SISU

PUBLIKATION 96:12

BEVAKNINGSRAPPORT – JUNI 1996

WWW'96

– den 5te internationella konferensen om WWW

Matts Ahlsén
Peter Rosengren

SVENSKA INSTITUTET FÖR SYSTEMUTVECKLING

SISU

Sammanfattning

The World Wide Web Conference är idag den största forskarkonferensen med inriktning på World Wide Web och dess användningsområden. WWW-96 hölls i Paris och lockade över 2000 deltagare. Årets WWW-konferens erbjöd en blandning av akademiska presentationer, kommersiella leverantörspresentationer och praktiska användarerfarenheter.

Den viktigaste trenden på årets WWW-konferens är det ökande intresset för *content management*. Motivet till detta är att det blir allt mer komplicerat att underhålla och vidareutveckla WWW-system, samtidigt som intresset för att effektivt kunna producera ny information och nya tjänster ökar. Lösningen ligger i att införa ett *content management system*. Hur ett sådant ska vara uppbyggt och vilka funktioner det ska ha är föremål för vidare forskning och utveckling. Men den teknik som idag ligger närmast till hands som plattform för *content management* är så kallade objekt-relationsdatabaser (OR-databaser).

Utöver content management kan man notera att *cachning* och *metadata* är i ropet. Med cachning avses att hämtade WWW-sidor temporärlagras i syfte att snabba upp nästa åtkomst av sidan. Funktionen finns redan i Internet-bläddrare som då temporärlagrar på persondatorn. Det som nu diskuteras är att konstruera gemensamma regionala och nationella cache-databaser. Det skulle kunna innebära att det i princip finns en lokal kopia av Internet här i Sverige som alla svenskar arbetar mot. Cache-databasen uppdateras till exempel varje natt.

Metadata är data om data. Detta är ett välkänt begrepp inom databas- och systemutvecklingsområdet där insikten om vikten av tydliga beskrivningar och scheman över databaser funnits länge. Samma insikt börjar nu dyka upp inom WWW-sfären. Sökmotorer, robotar, katalogtjänster etc behöver veta mera om vad som finns på WWW för att kunna förbättras. Hur WWW-sidor och servrar ska beskrivas ingår i metadata-arbete likväl som hur objekt ska namnges och hur deras innehåll ska beskrivas.

Det viktigaste som hänt de senaste månaderna när det gäller betalning över Internet är att två olika standardförslag för kreditkortsbetalning slagits samman till ett. Tidigare fanns *STT* (Secure Transaction Technology) från Microsoft och Visa samt *SEPP* (Secure Electronic Payment Protocol) från IBM, NetScape och MasterCard. Nu har de två lägren enats om standarden *SET* (Secure Electronic Transaction). Med tanke på de aktörer som står bakom SET så är det uppenbart att inga andra alternativ för kreditkortsbetalningar är aktuella. De första programdelarna för SET beräknas levereras under hösten 1996 och runt årsskiftet förväntas SET kunna vara i drift.

WWW i sig har också blivit ett forskningsområde. De flesta forskarna fokuserar på WWW som teknisk plattform och studerar antingen tekniska frågor kring WWW-arkitekturen eller användning av WWW på olika sätt till exempel för samarbete. Men tack vare WWW:s enorma utbredning i hela världen så har nu en del forskare börjat studera WWW ur ett rent statistiskt perspektiv. Hur stort är WWW, hur många hemsidor finns, vad pekar sidorna på, hur stor är en WWW-sida i genomsnitt är exempel på frågeställningar.

Intresset för design- och utvecklingsfrågor var stort på konferensen, ett tecken på att WWW har fått status som en plattform för systemutveckling. WWW-utveckling täcker idag ett mycket brett spektrum där det ingår allt ifrån utseendet på hemsidor och användbarhetsanalys till programvarukomponenter, databasstöd och systemadministration.

Innehåll

1 Introduktion

The World Wide Web Conference är idag den största forskarkonferensen med inriktning på World Wide Web och dess användningsområden. Eftersom utvecklingen inom WWW-området de senaste åren varit mycket intensiv har konferensen hittills anordnats två gånger om året. WWW-96 är den femte i ordningen och framöver ska man övergå till en konferens per år.

WWW-96 hölls i Paris och lockade över 2000 deltagare varav en förvånansvärt stor del kom från Sverige. Årets WWW-konferens erbjöd en blandning av akademiska presentationer, kommersiella leverantörspresentationer och praktiska användarerfarenheter.

Figur 1. CNit, La Defense, centret i Paris där konferensen WWW-96 hölls.

Den viktigaste trenden på årets WWW-konferens är det ökande intresset för *content management*. WWW har växt exponentiellt sedan 1994. Varje företag med självaktning har hoppat på tåget och skaffat sig en egen WWW-server för att presentera företaget och dess verksamhet. Nu när den första entusiasmen över att ha en egen WWW-server lagt sig börjar allt fler uppleva problem med att underhålla, uppdatera och vidareutveckla sina WWW-servrar. Projekt avslutas och nya startas, personer slutar och nya anställs. Produkter tillkommer och behöver presenteras på WWW-servern. Alla sådana förändringar kräver uppdateringar och förändringar av WWW-sidorna.

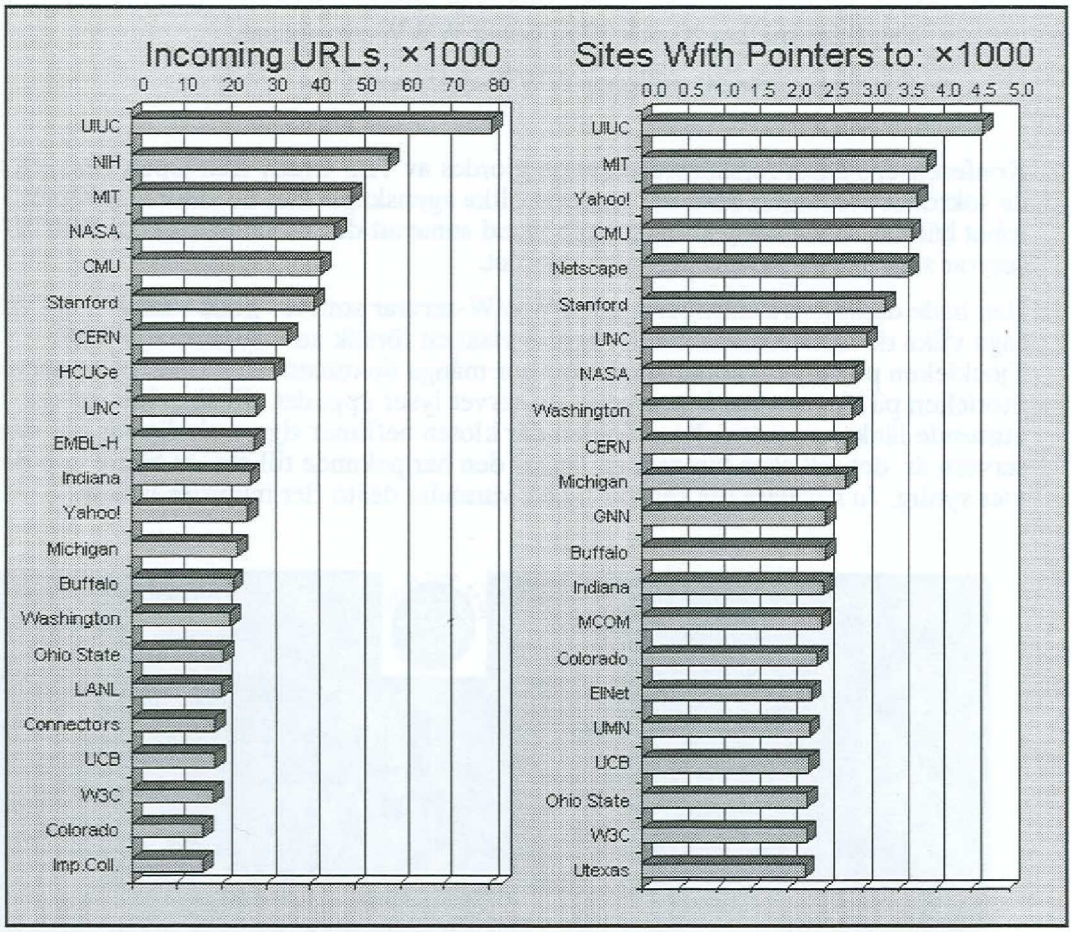
Med dagens serverteknik som bygger på en filsystemteknik är det ett tidsödande arbete att underhålla en WWW-server där innehållet ofta ändras och struktureras om. Idag finns det gott om verktyg för att editera enskilda HTML-sidor, men det är mer tunnsått med verktyg för att skapa WWW-serverns övergripande struktur eller för att enkelt hantera förändringar. Insikten om att det behövs någon form av strukturerad databas-

teknik har börjat sprida sig och det finns idag flera lösningar för att koppla ihop WWW med relationsdatabaser.

Nästa steg i utvecklingen är att införa ett *content management system.* Hur ett sådant ska vara uppbyggt och vilka funktioner det ska ha är föremål för vidare forskning och utveckling. Men den teknik som idag ligger närmast till hands som plattform för *content management* är så kallade objekt-relationsdatabaser (OR-databaser). Det är en sammansmältning av så kallade objekt-orienterade databaser och relationsdatabaser. De brukar även kallas *hybrid-databaser* (se SISU Rapport nr 96-11).

Utöver content management kan man notera att *cachning* och *metadata* är i ropet. Med cachning avses att hämtade WWW-sidor temporärlagras i syfte att snabba upp nästa åtkomst av sidan. Funktionen finns redan i Internet-bläddrare som då temporärlagrar på persondatorn. Det som nu diskuteras är att konstruera gemensamma regionala och nationella cache-databaser. Det skulle kunna innebära att det i princip finns en lokal kopia av Internet här i Sverige som alla svenskar arbetar mot. Cache-databasen uppdateras till exempel varje natt. Det är också möjligt att tänka sig att ett företag kopierar de WWW-servrar som företagets anställda oftast använder. Dessa kopior finns då internt inom företaget.

Med metadata avses data om data. Detta är ett välkänt begrepp inom databas- och systemutvecklingsområdet där insikten om vikten av tydliga beskrivningar och scheman över databaser funnits länge. Samma insikt börjar nu dyka upp inom WWW-sfären. Sökmotorer, robotar, katalogtjänster etc behöver veta mera om vad som finns på WWW för att kunna förbättras. Hur WWW-sidor och servrar ska beskrivas ingår i metadata-arbete likväl som hur objekt ska namnges och hur deras innehåll ska beskrivas.



Figur 2. De WWW-servrar som har flest länkar pekande till sig.

Intressant att notera är att WWW i sig har blivit ett forskningsområde. De flesta forskarna fokusera på WWW som teknisk plattform och studerar antingen tekniska frågor kring WWW-arkitekturen eller användning av WWW på olika sätt till exempel för samarbete. Men tack vare WWW:s enorma utbredning i hela världen så har nu en del forskare börjat studera WWW ur ett rent statistiskt perspektiv. Hur stort är WWW, hur många hemsidor finns, vad pekar sidorna på, hur stor är en WWW-sida i genomsnitt är exempel på frågeställningar. Figur 2 visar till exempel vilka WWW-servrar som har flest länkar pekande på sig.

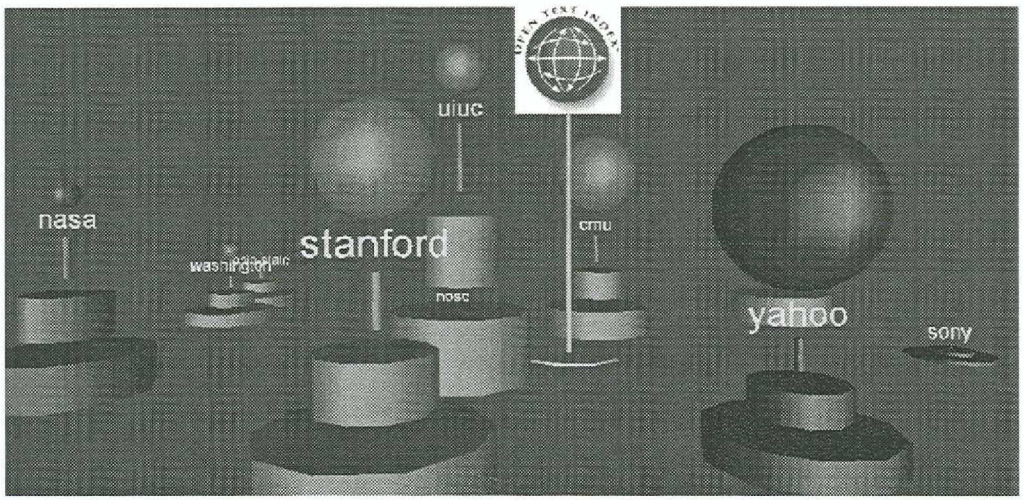
Hur ser då WWW ut? Ett axplock av siffror som presenterades på konferensen avslöjar ett inte så världsomspännande nät som vi kanske tidigare trott:

- 90 procent av Internet-trafiken går till eller ifrån en amerikansk server.
- 80 procent av alla WWW-servrar har inga utgående länkar till andra servrar.
- 56 procent av WWW-sidorna har bara en länk pekandes på sig.
- 25 procent av WWW-sidorna har ingen utgående länk.

- Det finns mellan 20 och 30 miljoner WWW-användare.
- Det finns cirka 50 miljoner WWW-dokument.

Konferensens bästa forskarpresentation gjordes av Tim Grady från Open Text, ett av de sökrobot-företagen. Han har studerat olika egenskaper hos de sidor som Open Text robot hämtat. Tim Grady konstaterade bland annat att det är endast ett fåtal WWW-servrar som sköter navigeringen på Internet.

Han hade dessutom kontrollerat vilka WWW-servrar som är "goda vänner", det vill säga vilka de har flest länkar till. Figur 3 visar ett försök att visualisera WWW. Tjockleken på varje cylinder illustrerar hur många dokument som finns på en server. Storleken på kloten visar hur mycket en server lyser upp, det vill säga hur många utgående länkar en server har. Höjden där kloten befinner sig symboliserar hur synlig servern är, det vill säga hur många länkar den har pekande till sig. Ju högre upp desto mer synlig. Ju närmare två servrar ligger varandra desto fler inbördes länkar.



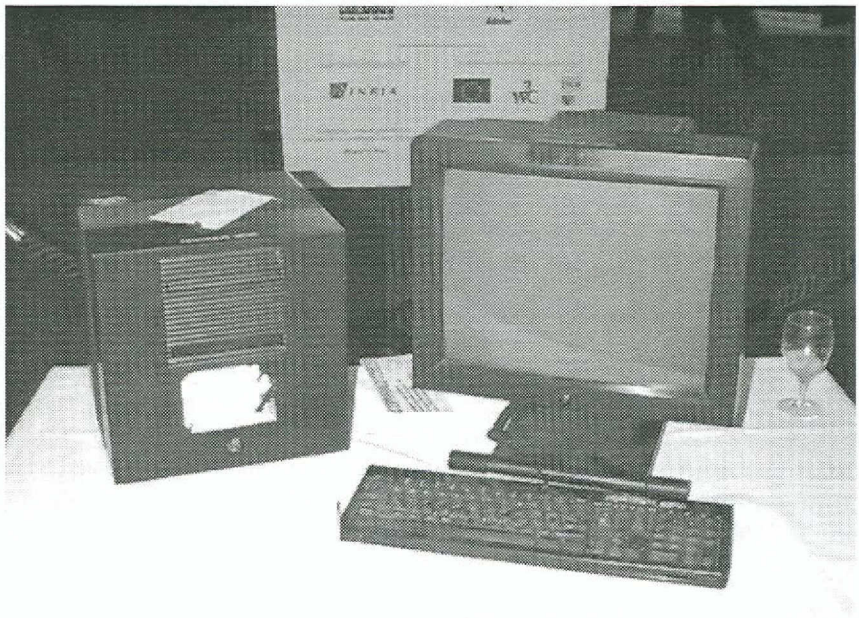
Figur 3. Så ser nätet ut. Ett försök att visualisera World Wide Web.

Tim Grady har specifikt undersökt vilka WWW-servrar som är bäst vänner med Playboy... Av omtanken om våra kollegor på andra forskningsinstitut och universitet vill vi dock inte avslöja detta i denna rapport.

Vi noterar att WWW-forskningen idag är mycket bred och diversifierad – användbarhet av WWW-sidor, åtkomst till gamla databaser via WWW, samarbete via WWW, informationssökning på WWW är bara några exempel på ämnen som behandlades under WWW-96. De flesta av de forskningsfrågor som togs upp på WWW-96 bearbetas redan inom andra fora och då också med ett större djup. Man får en känsla av att många av de forskare som idag arbetar med WWW som bas försöker lösa problem som i stor utsträckning studerats och även lösts inom andra områden, som t ex inom områden som databasteknologi, CSCW och hypertext. Vi noterar också att relativt få etablerade forskare inom dessa angränsande områden fanns med på denna konferens.

Vi ställer oss därför frågande till om det går att tala om någon egentlig WWW-forskning som tacklar WWW-specifika problem. Trots detta gav konferensen många intressanta inblickar i WWW-användning och vart den kommersiella utvecklingen är på väg, och även en insikt om att det ändå finns forskningsfrågor kring WWW. Som en kuriositet visades den dator upp på vilket World Wide Web utvecklades. Det är en

Next-dator. Tim Berners-Lee och Robert Calliau som då arbetade vid Cern talade nostalgiskt om tiden då World Wide Web bara bestod av de två som utbytte och kommenterade varandras forskningsrapporter. Så nostalgisk man nu kan vara om någonting som bara är fem år gammalt.



Figur 4. Datorn där World Wide Web skapades. Vid denna Next-dator utvecklade Tim Berners-Lee den första versionen av World Wide Web 1991.

2 Content Management

Det finns ett stort behov av verktyg och kunnande för att underhålla och vidareutveckla den stora mängd WWW-databaser som nu finns i företag och organisationer över hela världen.

WWW är ett interaktivt medium och som sådant är informationsinnehåll och presentation i fokus för de flesta tillämpningarna. Det finns många WWW-databaser som tillhandahåller mer eller mindre spektakulärt innehåll, vilka ofta används för att påvisa det fantastiska med WWW och Internet. En sådan tillämpning är den "skivade bankrånaren", d v s en i USA avrättad man som donerat sin kropp till forskning och som därför frysts ned, sågats i bitar och sedermera digitaliserats.

Under en av industripresentationerna som hölls av IBM, togs just denna tillämpning upp som ett exempel på den räckvidd som WWW fått. Sålunda kunde 2000 gapande konferensdeltagare se hur (delar av) denne djupfryste och itusågade man via en ATM-förbindelse transporterades från USA till föredragshållarens web-läsare. Utan WWW skulle denna typ av information förmodligen inte nå utanför ett fåtal specialiserade tillämpningsområden (kanske forskning, undervisning). Det unika med WWW är just kombinationen av produktion och distribution av information vars presentation, tolkning och användning kan variera oändligt.

2 .1 Från Data till Content Management

En av de viktigaste insikterna som denna konferens gav, var just att det nu finns en mängd tekniker såväl som designkunnande för att på allvar utveckla avancerad informationshantering på WWW-arkitekturen. På konferensen nämndes *content management* som ett samlande begrepp för informationshantering inom ramen WWW-tillämpningar, att jämföras med det mer traditionella *data management.*

När det gäller informationshantering så befinner sig WWW-utveckling idag ungefär i liknande situation som systemutvecklingen gjorde när databashanteringssystemen började ersätta filsystem. Man kan också konstatera att det finns en stor outnyttjad potential i forskningsresultat från områden, såsom databas- och objektteknik, metadatahantering och informationssystemarkitektur som kan kombineras med kunnande från områden som media och publicering.

En tolkning av begreppet *content management* är att det förenar ett traditionellt informationshanteringsperspektiv (repositories, databaser, client/server etc) med ett kommunikations- och publiceringsperspektiv, där WWW är den infrastruktur som låter detta göras.

I det traditionella och förhärskande datalagersynsättet på systemutveckling är kommunikation och informationsutbyte något som i allmänhet byggs på eller läggs till i efterhand eller åtminstone i ett relativt sent skede i en designprocess (undantag finns dock, t ex vissa ansatser workflow). De WWW-tillämpningar som nu utvecklas har kommit (eller är på väg) att vända på detta synsätt; kommunikation och innehåll först, struktur sedan.

2 .2 Teknik för Content Management

De viktigaste teknikområdena som förts samman genom WWW, och som ansatser till content management kan baseras på är,

- hyper(multi)mediala gränssnitt

- mobila programobjekt
- multimediala och aktiva databaser
- federerad databasarkitektur

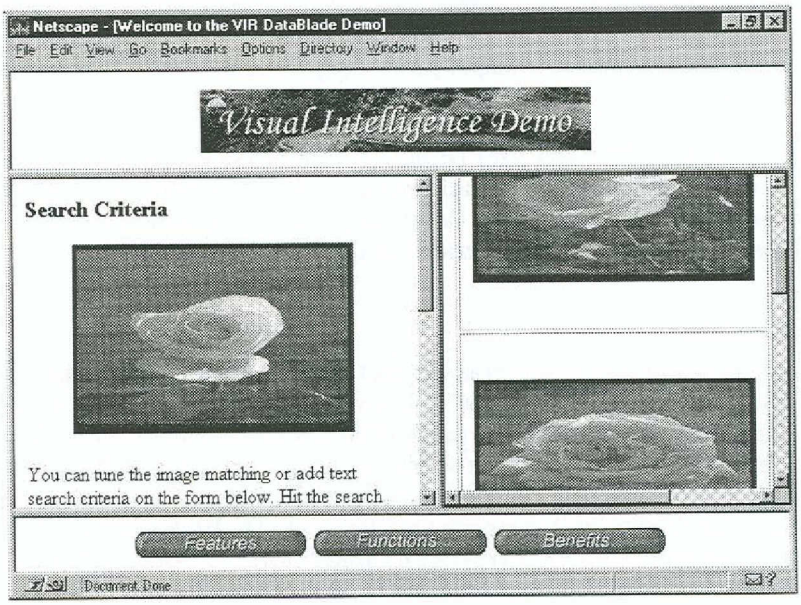
samt,

- teknik för gränssnitt mot existerande informationssystem.

Ortogonalt mot dessa teknikdimensioner har vi också skillnaden mellan intra- och inter-organisatoriska WWW-tillämpningar, som för med sig olika överväganden när det t ex gäller säkerhet, integritet och upphovsrätt.

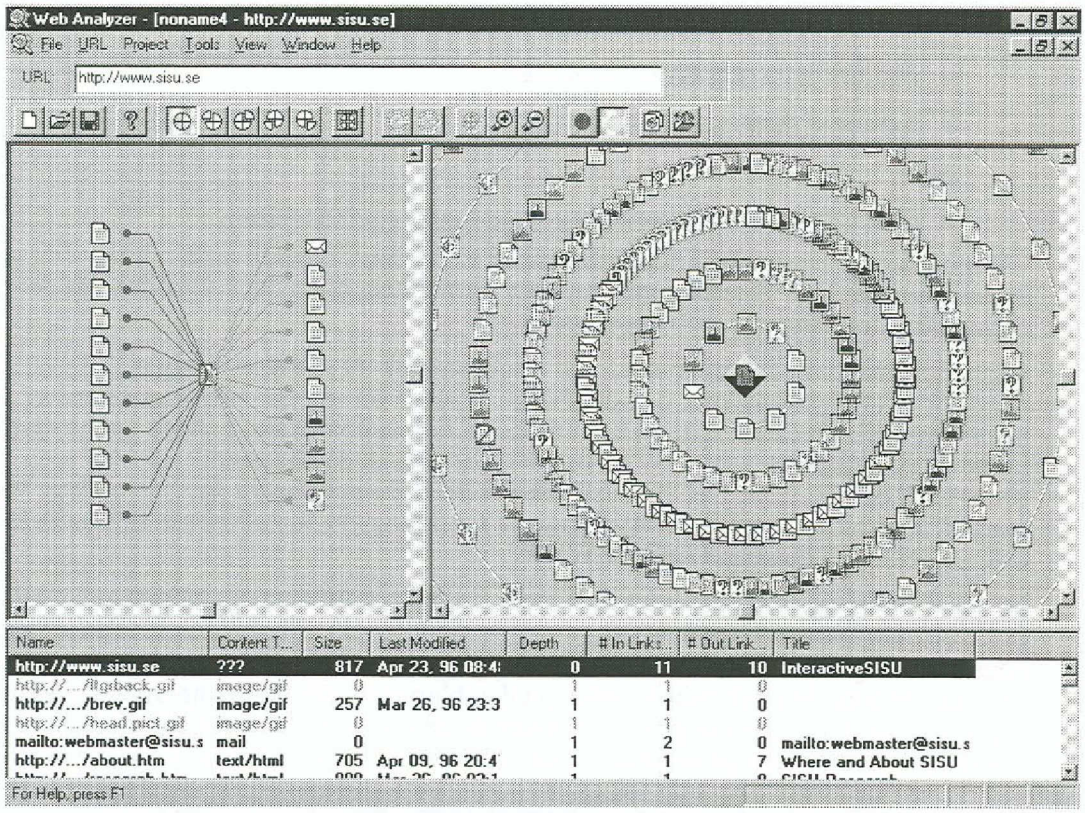
I detta sammanhang kommer etablerad databasteknik att spela en viktig roll. Tyvärr var det dock så att databasteknik och databaskompetens till stor del saknades i flera av de prototyper och modeller som presenterades på denna konferens. I något fall hävdades t o m att 90-talsteknik som WWW, inte bör baseras på databasteknik som har sina rötter i 60- och 70-talet! Det är också slående att få namnkunniga databasforskare var med på denna konferens, trots att många av forskningspresentationerna behandlade olika aspekter på databasteknik för WWW.

Ett databasföretag som försöker positionera sig inom området content management är Informix. De karaktäriserade tillämpningsområdet för sin OR-databas Illustra som just content management. Illustra är modulärt uppbyggt med så kallade *datablad* (data blades) som kan pluggas in efter behov. Separata datablad finns för hantering av olika media-typer. I Figur 5 ses ett exempel på användning av en speciell modul för bildsökning som gör det möjligt att söka på innehållet i en bild, istället för att använda sig av nyckelord.



Figur 5. Exempel på WWW-tjänst som byggts med hjälp av den objekt-relationella databashanteraren Illustra.

En annan typ av verktyg som nu börjar spridas är olika hjälpmedel för att analysera struktur och innehåll i en WWW-server. Behovet av sådana program motiveras främst av att dagens WWW-servrar bygger på primitiva filsystem som blir oöverskådliga i takt med att antalet WWW-dokument ökar.



Figur 6. Webanalyzer: exempel på verktyg för analys av WWW-servers.

Vissa verktyg, som Webanalyzer (Figur 6), kan också skapa lokala kopior av externa WWW-servrar. Figur 6 är en visualisering av SISUs publika WWW-server, där cirklarna representerar successiva länknivåer med hemsidan i centrum. Till vänster visas hemsidans alla in- och utgående länkar.

2 .3 Metadata

Ett problemområde som diskuteras mycket just nu är behovet av metadata på WWW. Med metadata avses "data om data". Behovet av metadata kommer av att man på olika sätt vill kunna beskriva innehållet på WWW. Metadata kan omfatta allt från författarnamn på ett dokument, logisk adress för ett dokument till en subjektiv gradering av innehåll, ur till exempel ett känslighetsperspektiv. Betydelsen av metadata är o-omtvistad inom traditionell systemutveckling och databashantering. Samma insikt börjar nu dyka upp inom WWW-sfären. Sökmotorer, robotar, katalogtjänster etc behöver veta mera om vad som finns på WWW för att kunna förbättras. Hur WWW-sidor och servrar ska beskrivas ingår i metadata-arbete likväl som hur objekt ska namnges och hur deras innehåll ska beskrivas.

Ett problem idag är namngivning. Navigering på WWW bygger på s k URL (Uniform Resource Locator), ett adresseringssystem som unikt pekar ut en fysisk plats på WWW där en viss resurs (vanligen ett dokument) förväntas vara tillgängligt. Om objektet som finns på den platsen flyttas så pekar URL:en fortfarande på platsen trots att inget objekt finns där längre. Därför diskuteras begreppet URN (Uniform Resource Name) istället. Detta ska bli ett sätt att unikt namnge ett objekt som sedan kan hittas oavsett var det fysiskt finns lagrat. Arbete med namngivning sker inom både W3C och IETF (Internet

Engineering Task Force). Identifiering och namngivning har också sedan länge studerats inom ramen för objektorienterade system.

Ett förslag till URN-standard är så kallade PURL:s (Persistence Uniform Resource Locator). De fungerar som URL med den skillnaden att en PURL pekar på en URL som sedan pekar på ett objekt. URL:erna lagras då istället i en databas. Om ett objekt flyttas behöver bara databasen uppdateras en gång sedan kommer alla länkar till objektet att fungera.

En annan aktivitet inom detta område är Platform for Internet Content Selection, PICS, som syftar till att möjliggöra känslighetsklassificering av WWW-information. Upprinnelsen till projektet är den heta debatten i USA som föranletts av den stigande mängden osedliga HTML sidor samtidigt som allt fler barn använder WWW.

Inför hotet om ett lagförslag om censur av information på Internet, tog W3C med ett 20-tal företag initiativet till PICS för att visa industrins vilja att hantera problemet. Man påpekar dock att PICS inte är centralstyrd censur, utan "ethical censorship without central control". PICS skall göra det möjligt för användare och producenter att välja och erbjuda olika typer av klassificeringstjänster (rating services) för WWW-information.

I ett vidare perspektiv är PICS ett exempel på stöd för metainformation på WWW, vilket inkluderar olika tekniker för kataloger, klassificering med mera, som idag saknas i stor utsträckning.

Metadata är ett typiskt exempel på ett område där tidigare databasforskning har producerat många användbara resultat som skulle kunna återanvändas inom WWW-området. Forskare inom områden som distribuerade databaser, federerade system, frågespråk etc har tacklat olika delar av metadataproblematiken.

En skillnad är att WWW-forskarna fortfarande i allmänhet fokuserar på metadata för att beskriva objektens yttre attribut t ex var det finns och vem som skrev, och inte så mycket metadata för att beskriva innehållet i ett objekt.

3 Elektronisk betalning

Elektronisk betalning över Internet är ett hett ämne och under WWW-96 ägnades flera sessioner åt detta. När man talar om betalsystem för Internet brukar man skilja på slutna och öppna system. Ett *slutet system* bygger på att kunden och handlaren har en långsiktig relation. Det kan till exempel vara att kunden får ett konto hos handlaren som debiteras alltefter som kunden förbrukar varor. Själva betalningstransaktionen sker sedan utanför Internet, till exempel via en vanlig kreditkortsnota eller faktura. Slutna system lämpar sig för en viss typ av försäljning, men förutsätter att kunden och handlaren känner till varandra. Idag är slutna system det vanligaste på Internet.

I ett öppet system där vem som helst ska kunna handla och dessutom när som helst krävs andra lösningar, t ex kreditkort, elektroniska checkar och digitala kontanter. I sådana system sker inte bara beställningen av varor, utan även själva betalnings-transaktionen över Internet. Om betalningen görs med hjälp av kreditkort så är handlarens server kopplad mot en kreditkortsbrygga (gateway), där kontroll och godkännande av betalningstransaktionen sker. Det fungerar på samma sätt som när vi betalar med kort på till exempel ICA, där det dras genom en kortläsare.

Det viktigaste som hänt de senaste månaderna är att två olika standardförslag för kreditkortsbetalning slagits samman till ett. Tidigare fanns *STT* (Secure Transaction Technology) från Microsoft och Visa samt *SEPP* (Secure Electronic Payment Protocol) från IBM, NetScape och MasterCard. Nu har de två lägren enats om standarden *SET* (Secure Electronic Transaction).

Med tanke på de aktörer som står bakom SET så är det uppenbart att inga andra alternativ för kreditkortsbetalningar är aktuella. De första programdelarna för SET beräknas levereras under hösten 1996 och runt årsskiftet förväntas SET kunna vara i drift. Exempel på företag som kommer att leverera SET-program är Microsoft, IBM, NetScape och Terisa Systems.

De kvarstående problemen är inte av en teknisk karaktär utan handlar om att bygga upp en certifierings-organisation. Varje Internet-handlare ska certifieras som SET-operatör. Samtidigt ska respektive kreditkortsföretag utfärda elektroniska certifikat till varje köpare. När ett köp genomförs på nätet överförs kreditkortsnummer på ett krypterat sätt tillsammans med ett certifikat som till exempel säger att "detta är ett giltigt Visa-kort och kortet har registrerats för Internet-handel". Men att få ut alla dessa certifikat till miljoner användare kommer inte att bli en lätt uppgift för Visa och MasterCard.

Fördelen med kreditkortsbaserad Internet-betalning är den stora tillgänglighet ett sådant betalsystem har. De flesta har kreditkort och är vana att använda det för köp. Nu återstår "bara" att övertyga folk om att kreditkort kan användas på ett säkert sätt över Internet. Med tanke på den uppmärksamhet som säkerhetsproblemen fått under den senaste tiden är detta en svår uppgift för de företag som vill få igång handel över Internet.

Ett problem med kreditkortstransaktioner är att hanteringskostnaderna är höga. Eftersom köp över Internet ofta kan förväntas röra små summor, till exempel 5 kronor för en databassökning, kan själva transaktionskostnaden då överstiga värdet på det köpta. Ytterligare en nackdel är att kreditkort inte fungerar för betalningar mellan två personer. Dessa resonemangen har lett fram till att teknik utvecklats för *digitala pengar*.

Inom området digitala pengar är inte situationen lika klar som när det gäller kreditkort. DigiCash ligger längst fram med sina *ecash*, men flera olika alternativ finns.

CyberCoins från CyberCash och *Millicent* från Digital är exempel på alternativ som seglat upp. Teknik för digitala pengar är avsedd för framför allt små belopp.

Digitala pengar tycks dock ha det trögt. Trots att ecash funnits tillgängligt sedan november, och innan dess i en experimentversion i cirka ett halvår, så finns det än så länge bara två banker i världen som accepterar ecash – Mark Twain Bank i USA och Merita Bank i Finland. Det finns inte mer än ett drygt 20-tal ecash-handlare registrerade hos Mark Twain Bank i USA. Posten har licens för ecash i svenska kronor men har ännu inte kommit igång med någon verksamhet.

Nytt på senare tid är också att elektroniska checkar börjat diskuteras allt mer som en möjlig betalform. GlobeID är ett exempel på ett checkbaserat system som prövas i Frankrike. Principen är att checkar kan köpas via kreditkort och sedan användas vid betalning.

I och med att fler än en betalningsform kommer att finnas tillgänglig på nätet uppstår situationer då köpare och säljare har flera alternativ att välja mellan för att genomföra en transaktion. Ungefär som när man idag handlar och kan välja mellan att betala med kort, check eller kontanter. Vissa kort kanske ger någon form av bonuspoäng i en viss affär, eller att vissa varurabatter ges.

JEPI (Joint Electronic Payment Initiative) är ett stort projekt som startats av W3C (World Wide Web Consortium) som tacklar problemet med flera betalningsformer. Syftet med JEPI-projektet är att utveckla system för att två parter på nätet ska kunna förhandla om olika betalningsformer. Flera stora företag, till exempel Microsoft, NetScape, IBM och CyberCash deltar och under hösten kommer ett prototyp-system att demonstreras. JEPI bygger på W3C-protokollet PEP (Protocol Extension Protocol) som är ett standardiserat sätt att göra tillägg till http-protokollet.

Utvecklingen mot elektroniska betalningsmetoder på Internet är inte oväntad. Ska företag kunna tjäna pengar och göra affärer på Internet krävs också att de kan ta betalt på ett smidigt och säkert sätt. Det som händer nu är att teknik och standarder för betalning över Internet kommer på plats. Det innebär inte automatiskt att Internet kommer att blomstra som kommersiell marknadsplats. Att göra affärer är inte enbart en fråga om att ta betalt, man ska ha något att sälja också och det ska vara vettigt prissatt. Att hitta rätt prismodeller kommer att vara en stor utmaning om Internet ska kunna fungera som en informationsmarknad.

Det finns idag många existerande betalsystem i drift eller under utprovning på Internet. Se också rapporten "Betalsystem för Internet", (SISU Rapport nr 96-04).

4 Att hitta och hittas på WWW

Den stora informationsrymd som idag i princip är tillgänglig för vem som helst med en WWW-läsare är mycket heterogen och oöverblickbar. Det finns nu en mängd olika söktjänster att tillgå via Internet/WWW och sökverktygen blir allt kraftfullare.

I takt med att både mängden och mångfalden av information på WWW ökar, så blir det inte bara viktigt att se (hitta) utan minst lika viktigt att synas och bli hittad. Två sidor av samma mynt. Sökning och indexering av WWW-information är alltjämt ett viktigt område för vidareutveckling. På konferensen fokuserades detta i ett flertal presentationer och tutorials.

Informationssökning, "Information Retrieval", är ett forskningsområde som funnits i 30 år. Därför är problemen väl förstådda, dock inte nödvändigtvis lösta. En effektiv söktjänst, oavsett om den är WWW-baserad eller inte, förutsätter dels en effektiv och heltäckande indexering av informationsmängden, dels ett lättanvänt men kraftfullt sökspråk för återsökning av informationsobjekt.

4 .1 Indexering och robotar

Indexering innebär att orden i ett dokument extraheras och lagras i en databas. Det är sedan indexet som används när återsökning av dokument görs. För att indexet inte ska bli större än dokumentet i sig tas så kallade stoppord bort, t ex "och", "men", "så", "jag", "i", "på". Indexering görs av en *index-motor*. En bra index-motor kan hantera fraser, böjningar, sammansatta ord med mera.

De söktjänster som finns idag på WWW är uppbyggda enligt samma principer som vilket normalt textsökningssystem som helst, dock med den skillnaden att såväl index-motorer som sök-motorer är förhållandevis primitiva jämfört med vad som normalt finns tillgängligt i ett textsökningsprogram. Indexering är det som är mer problematiskt för en WWW-söktjänst jämfört med ett vanligt textsökningssystem. I ett vanligt textsökningssystem indexeras dokumenten vartefter de lagras i textdatabasen. I en WWW-tjänst är det stora problemet att hitta de dokument som ska indexeras.

Indexeringen i en WWW-tjänst sköts av en så kallade *robot*. Det är en programvara som har till uppgift att hitta så många WWW-dokument som möjligt och indexera dem. Robotarna extraherar alla ord i ett WWW-dokument och lagrar dem sedan i en central databas. Det är sedan i denna databas som sökningar görs. Den största index-databasen för närvarande finns hos AltaVista, Digitals söktjänst. I princip kan man säga att Digital har en egen kopia av Internet på en mycket snabb och kraftfull Alpha Server.

Eftersom WWW är ett helt decentraliserat system utan någon central kontroll finns det idag ingen som vet hur många WWW-dokument som finns än mindre exakt **var** alla finns. Robotarna fungerar så att de helt enkelt börjar på ett antal kända servrar, typ Yahoo, och därifrån börjar med att läsa ett dokument och extrahera orden från det dokumentet. Därefter följs alla länkar från det dokumentet till andra dokument där samma procedur upprepas. Detta pågår sedan dygn efter dygn tills roboten inte kommer längre. Därefter börjar roboten om från början...

Det som påverkar prestanda hos en robot är dels hur snabbt ett dokument kan flyttas från källan till indexeringsdatorn, dels hur snabbt indexeringen kan göras. För att hinna indexera hela WWW krävs att många dokument hämtas och indexeras parallellt. Om roboten klarar av att hitta och indexera ett (1) dokument per sekund så tar det ungefär 8 månader att indexera hela WWW. Detta är inte acceptabelt eftersom en stor del av dokumenten som indexerades under dessa månader kommer att ha förändrats avsevärt

innan roboten är färdig. Om man däremot klarar 250 dokument parallellt, vilket motsvarar en konstant beläggning av en 1,5 Mbit/s-förbindelse, så tar det inte mer än 8 dagar att indexera hela WWW.

Hastigheten och parallell laddning och indexering är två av de problem som robot-forskare brottas med. Ett annat problem är att undvika att roboten börjar gå i cirklar. Detta kan inte undvikas eftersom WWW är ett nätverk och det inte finns några garantier för att roboten inte kommer tillbaka till samma dokument, men det gäller att ha tekniker för att upptäcka cirkel-sökningar.

Ytterligare problem är prestanda på server-sidan. Om en robot plötsligt börjar hämta flera hundra dokument från en och samma server kommer prestanda där att dramatiskt försämras med mindre glada Web-administratörer som resultat. Därför gäller det för roboten att hämta lagom mycket i taget från en och samma server.

En allt större del av WWW-information lagras idag i databaser och istället för att ha statiska sidor genereras sidorna dynamiskt när man klickar på en knapp. Detta påverkar naturligtvis söktjänster som använder sökrobotar för att bygga sina index. Det har därför föreslagits att man skall beskriva innehållet i en WWW-server, så att robotarna kan hitta den information som man vill indexera eller kategorisera.

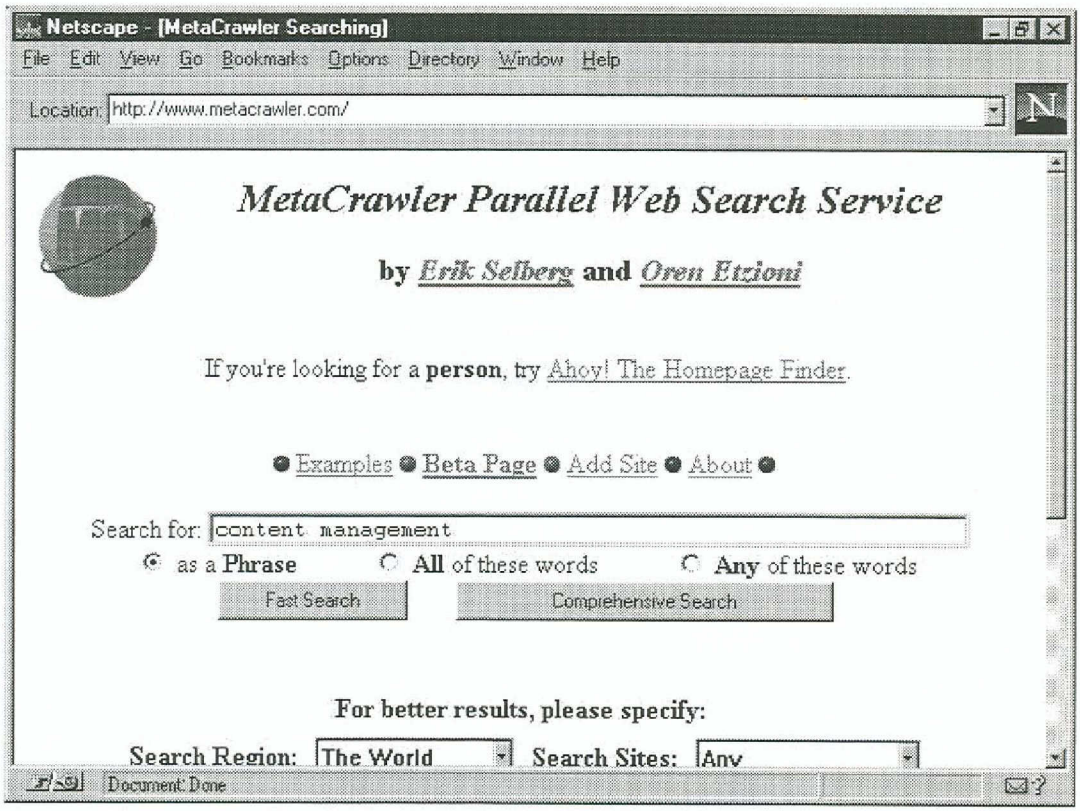
4 .2 Sökning på WWW

Allmänt kan sägas att WWW-sökning idag är helt textbaserad. Innehållsbaserad sökning (content-based retrieval) i till exempel bilder förekommer inte idag i någon WWW-söktjänst. De sökhjälpmedel som idag finns kan grovt delas in i följande grupper:

Ämneskataloger, d v s WWW-resurser organiserade enligt någon ämnesklassificering. Dessa index varierar mycket när det gäller information om olika resurser och i valet av ämnesindelning. Dessa ämnesindex byggs och administreras till stor del av mänsklig hand, Yahoo har t ex ett 30-tal personer (av vilka flera är bibliotekarier) för att administrera sitt index. Exempel på index är SUNET, Yahoo, Galaxy, The WWW Virtual Library. Till ämneskatalogernas fördelar hör att de ger en känsla av struktur ofta kombinerat med möjligheter till fritextsökning. Ett problem kan vara att de bygger på olika klassificeringsprinciper när det gäller ämnesindelning, urval och detaljerings-grad.

Sökverktyg erbjuder textsökning mot databaser baserade på index som konstruerats och upprätthålls med sökrobotar. Exempel på etablerade verktyg är AltaVista, Lycos, InfoSeek och Open Text Index. AltaVista indexerar idag ca 30 miljoner web-dokument med ett index som är ca 100GB. Fördelen med dessa verktyg att de är lättanvända och att man indexerar stora mängder dokument. Det kan däremot vara ett problem att fokusera sökningar för att undvika stora resultatmängder, samt att sökspråken kan skilja sig mellan olika verktyg.

Metasökverktyg använder sig av flera andra sökverktyg samtidigt och ev parallellt för att utföra en viss sökning. Delresultaten aggregeras, rangordnas och presenteras som ett integrerat sökresultat. Den här typen av verktyg är under utveckling, exempel på två metasökare är Meta Crawler och Savvy Search. En fördel med denna typ av verktyg är naturligtvis möjligheten att använda dem som gränssnitt mot ett flertal andra sökverk-tyg. Problem som framkommit är bl a att olika sökverktyg använder olika principer för rangordning vilket gör att det integrerade resultatet kan vara svårt att ge en meningsfull innebörd.



Figur 7. Meta Crawler är ett exempel på ett så kallat metasökningsverktyg.

Tjänster som baseras på ovanstående vektygskategorier har i allmänhet WWW som sitt sökdomän. Det finns även mer specifika söktjänster med WWW-gränssnitt mot kommersiella och/eller publika databaser. Många söktjänstföretag erbjuder också individ/företagsanpassade söktjänster, antingen baserat på en frågeprofil (som t ex InfoSeek) eller genom att användaren väljer ett antal kategorier om skall bevakas. Dessa informationstjänster inkluderar ofta även information från källor och media utanför WWW (tryckta tidningar, nyhetsbyråer m m). Ett exempel på en sådan tjänst är NewsPage Direct, där man som prenumerant får kortare sammandrag via epost av nyheter inom ett antal på förhand valda kategorier.

Allmänt kan sägas att informationssökning på WWW försvåras av bristen på gemensamma klassificeringsprinciper och enhetliga söksprak. I en tutorial om websökning pekade Roy Tennant från Berkeleyuniversitetets bibliotek ut bristen på:

- gemensam vokabulär för söksprak och resultattolkning, där t ex innebörden i begrepp som precision, relevans, närhet definieras.
- generella ämnestermer som kan användas för klassificering och nyckelord.
- struktur- eller fältbaserad sökning, motsvarande sökning i strukturerade databaser. En söktjänst som AltaVista ger ett visst stöd för detta, dock begränsat till vissa HTML-element.

- intelligenta sökfunktioner, där söksystemet har viss kunskap om det som söks, t ex vid sökning på ett visst författarnamn så kan även dennes pseudonym inkluderas. Inom IR-området används begreppet "authority control" för detta.

Det här är ju funktioner som inte är nya vare sig inom datorstödd informationssökning (IR) eller databasområdet. Det intressanta är att WWW har kommit att kombinera klassisk IR med distribuerade informationssystem och databaser. Brist på gemensamma begrepp och terminologi har ju sedan lång tid varit det stora hindret när det gäller att åstadkomma samverkan och interoperabilitet mellan informationssystem.

I sådana här sammanhang brukar ofta standardisering anföras som en lösning. Det påpekades dock på konferensen att existerande standarder för klassificering och sökning inom IR och biblioteksområdet är historiskt belastade och inte riktigt passar för att strukturera det heterogena informationsinnehållet i dagens WWW-servrar.

5 Webutveckling

Överhuvudtaget var intresset på konferensen stort för design- och utvecklingsfrågor, ett tecken på att WWW har fått status som en plattform för systemutveckling. WWW-utveckling täcker idag ett mycket brett spektrum där det ingår allt ifrån utseendet på hemsidor och användbarhetsanalys till programvarukomponenter, databasstöd och systemadministration.

En av konferensens presentationer som fick pris i klassen "bästa tutorial", hölls av Jacob Nielsen, SUN, med titeln "Designing and Maintaining a Highly Usable Site". Nielsen, känd som forskare inom gränssnitt och användbarhet, beskrev erfarenheterna från gränssnittsdesignen för SUNs egna hemsidor (www.sun.com), där ett omfattande arbete lades ned på utformningen av ikoner, knappar och länkstruktur.

Underhåll och vidareutveckling av WWW-servrar kräver, som vi tidigare konstaterat, allt större resurser. En mängd olika verktyg börjar nu också att dyka upp avsedda för webserveradministration. Exempel på detta är olika produkter för att administrera referenser mellan WWW-dokument och skapa buffertar (caching) av externa WWW-serverar.

5.1 Databasstöd

Många av forskarpresentationerna handlade om olika aspekter av databasteknik. Allmänt kan sägas att de flesta av problemen som nu tas upp inom ramen för WWW är gamla och välkända (om dock ej lösta) från tidigare forskning och tillämpning inom databasområdet. Områden som togs upp under konferensen var bl a,

- Objektidentifiering och referentiell integritet
- Distribuering, caching och replikering
- Datastrukturering och dynamisk generering av HTML-objekt.

Konferensens enda svenska forskningsbidrag presenterades av Erik Sandevall, Linköpings Universitet, som beskrev ett verktyg för att strukturera och administrera HTML-textfiler baserat på en enkel objektmodell (http://vir.liu.se/brs). Man använder här ett eget språk för att definiera en uppsättning objekttyper som verktyget använder för att strukturera och generera HTML-dokument. Någon ytterligare databasfunktionalitet i konventionell mening innehöll dock inte detta verktyg.

Några presentationer beskrev exempel på hur WWW-arkitekturen skulle kunna fås att samverka med objektorienterade system. Relativt grundläggande principer för namngivning och referenshantering relaterades till WWW, så också aktiva objekt.

Om vi ser till produkter på databasområdet, så framstår Illustra från Informix (se ovan) som ett mycket intressant verktyg för WWW-tillämpningar med databasstöd, eller content management. Illustra är ett exempel på en s k hybriddatabas, eller objekt/relations- (OR) databas, som förenar relationsdatabasens flexibilitet med objektmodellens struktureringsegenskaper.

Michael Stonebraker, med ett förflutet som databasforskare bl a inom Ingress och Postgress-projekten, startade företaget Illustra som dock numera ingår i Informix. Stonebraker är teknisk strateg på Informix och arbetar med att integrera Illustra med Informix' databasprodukt. Resultatet blir Informix Universal Server som beräknas att släppas hösten 96, som bl a skall stödja Java, Visual Basic och Embedded SQL.

Hybriddatabaser är ingen ny företeelse, vissa tidigare system har dock varit objekt-orienterade påbyggnader (eller "wrappers") mot en konventionell relationsdatabas, vilket bl a får konsekvenser för prestanda. Illustra är dock, enligt Stonebraker, en mer komplett OR-databas (dock med en intern lagringsmodul från en relationsdatabas) vilket skall borga för både prestanda och flexibilitet. För en översikt av hybriddatabaser hänvisas till en separat SISU-rapport (SISU-rapport 96-11).

Intressant är de möjligheter som en objektorienterad databas ger, att förutom multi-mediaobjekt, även kunna lagra aktiva objekt (t ex programkomponenter som applets) i databasen. Det förefaller som den objekt-orienterade databasen till slut har hitta ett större tillämpningsområde i content management för WWW.

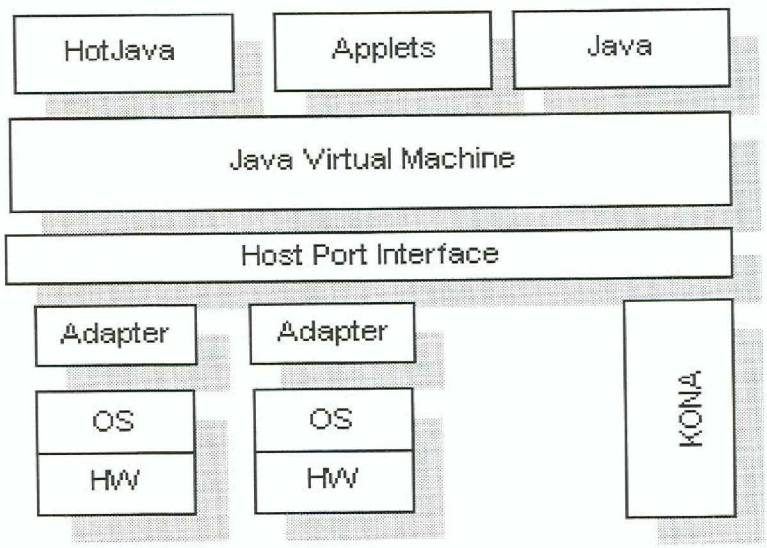
5 .2 WWW-komponenter: ActiveX vs. Java

Flera av de tekniskt inriktade panelerna och industripresentationerna handlade om valet mellan Java och ActiveX som miljö för komponentprogram på WWW.

Microsoft propagerade på konferensen för sin miljö ActiveX som i princip består av WWW-läsaren Internet Explorer samt WWW-anpassade OCX-komponenter, medan SUN och de flesta av de akademiska konferensdeltagarna såg Java som den bästa komponentmiljön för WWW.

Portabilitet och plattformsoberoende framstår som det dominerande argumentet i valet mellan Java och ActiveX. Många menar att detta gör Java till ett självklart val, Microsoft hävdade dock att man kommer att uppnå plattformsoberoende genom att utnyttja objektgränssnittet COM/DCOM mellan komponenter baserade på Java eller ActiveX, och andra tillämpningar t ex utvecklade i C++.

Javas pappa, James Gosling från SUN, höll ett mycket uppskattat föredrag om erfarenheterna från och planerna för Java. Gosling, som har en viss guru-status, utvecklade bl a Emacs för UNIX på 80-talet och har arbetat med gränssnittssystem som News (på konferensen hävdade någon att WWW är 90-talets Emacs, hur nu det skall tolkas). Gosling hävdar att de tillämpningar som idag utvecklas med Java spänner över hela fältet från de ursprungligen avsedda små gränssnittskomponenterna (applets) till större tillämpningar där Java använts som ett alternativ till C++.



Figur 8. Principiell arkitektur för Java enligt Gosling.

SUNs satsning på vidareutveckling av Java-miljön framstår som mycket intressant, bl a räknade Gosling upp en mängd nya APIer. Det skall finnas stöd för s k "servlets", motsvarigheten till applets men på serversidan som bl a kan ersätta de idag dominerande CGI-programmen. Andra APIer skall stödja kommunikation mellan applets och gränssnitt mot protokoll för elektronisk handel såsom betalningssystem. För att bättra på Javas något skamfilade rykte när det gäller säkerhet, skall det även vara möjligt att skapa signerade applets, för att kunna verifiera identitet och ursprung hos externt skapade programkomponenter.

Java-anhängare menar att språket och miljön kring det, i kombination med spridningen av WWW, möjliggör en ny typ av programvaruproduktion och distribution. Man jämför här med tidigare stordatorarkitekturer där all exekvering och administration var lokaliserad till server-sidan. Med client/server-arkitekturer kunde exekvering distribueras, samtidigt som detta medförde administration på såväl klient som serversidan. Java, menar man, ger det bästa av dessa två världar.

Enligt SUN är Java "completely platform independent" och man strävar efter att i så hög grad som möjligt reducera beroendet av operativsystemet på de datorer som idag används som klienter på WWW. Som ett led i denna strävan nämnde Gosling ett slags operativsystem avsett för Java kallat KONA. Enligt Gosling är KONA egentligen inte att betrakta som operativsystem i konventionell mening då det inte skulle ha något synligt API och sakna systemanrop. En tanke med KONA var att det skulle utgöra run-time miljö i specialiserade nätdatorer som den s k nätverks-PCn, eller olika mobila enheter som handdatorer (PDAer). Nätverks-PCn, idén om en förenklad dator enbart avsedd för nätverksåtkomst och riktad mot hemmamarknaden, har i debatten setts som alternativet persondatorn. Få tror dock på allvar idag att en sådan nätdator kan ersätta den generella persondatorn som WWW-klient vare sig funktionellt eller prismässigt.

Microsoft ser Java som en av komponenterna vilka kan ingå i tillämpningar som utvecklas inom ramen för ActiveX-miljön. Man presenterade också sitt prototypsystem "Nashville", i vilket Windows integrerats helt med WWW-klientfunktionalitet (eller omvänt). Denna lösning framstod som mycket tilltalande, trots att systemet kraschade under demonstrationen.

6 Sammanfattning

Nedan följer en sammanställning av de produkter, tjänster, prototyper och projekt som diskuterats i denna konferensrapport och/eller förekom på konferensen. Delar av konferensdokumentationen kan nås på URL: http://www5conf.inria.fr.

Produkt/tjänst/prototyp	*Kommentar*	*URL*
ActiveX (Microsoft)	Microsofts komponentarkitektur för WWW-tillämpningar.	www.microsoft.com
Alta Vista	Den idag största och mest omfattande söktjänsten för WWW-information.	www.altavista.com
CyberCoins	Digitala pengar för betalning över Internet.	www.cybercash.com
Ecash	Digitala pengar för betalning över Internet.	www.digicash.com
Illustra (Informix)	OR-databas med stöd för multimediaobjekt och WWW-gränssnitt	www.illustra.com
Incontext Spider (Incontext) – HTML-editor för Windows	HTML-editor som kan integreras med blädrare som Netscape, Explorer, Mosaic.	www.incontext.com
Incontext Webanalyzer (Incontext)	Verktyg för att analysera strukturen hos WWW-servrar såsom länkstrukturen och typer av resurser. Caching av externa servrar.	www.incontext.com
InfoSeek	Kommersiell söktjänst.	www.infoseek.com
Java (SUN)	SUNs komponentarkitektur för WWW-tillämpningar.	www.sun.com
JEPI	Joint Electronic Payment Intitiative. Protokoll för förhandling om olika betalningsystem. Samarbete mellan W3C-konsortiet och CommerceNet.	www.w3.org/pub/WWW/Payments/
Lycos	Populär sökmotor med tillhörande indexeringsrobot.	www.lycos.com
MetaCrawler	Metasökningsverktyg som kombinerar sökning via flera olika söktjänster.	www.metacrawler.com
Millicent	Digitala pengar för betalning över Internet. Specifik avsett för små belopp.	www.research.digital.com/SRC/millicent/
PEP	Protocol Extension Protocol. Protokoll som möjliggör standardiserade utökningar http-protokollet. Används både för JEPI och PICS.	www.w3.org/pub/WWW/TR/WD-http-pep.html
PICS	Platform for Internet Content Selection. Kommande standard för gradering av WWW-sidor utifrån till exempel känslighetssynpunkt.	www.w3.org/pub/WWW/PICS/

Savvy Search	Metasökningsverktyg som kombinerar sökning via flera olika söktjänster.	www.cs.colostate.edu/~dreiling/smartform.html
SET	Secure Electronic Transactions. Standard för betalning med kreditkort.	www.visa.com
UPP	Universal Payment Preamble. Protokoll för förhandling om betalprotokoll. Kommer att användas i JEPI-projektet. Från CyberCash.	www.commerce.net/public/ElectronicCommerce/ansi-epay/UPP.txt
WWW-96	Den årliga forskarkonferensen med inriktning på World Wide Web och dess användning.	www5conf.inria.fr
Yahoo	Den mest använda katalogtjänsten på WWW.	www.yahoo.com