The URL contains access and retrieval information for a given resource. However, since the resource may move around quite rapidly, the URL is brittle and is easily broken without the user's notice. The URN, on the other hand, is intended to provide a persistent, location independent reference to a resource. One analogy to the text world is that the URN is essentially an ISBN number; a valid reference to a book no matter where it happens to sit on the library shelves. The URL(s) associated with a given URN would then tell you where the resource was currently located, and how to retrieve it.

This map is just starting to be deployed. The URL and URN specifications are in the final throes of standardization, and there is a fairly clear idea of what is necessary to build the URN -> URL mapper which will maintain these persistent references. Once it is in place, information producers will be able to use persistent references to their own and to other's resources, with the guarantee that they will be able to retrieve the resource at any time in the future, if it still exists on the Internet.

However, URNs and URLs are only the first part of an Internet information architecture. A small group of researchers associated with the IETF is attempting to build a functional specification for services required by a generic Internet information tool, such as security, caching, and update capabilities. Once this functional specification is in place, new information tools will be able to be build on existing components rather than reinventing everything from scratch. As this group just started its work, it is difficult to determine what the final outcome will be.

III. Unapproachable Bogs

As this architecture for support of information tools becomes clearer and starts to be implemented, there are a number of features we should explore about our interactions with the tools, so that we can design new tools which meet our information needs.

The first is that the 'come get it' tools, particularly Gopher and WorldWideWeb, are still heavily text based, and are essentially text browsing tools. Despite the fact that they can display non-text resources, the vast majority of the navigational cues are based on text and require reading to select the desired choice. While text is indeed a very good way to rapidly impart a lot of information, I suspect that tools which use mostly or solely visual cues will become increasingly important to our exploration of the Internet. Just as the presentation of scientific data was revolutionized when artists started using color maps to indicate different quantities in observed data, a similar revolution is required if we are to extract as much information as possible from the vast masses of data on the Internet.

The second limiting factor is that information tools require one to follow the information provider's concept of how resources are arranged and connected. Although clients such as Mosaic allow some limited annotation and bookmarks, the URN infrastructure needs to be in place if we wish to allow any user to easily establish their own connections between resources. The proposed infrastructure will allow people to create virtual databases of material of interest to them, where all the requisite indexes and links are kept on their local machine while the actual resources remain somewhere else on the Internet.

The third limiting factor is the paltry state of the navigation tools. The descriptive information contained by (for example) Gopher, for each of the resources it points to, is rather thin to support the kind of navigation we might hope for. A much more knotty problem is the limits of the current search techniques themselves. Keyword searches only take you so far in the world of the Internet, where there is no vocabulary control, no consistent metainformation for various resources, etc. We need navigational tools which can understand semantics as well as syntax, and a consistent infrastructure for information about resources. The architecture being developed in the IETF does address at least some of these problems.

The fourth limiting factor, and one which may never be removed, is simply the vast mismatch between the size of the resources and the bandwidth to the average user. Many information providers tend to test their new resources locally, and notice no problems. However, anyone who has attempted to use a 9600 baud connection to access a

WorldWideWeb home page with embedded graphics know how frustrating it can be to wait many minutes for the entire download. As local nets get faster, the temptation to create ever bigger resources increases, which may not be good news for many users.

IV. Conclusion

It will be quite a while before we are able to map out all of the landscape of the Internet. We'll need new surveying tools. We'll need to make sure that the tools we do have work together. We'll need to build a set of signposts so that we can all use the same techniques to explore the Internet, and so that we can refer to specific points of interest in a consistent fashion. And, perhaps most importantly, we need to build flexibly so that each person is capable of getting the most out of the Internet. If we do this, we can tame the wild beasts and chart the unapproachable bogs.

V. References

[1] C. Weider and P. Deutsch, "A Vision of an Integrated Internet Information Service", Internet Draft, October 1993. URL:ftp://nic.merit.edu/documents/internet-drafts/draft-ietf-iiir-vision-01.txt.

[2] T. Berners-Lee, "The Uniform Resource Locator", Internet Draft, March 1994. URL:ftp://nic.merit.edu/documents/internet-drafts/draft-ietf-uri-url-03.txt.

[3] C. Weider and P. Deutsch, "Uniform Resource Names", Internet Draft, October 1993. URL:ftp://nic.merit.edu/documents/internet-drafts/draft-ietf-uri-names-02.txt.

Author Information

Chris Weider joined Bunyip Information Systems in December 1993. He leads development of directory protocols and is also manager of the Educational Services division. He has been a leader in Internet information systems architecture and design since 1990, and is also well known for his work on directory services protocols, particularly his creation (with Peter Deutsch, Jim Fullton, and Simon Spero) of the WHOIS++ directory service. He received B. Sc. degrees in Mathematics and in Computer Science from the University of Missouri in 1987, and a M.A. in Mathematics from the University of Michigan in 1992.

Mr. Weider co-chairs the Integrated Directory Services Working Group and the Integration of Internet Information Resources Working Group of the IETF. He is a member of ISOC and of the ACM