World Wide Web: O universo da informação[1,2]

Tim Berners-Lee, Robert Cailliau, Jean-François Groff, Bernd Pollermann
CERN, 1211 Geneva 23, Switzerland

Resumo

A iniciativa World Wide Web (W^3) é um projeto prático para criar um universo global de informação usando a tecnologia disponível. Este artigo descreve os objetivos, modelo de dados e protocolos necessários para implementar a "web" e a compara com vários outros sistemas contemporâneos.

O SONHO

Pegue sua caneta, mouse ou dispositivo apontador favorito e pressione em uma referência neste documento – talvez no nome do autor, da sua instituição ou em algum trabalho relacionado. Suponha que você seja diretamente levado a material de suporte – outros artigos, as coordenadas do autor, o endereço da instituição ou toda a sua lista telefônica. Suponha que cada um desses documentos tenha a mesma propriedade de estar ligado a outros documentos ao redor do mundo. Você teria na ponta dos seus dedos tudo o que precisa saber sobre publicação eletrônica, física de alta energia ou, por qualquer motivo, cultura asiática. Se você está lendo este artigo em papel, só pode sonhar, mas continue lendo.

Desde o artigo de Vannevar Bush [1], os homens têm sonhado estender seu intelecto tornando seu conhecimento coletivo disponível a cada indivíduo através das máquinas. Os computadores nos dão duas técnicas práticas para a interface homem-conhecimento. Uma é o hipertexto, em que ligações entre pedaços de texto (ou outra mídia) imitam a forma humana de associação de ideais. A outra é a recuperação de texto, que permite que as associações sejam deduzidas a partir do conteúdo do texto. No primeiro caso, a operação do leitor é tipicamente clicar com o mouse (ou digitar um número de referência) – no segundo caso, é fornecer algumas palavras representando aquilo que deseja. O mundo ideal W^3 permite ambas as operações e oferece acesso de qualquer plataforma de navegação.

A REALIDADE

Projetos de pesquisa e produtos comerciais existentes não estão longe de alcançar parte deste sonho. O sistema Xanadu [2] é um ambicioso projeto de hipertexto distribuído. Sistemas hipertexto existentes (veja, por exemplo, [3, 4]) tendem a serem restritos ao sistema de arquivos local ou distribuído e geralmente são desenvolvidos com um limitado conjunto de

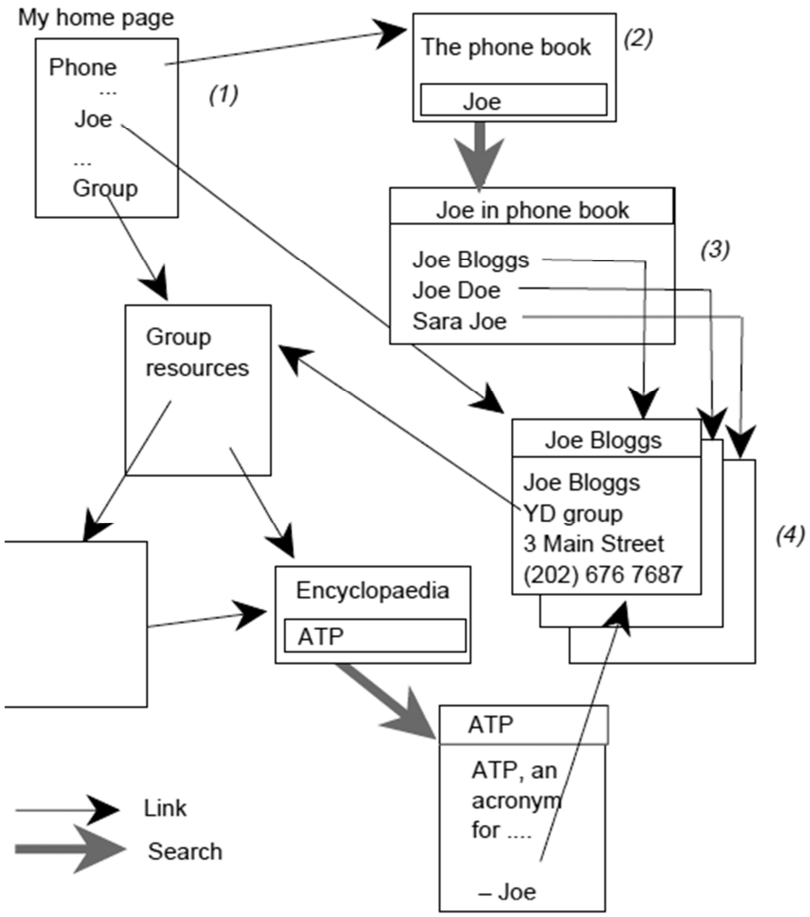
[1] BERNERS-LEE, Tim; CAILLIAU, Robert; GROFF, Jean-François; POLLERMANN, Bernd. World-Wide Web: The information universe. **Electronic Networking**, v.2, n.1, p. 52-58, 1992.

[2] Traduzido por Marcos André S. Kutova

plataformas em mente. Sistemas de acesso e recuperação de informação contemporâneos, como Alex [5], Gopher [6], Prospero [7] e WAIS [8], cobrem uma vasta área sem a funcionalidade do hipertexto. A fusão das técnicas do hipertexto, recuperação de informação e redes de grande escala produz o modelo W^3. Isso impõe requisitos específicos de esquemas de nomes de documentos, protocolos e representação de dados.

O MODELO DE DADOS W^3

O modelo W^3 usa ambos os paradigmas de ligações hipertexto e busca de texto de uma forma complementar, pois nenhuma pode substituir a funcionalidade da outra. A figura 1 mostra como uma *web* de informação personalizada é construída com esses operadores.



O modelo W3 envolve ligações hipertexto e buscas indexadas. O leitor parte de uma página inicial (1) e rapidamente usa suas próprias ligações, de grupos ou ligações públicas para encontrar recursos. Os índices como a lista telefônica (2) são representados como documentos com a possibilidade da digitação de palavras de busca. O resultado é um documento hipertexto virtual (3) que aponta para os documentos encontrados (4).

Figura 1: Uma *web* de ligações e índices

As características a observar são:

- A informação precisa ser representada apenas uma vez, pois uma referência pode ser usada ao invés de se fazer uma cópia;
- As ligações permitem que a topologia da informação evolua, modelando então o estado do conhecimento humano a qualquer hora sem restrições;
- A *web* se estende de forma transparente desde pequenas notas pessoais na estação de trabalho local a grandes bancos de dados em outros continentes;

- Os índices são documentos e podem por sua vez serem encontrados nas buscas e/ou através das ligações. Um índice é representado para o usuário como uma "página de capa" que descreve os dados indexados e as propriedades da máquina de busca.
- Os documentos na *web* não precisam existir como arquivos: eles podem ser documentos "virtuais" gerados pelo servidor em resposta a uma consulta ou nome de documentos. Eles podem representar visões do banco de dados ou imagens instantâneas de dados variáveis (como uma previsão do tempo, informações financeiras, etc.).

Um aspecto agradável e útil é que quase todo sistema de informação existente pode ser representado nos termos do modelo W^3. Um menu se torna uma página de hipertexto em que cada elemento é ligado a um diferente destino. O mesmo acontece com um diretório, seja parte de sistema hierárquico ou de referência cruzada. A possibilidade de se dar nomes aos índices na *web* permite que uma máquina de busca ou banco de dados estejam visíveis por vários endereços diferentes, cada um representando diferentes opções para o algoritmo de busca. Por exemplo, o índice `/library/books/ti+au/substring` pode representar uma busca de título e autor, enquanto que `/library/books/text/exact` pode representar uma busca de texto completo com palavras exatas. Os endereços são discutidos mais detalhadamente abaixo.

PUBLICAÇÃO

Do ponto de vista do provedor de informação, as informações existentes nos sistemas atuais podem ser publicadas na *Web* simplesmente oferecendo-se acesso aos dados através de um pequeno programa servidor. Os dados propriamente ditos, assim como o software e os processos dos usuários que os gerenciam, permanecem inalterados. Esta abordagem permite, por exemplo, que o armazenamento de documentos em um mainframe e o seu sistema de indexação sejam abertos para serem acessados por todas as plataformas da empresa. Para se entender como isso é feito, é necessária uma breve revisão da arquitetura W^3.

ARQUITETURA W^3

Os sistemas de hipertextos e de recuperação de textos têm estado disponíveis há muitos anos e uma pergunta válida é por que um sistema global ainda não foi criado? As respostas tradicionais para essa pergunta são a falta de:

- Um esquema comum de nomes dos documentos
- Protocolos comuns de acesso à rede
- Formatos de dados comuns para os hipertextos

Muito da pesquisa em sistemas de hipertexto (à exceção do projeto Xanadu) tem focado em questões da interface do usuário e de autoria, ao invés de questões de distribuição por áreas extensas ou longo tempo. Essas arquiteturas assumiram que os usuários compartilham um programa aplicativo comum rodando em computadores (geralmente do mesmo tipo) que, por sua vez, compartilham um sistema de arquivos comum. A arquitetura W^3 deve operar com um conjunto de computadores heterogêneos, distribuídos por uma grande área, rodando

aplicações diferentes que possuem formatos de dados preferenciais. Isso requer um modelo cliente-servidor. O cliente é responsável por traduzir o endereço do documento em um documento usando seu repertório de protocolos de rede. O servidor provê dados em um hipertexto simples ou no formato básico de texto ou, ainda, através da negociação com o cliente de qualquer outro formato de dados.

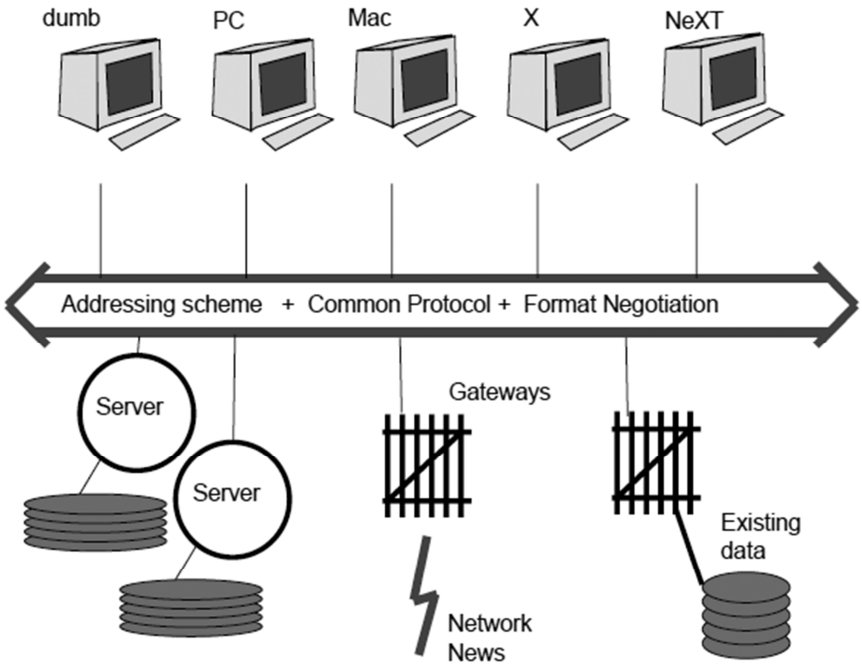


Figura 2: A arquitetura W^3 esboçada

Inicialmente, pode ser mais difícil inicialmente desenvolver um navegador hipertexto genérico que uma forma de acesso específica para um sistema de informação em particular. Entretanto, a separação entre os programas cliente e servidor pelo "barramento de informação" se paga quando mais clientes e servidores se conectam e o acesso universal é alcançado. Escrever um servidor para novos dados é geralmente uma tarefa simples, porque não requer a programação de uma interface homem-máquina.

NOMES DE DOCUMENTOS

O ponto de apoio em que o universo de documentos se apoia é o esquema para os nomes de documentos. Um nome de documento oferece um método para o cliente encontrar o servidor e para o servidor encontrar o documento. No modelo W^3, um nome também pode especificar uma parte do documento a ser selecionada pela aplicação de apresentação.

Apesar de um nome de documento normalmente estar escondido na sintaxe do hipertexto transferido pela ligação, na prática deve ser algumas referenciado pelas pessoas e passado por aplicações (como o e-mail) que ainda não compreendem hipertextos. Ele deve, portanto, ser idealmente composto por caracteres imprimíveis e ser pequeno.

Qualquer referência duradoura a um documento deve ser um nome lógico ao invés de um endereço físico. Isto é, ele deve referenciar o registro do documento com alguma organização

de “publicação” ao invés de um local físico, de tal forma que sua localização possa ser movida mais tarde. O cliente é, portanto, preparado para seguir vários estágios de tradução por servidores de nomes antes de encontrar o servidor final do documento. Similarmente, um nome de documento não deve contar informações que sejam transitórias, como os formatos específicos disponíveis para o documento ou seu tamanho, por exemplo.

O esquema de nomes W^3 atende a esses requisitos, mas é por outro lado aberto à adição de novos protocolos à medida que a tecnologia evolui. Para isso, um prefixo é usado para identificar o protocolo (e consequentemente o esquema de nome) a ser usado. Os clientes que não possuírem aquele protocolo em seu repertório usarão um *gateway* para tradução.

PROTOCOLOS

Os clientes W^3 são construídos a partir de um núcleo comum de códigos de rede para acesso às informações. Este núcleo provê acesso através protocolos da Internet bastante difundidos como:

- File Transfer Protocol – FTP [9]
- Network News Transfer Protocol – NNTP [10]
- Acesso a sistemas de arquivos instalados

Um novo protocolo para busca e recuperação se fez necessário, conhecido como HTTP. Mais rápido que o FTP para recuperação de documentos, ele também permite a busca indexada. O HTTP é similar em implementação aos protocolos da Internet acima e similar em funcionalidade ao protocolo WAIS. Algumas diferenças são discutidas abaixo.

FORMATOS DE DOCUMENTOS

O modelo de dados Dexter [11] oferece um modelo conceitual para sistemas hipertexto e o padrão HyTime [12] formaliza o hipertexto em um alto nível. O projeto W^3 define uma sintaxe concreta no estilo SGML para hipertexto básico como o usado para menus, resultados de busca e documentação hipertexto online. Cada aplicação de navegação W^3 é capaz de traduzir esse simples formato (figura 3).

Na fase piloto do projeto, esse formato era tudo que era necessário, mas na segunda fase, a negociação de formato entre cliente e servidor permitirá a troca de informação em qualquer meio usando qualquer representação mutuamente aceitável.

WAIS E A *WEB*

Do ponto de vista do sonho W^3, o protocolo WAIS representa um significativo avanço ao padrão de protocolo de busca e recuperação Z39.50/ISSO-10163, por ser sem registro de estado (*stateless*) e introduzir um nome persistente. Os nomes de documentos usados são locais ao banco de dados que os contém, mas a esses nomes podem ser acrescentados o nome do banco de dados e o endereço do servidor para formar um endereço W^3 universal. Desta forma, os índices e servidores WAIS podem ser representados na *web*. Um programa *gateway*,

rodando no CERN e disponível para uso geral, fornece esse mapeamento. O modelo WAIS usa arquivos “fonte” separados para descrever índices. O *gateway* WAIS-W^3 mantém uma cópia cachê desses arquivos usando-os para construir “páginas de capa” descritivas para os índices.

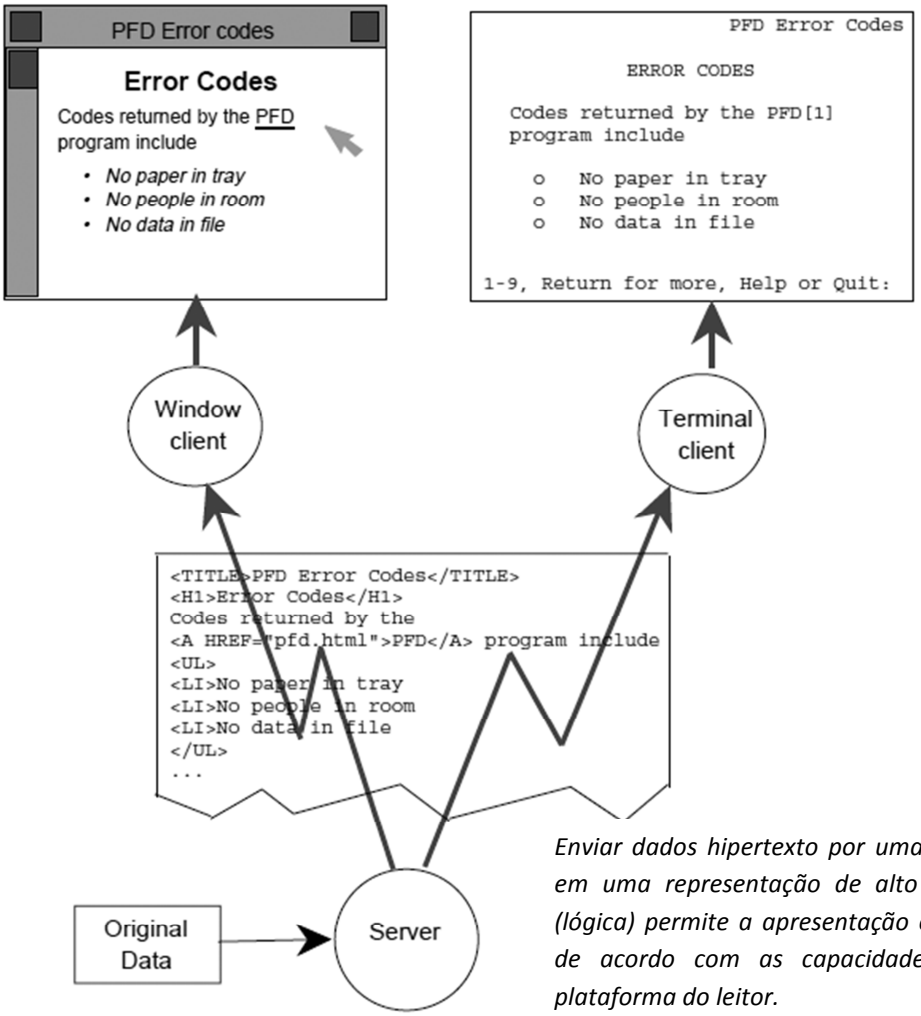


Figura 3: Um esquema da codificação de dados hipertexto. A ligação é representada em uma janela pelo texto sublinhado ou no terminal por um número de referência.

O modelo WAIS atual requer que os resultados de uma busca apontem para documentos disponíveis no mesmo servidor. Isto é, o mesmo servidor é responsável por indexar e, na realidade, prover os dados. No mundo W^3, essa restrição não existe. Uma vantagem prática dessa abordagem é que, como Yeong aponta [13], um documento multimídia extenso pode ser mais eficientemente recuperado de um servidor diferente e através de um protocolo diferente àquele usado na consulta original. Além do mais, à medida que a informação *online* proliferar, uma função importante será a do revisor, indexador e resumidor “terceirizado” que referem a dados que não os armazenam. É esperado que esses serviços sejam chave no controle da explosão da informação e valiosos recursos da comunidade.

Um usuário W^3 constrói uma *web* personalizada de informações fazendo ligações de seus próprios dados na *web*. Ele pode fazer uma ligação ao resultado de uma busca a ser realizada, de tal forma que a próxima vez que seguir a ligação a busca será reavaliada. Isso é equivalente a se armazenar uma “pergunta” WAIS – há um bom mapeamento entre os modelos. Os

clientes W^3 ainda não suportam a retroalimentação da relevância, mas isso não é estranho ao modelo.

Há duas ocasiões em que o hipertexto poderia particularmente melhorar o modelo WAIS. Em primeiro lugar, os usuários frequentemente gostariam de poder navegar pelos índices WAIS disponíveis. Tanto o WAIS quanto a W^3 tratam os índices como documentos e, portanto, permitem que sejam encontrados através das mesmas técnicas usadas para documentos. Na verdade, o *gateway* WAIS-W^3 permite que um documento hipertexto geral W^3 seja criado com apontadores para os índices WAIS. Em segundo lugar, quando alguém encontra um pedaço de texto, o WAIS entrega apenas aquela parte de arquivo que foi encontrada. Muitas vezes esse alguém gostaria de ligações para as informações ao redor no mesmo banco de dados.

A popularidade do WAIS tem sido uma grande impulsionadora do mundo da informação *online*. Sua integração com uma nomenclatura universal e com o hipertexto deve ser muito encorajada.

SISTEMAS DE MENU E A *WEB*

Os sistemas Alex [5], Internet Gopher [6] e Prospero [7] usam o modelo de diretórios e arquivos (ou menus e documentos) para implementar um sistema de informação global. Isso se mapeia muito naturalmente na *web*, em que cada diretório (menu) é representado por uma lista de elementos textuais ligados a outros diretórios ou arquivos (documentos). Esses sistemas são muito confortáveis para os leitores que estão acostumados com sistemas de arquivos hierárquicos, para quem diretório é um conceito bem estabelecido. Mesmo quando a estrutura é, na verdade, de referências cruzadas, o leitor se sente em casa quando a considera uma estrutura de árvore. Além disso, para o provedor de informação, tais sistemas são fáceis de serem construídos através do cruzamento de referências entre sistemas de arquivos existentes.

Um exemplo de mapeamento de um sistema de menu na *web* é feito pelo software cliente W^3 que incorpora o simples protocolo Gopher e, portanto, permite ligações no sistema Gopher. A inicialização rápida desses sistemas os tornou razoavelmente populares. É verdade que um menu é necessariamente uma mídia mais restrita que um hipertexto genérico: uma página de hipertexto pode conter várias informações para o leitor sobre as opções a serem seguidas, através do uso de uma formatação mais flexível. O hipertexto permite que os menus de ligações levem a nodos com progressivamente mais conteúdo textual. Entretanto, o mundo restrito do texto simples e do menu, com sua facilidade de publicação, é adequado para muitos provedores de informação.

Similarmente, os clientes W^3 também possuem a habilidade embutida de navegar pelo mundo dos arquivos FTP anônimos e um *gateway* que provê acesso ao sistema de ajuda VMS™ da Digital™.

X.500 E A *WEB*

O padrão x.500 para servidores de nome oferece uma forma interessante para nomes de longo prazo de documentos. Inicialmente planejada para coordenadas de pessoas e empresas, ao ser usada em documentos exigirá extensões similares (mas mais simples) àquelas propostas, por exemplo, por Yeong [14]. O principal atributo de um documento para os propósitos W^3 é o endereço físico W^3. Assim que o acesso aos servidores de nomes estiver amplamente disponível, "nomes amigáveis" oferecerão um formato de nome de documentos W^3 para endereços lógicos.

A EXPERIÊNCIA COM O PROJETO PILOTO W^3

O primeiro software cliente escrito com os requisitos W^3 rodou em uma máquina NeXT usando as ferramentas NeXTStep™ de interface gráfica com o usuário. Esse editor/navegador de hipertexto demonstrou a facilidade de uso de uma interface de hipertexto baseada em janelas para as informações globais. Também permitiu que um banco de dados de hipertextos sumário fosse construído para apontar para dados na *web* por assunto ou empresa. O segundo cliente escrito foi um navegador no modo de linhas para terminais não gráficos. Sendo portátil para qualquer máquina, fica assegurada a legibilidade universal para todos os documentos publicados. A documentação hipertexto foi posta *online* e os *gateways* foram montados em vários sistemas de informação existentes.

Usuários entusiastas de softwares de navegação apreciaram particularmente a interface com o usuário consistente para todos os tipos de dados. Ler artigos de notícias como hipertexto foi um bom exemplo: a mesma interface com o usuário foi oferecida e as referências entre os artigos e as referências entre os artigos e os grupos de notícias em que eram publicados eram todas consistentemente representadas por ligações.

Ficou evidente que tantos as ligações hipertexto quanto a busca de texto são partes importantes do modelo. Uma procura típica de informação começa em uma página hipertexto padrão seguindo ligações de um índice. Uma busca nesse índice pode retornar os dados desejados ou mais ligações para serem seguidas. Algumas vezes, um novo índice pode ser encontrado e vasculhado, e por aí vai. Quando o usuário de um editor hipertexto encontrar o que procura (não importa o quão remoto), ele pode fazer uma nova ligação para isso da sua página inicial para que possa localizar numa próxima vez quase instantaneamente. Isso é geralmente preferível que fazer uma cópia que logo pode estar desatualizada.

O FUTURO

O sucesso do projeto piloto estimulou a continuação do desenvolvimento de softwares e informações compatíveis com a W^3. Os projetos atuais de clientes por várias empresas incluem três navegadores baseados no X11 e um navegador Macintosh. Vários *gateways* de servidores para outros sistemas de informação foram produzidos e o total de informação disponível na *web* está se tornando significativo, especialmente por incluir todos arquivos FTP anônimos, servidores WAIS e servidores Gopher, bem como específicos servidores W^3. Percebemos que

as funções de cada um desses servidores poderia ser oferecida por servidores W^3 e que assim um único protocolo poderia ser usado por toda a comunidade.

O projeto Archie oferece um índice para os arquivos na Internet e é um excelente exemplo de serviço que esperamos que esteja disponível na *web*. Podemos imaginar tal indexação sendo estendida para abranger outras formas de dados. A W^3 oferece a infra-estrutura básica para acesso a informação. Todos os tipos de ferramentas de indexação, busca, filtragem e análise poderiam ser utilmente construídas usando o mecanismo genérico de acesso da W^3 e assim serem aplicadas aos vários domínios dos dados. Os resultados disso poderiam estar disponíveis na *web*. Muitos projetos de pesquisa em hipertexto seriam possíveis a partir da existência de uma grande e interligada base de informações.

Enquanto isso, a equipe da W^3 no CERN e seus colaboradores ao redor do mundo convidam qualquer provedor de informação a se juntar à *web*, contribuindo com informação ou software. Informações detalhadas sobre os protocolos, formatos de dados, etc. da W^3 estão disponíveis no nosso servidor W^3. A forma mais direta de acessar isso é por telnet no endereço `info.cern.ch`. Uma forma melhor é rodar nosso software navegador (disponível por FTP no mesmo servidor) na sua máquina local. Se você usar um navegador orientado a janelas, então poderá ler artigos como esse na sua tela. Quando o fizer, pegue sua caneta, mouse ou dispositivo apontador favorito e clique em uma referência neste documento... o sonho está se tornando realidade.

REFERÊNCIAS

[1] Bush, Vannevar, "As We May Think", *The Atlantic Monthly*, July 1945

[2] Nelson, Theodor H., *Literary Machines* version 90.1, Mindfull press 1990.

[3] "Beyond Hypertext: The DECWindows Hyperenvironment Vision", Digital Equipment Corporation, Maynard, MA., 1990

[4] Kahn, Paul and Normal Meyrowitz. "Guide, HyperCard, and Intermedia: A Comparison of Hypertext/Hypermedia Systems", *IRIS Technical Report* 88-7. Brown University, Providence RI, 1988.

[5] Cate, Vincent, Carnegie-Mellon Univerity, private communication.

[6] Alberti et.al. "Notes on the Internet Gopher Protocol" Univeristy of Minnesota, December 1991.

[7] Neuman, Clifford B., "The Prospero File System: User's manual". Department of Computer Science and Engineering, University of Washington.

[8] Kahle, B., et. al., "WAIS Interface Prototype Functional Specification", Thinking Machines Corporation, April 1990

[9] Postel, J. and Reynolds, J. "File Transfer Protocol (FTP)", Internet RFC-959, October 1985.

[10] Kantor, B., and Lapsley, P., "A proposed standard for the stream-based transmission of news", Internet RFC-977, February 1986

[11] Halasz, F. & Schwartz,M., "The Dexter Hypertext reference Model", *Proceedings of the Hypertext Standardization Workshop January* 16-18, 1990, National Institute of Standards and Technology.

[12] GoldFarb, Charles F., *Information Technology – Hypermedia/Time-based Structuring Language (HyTime)*, ISO/IEC CD 10744 (Draft).

[13] Yeong, W., "Towards Networked Information Retrieval", *Technical report 91-06-25-01*, Performance Systems International, Inc.

[14] Yeong, W., P.S.I., "Representing Public Archives in the Directory", *Internet Draft*, November 1991.

[15] Emtage, A and Deutch, P, "archie – and Electronic Directory Service for the Internet", to be presented to the 1992 *usenix* conference.