neofonie
INFORMATIONSARCHITEKTUR

DFN Science-To-Science:
Suche auf Peer-to-Peer Basis für die Wissenschaftliche Gemeinschaft

12.05.2003@fu-berlin.de

(Colloquium: XML Clearinghouse für Berlin und Brandenburg)
präsentiert von Ronald Wertlen
rrrw@neofonie.de

Agenda

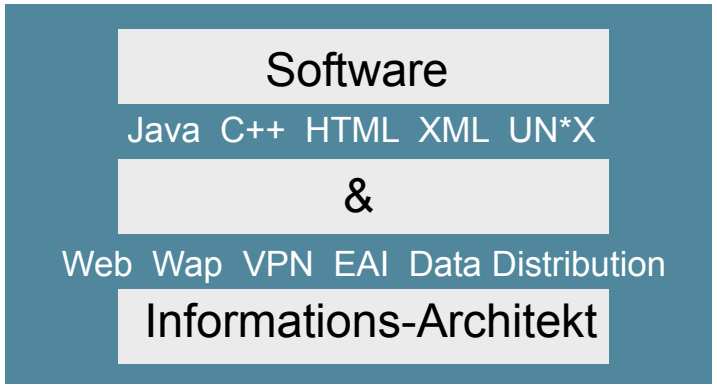
- Einführung
  - Wer sind wir?
  - Was ist P2P?
- Übersicht P2P Suche
  - Neurogrid
  - EDUTELLA
  - JXTA Search
  - Platzierung S2S
- Projektübersicht
  - DFN S2S Ziele
  - Was bringt S2S?
- Grundlegende Technik
  - JXTA
  - JXTA Search
  - neofonie search
- XML in S2S
- Die Zukunft
  - Probleme
  - Projekterweiterungen
- Beta-Tests
  - Einladung

Vorstellung: Kompetenzen

Beratung
herstellerunabhängig
bedarfsgerecht
kompetent

Dienstleistung
Entwurf & Design
Maßgeschneiderte Entwicklung
Integration
Wartung, Pflege & Betrieb

Produkte
Information Gathering/Retrieval
(Content Management)
Standard-Komponenten



Forschung
Information Gathering/Retrieval
Peer-to-Peer
C++-Servlets
XML

Entwicklung
Portale
Online-Dienste
Knowledge Management

neofonie – Forschung & Entwicklung

- neofonie kombiniert…
  - Kompetenz- und Entwicklungsvorsprung durch Forschung, finanziert aus dem operativen Geschäft sowie durch regionale und europäische Förderprogramme
  - Know-How und Erfahrung aus der Durchführung kommerzieller Großprojekte
  - Full Service für den Kunden: Beratung, Entwurf, Entwicklung, Installation, Wartung und Weiterentwicklung
- Forschung
  - neofonie 1998 aus Technischer Universität Berlin ausgegründet
  - Untersuchung neuester Technologien und Softwarearchitekturen. Use Cases und Praxistauglichkeit.
  - Bisher fünf erfolgreich durchgeführte Forschungsprojekte. Förderung durch EU, DFN, Senat Berlin.
  - Fokus 2003/2004: Information Retrieval, EAI und P2P basiertes, Workflow-unterstütztes Knowledge Management
- Beratung und Entwicklung
- sowie Produkte

Meilensteine

- 1998: Gründung der infonie GmbH mit Sitz in Berlin
- 1998: Internet-Suchdienst Fireball; Printmedien-Suchdienst Paperball
- 1999: Kostenpflichtige ASP-Suchplattform für Unternehmensgruppe
- 1999: Ausschreibungsdienst Baubranche (ibau)
- 2000: WAP-Suchmaschine (e-plus); Online-Handelsplattform (Cnited)
- 2000: Suche und Katalog (bund.de)
- 2000: Launch Compuserve-Portal
- 2001: Konsortium <xmlcity:berlin>
- 2001: DFN S2S (Peer-to-Peer-Suche)
- 2002: Umbenennung in neofonie GmbH
- 2002: Produktfamilie neofonie:search
- 2002: Zusammenführung berlin.de und berlinonline.de
- 2003: Launch AOL Kleinanzeigen

Referenzen

NET SEARCH	NETWORK SEARCH	SITE SEARCH	INTRANET SEARCH	CONTENT MANAGEMENT	OTHER
FIREBALL Netsearch GmbH	BertelsmannSpringer	Electronic Media Sales	T-Nova	AOL.de	LYCOS
Dalim	Die Bundesregierung	n-tv	Pixelpark	CompuServe	BerlinOnline Berlin auf einen Klick!
e-plus	PrimusOnline.de	ebay		Netscape	Cnited.com
baynet.de		TOMORROW INTERNET AG		Electronic Media Sales	DFN DFN-Verein
		ricardo		ibau	Bundesministerium für Bildung und Forschung

P2P - Ein paar Definitionen

- "Peer" bezeichnet einen Rechner der
  - gleichzeitig Client und Server ist - darf Verbindungen annehmen sowie initiieren
  - anderen "Peer" Rechnern aufgrund ähnlicher Funktionalität gleichgestellt ist
  - oft am Rande des Netzwerkes ist: hat keine feste IP Addresse - befindet sich hinter einer Firewall/NAT
  - nicht immer online ist

- Merkmale von P2P Netzwerken:
  - Sie formen sich „von alleine" (ohne administrativen Aufwand) da die software einfach zu installieren und bedienen ist
  - Durch verbreitete Nutzung des Netzwerkes steigt der Wert des Netzwerkes
  - Ungenutzte Ressourcen (Speicher, Festplatte, Netzwerkbandbreite) werden eingespannt

P2P Umfeld Übersicht

- Das Thema drang an die Öffentlichkeit durch Napster… Datenaustausch mit einfacher Suche nach Dateinamen.
  - Napster kombinierte dezentrale mit zentralen Elementen
  - Vereinfachte den Austausch von Dateien
  - Je mehr Teilnehmer desto besser die Auswahl für den Nutzer
  - Gnutella + Derivate (komplett dezentral - wer sollte da verklagt werden?)

- P2P wurde ein Forschungsthema. Aus Erfahrungen mit den ersten grossen P2P Netzwerken liessen sich folgende Themenbereiche feststellen:
  - Routing - wie kommt man am besten von Punkt A nach B? (PASTRY, HyperCUP)
  - Sicherheitsmodelle - ohne zentrale Zertifizierungsauthorität, wie kann man ein sicheres Netzwerk aufbauen? (Mojonation - reputation)
  - Anonymität durch P2P (Freenet)
  - Dezentrale Metadaten und Suche - hier ordnet sich S2S ein

P2P Umfeld Übersicht (2)

- Verwandte P2P Anwendungsbeispiele in der Forschung …
  - JXTA Search - eine Architektur und XML-Protokolle die von einen ad-hoc Suchnetzwerk als Standard benutzt werden können.
  - EDUTELLA - Ein semantisch korrektes Metadaten und Such System, benutzt die JXTA Search Architektur, Ziele sind aber sehr unterschiedlich.
  - Neurogrid - Ad-hoc Methode Metadaten zu Suchen welche vorsieht dass ein Peer vom Verhalten des Peer-Nutzers und der Peer-Nachbarn lernt um so Suchergebnisse zu verbessern.

- … und in der Praxis
  - Groove - software zur P2P Kollaboration ermöglicht die ad-hoc Erstellung von Gruppen, Datenaustausch, Chat, Kollaboratives Zeichnen usw.
  - Grub - wie SETI soll man sich einen Screensaver herunterladen der durch das Web spidert nach angaben eines zentralen Servers. Der Nutzer kann das Verhalten des Software direkt nicht beeinflussen.

Neurogrid Eckpunkte

- Funktioniert nach sehr anthropologischen Prinzipien. Peers kennen nur die Peers in ihrer Umgebung. Wenn jemand etwas sucht dann beantwortet ein Peer selbst oder leitet die Frage weiter.
- Jeder darf auf seine Art und Weise antworten.
- Lernen ist hier wichtig: mit der Zeit lernt ein Peer seine Nachbarn kennen und kann gezielt Fragen weiter leiten.
- Gelernt wird auch die Nutzer Reaktion. Mit der Zeit baut die Software ein Profil des Nutzers und benutzt dieses, um Suchanfragen zu präzisieren und Ergebnisse besser zu filtern.
- Flooding ist der anfänglich verwendete Routing Algorithmus, der mit der Zeit durch einen semantischen Algorithmus ersetzt wird.

JXTA Search Eckpunkte

- Infrasearch, gekauft von SUN Microsystems.
- Architektur Hub, Provider und Consumer
- Netzwerk Aufbau erfolgt in dem ein Provider sich bei einem Hub mittels einer Registrierungsnachricht anmeldet.
- Suchanfragen kommen von einem Consumer und werden vom Hub zum passenden Provider weitergeleitet. Der Provider kann auch direkt angesprochen werden.
- Das Protokoll definiert wie Suchanfragen, Registrierungen und Suchergebnisse in XML aussehen. Es darf außer einem XML Umschlag immer auch beliebiges XML eingebettet werden.

EDUTELLA Eckpunkte

- LearningLab.de, Niedersachsen + Stanford, Massachusetts
- Basiert ursprünglich auf JXTA Search - die Architektur bleibt erhalten aber die Semantik ändert sich.
- Ziel ist es Lern- und Lehrinhalte in RDF ausgedrückt suchbar zu machen
- Weitere Ziele sind Ontologie-Transformatoren und weitere Semantisch-Korrekte Prozesse in ein P2P Netzwerk einzubauen.
- Erfordert einen hohen Verwaltungsaufwand im Netzwerk, sowie hohe Aufwände für den Nutzer (RDF oder zumindest XML Schemata werden verlangt).

DFN S2S

- DFN S2S implementiert die JXTA Search Protokolle und Architektur, hat Berührungspunkte mit Neurogrid im vorgehen.
- JXTA Search ist aber nicht fertig - seit 2001 (Projektantrag) wird angeblich nicht mehr daran gearbeitet, außer von uns. Unsere Resultate sind (noch) nicht veröffentlicht.
- Wir haben JXTA Search um die Funktion zum Download erweitert und im Rahmen einer Diplomarbeit wird daran gearbeitet Hubs verlinken zu können.
- S2S wird dahingehend erweitert, dass Provider direkt miteinander Terme austauschen können, um die Relevanz zu steigern.
- Viele weitere Features beeinflussen die Funktionalität des Netzwerkes, obwohl oft auf eine protokoll-neutrale Art (z.B. Relevanz-Sortierungstechnologie)

Ziele von DFN S2S

- Das Ziel von DFN S2S ist es, Wissenschaftlern eine einfache Möglichkeit für den Wissensaustausch anzubieten. Der Wissensaustausch erfolgt so:
  - Wissenschaftler geben an, welche Dokumente sie suchbar machen wollen. Die Dokumente können auf der eigenen Festplatte aber auch auf einen FTP- oder Webserver liegen.
  - Wissenschaftler können per Volltext- und Feld-Suche im Netzwerk recherchieren
  - Wenn ein interessantes Dokument gefunden worden ist, kann es per S2S auch heruntergeladen werden (falls erlaubt).

Was bringt das eigentlich?

- 1. Bessere Suche!
- 2. Ein Wissenschaftler kann seine eigene Dokumente bzw Dokumentbestände durchsuchbar machen, für sich selbst!
- 3. Wissenschaftler können sich in Gemeinschaften (Communities) zusammenschliessen mittels der Queryspaces.

Herkömmliche Recherche im Internet

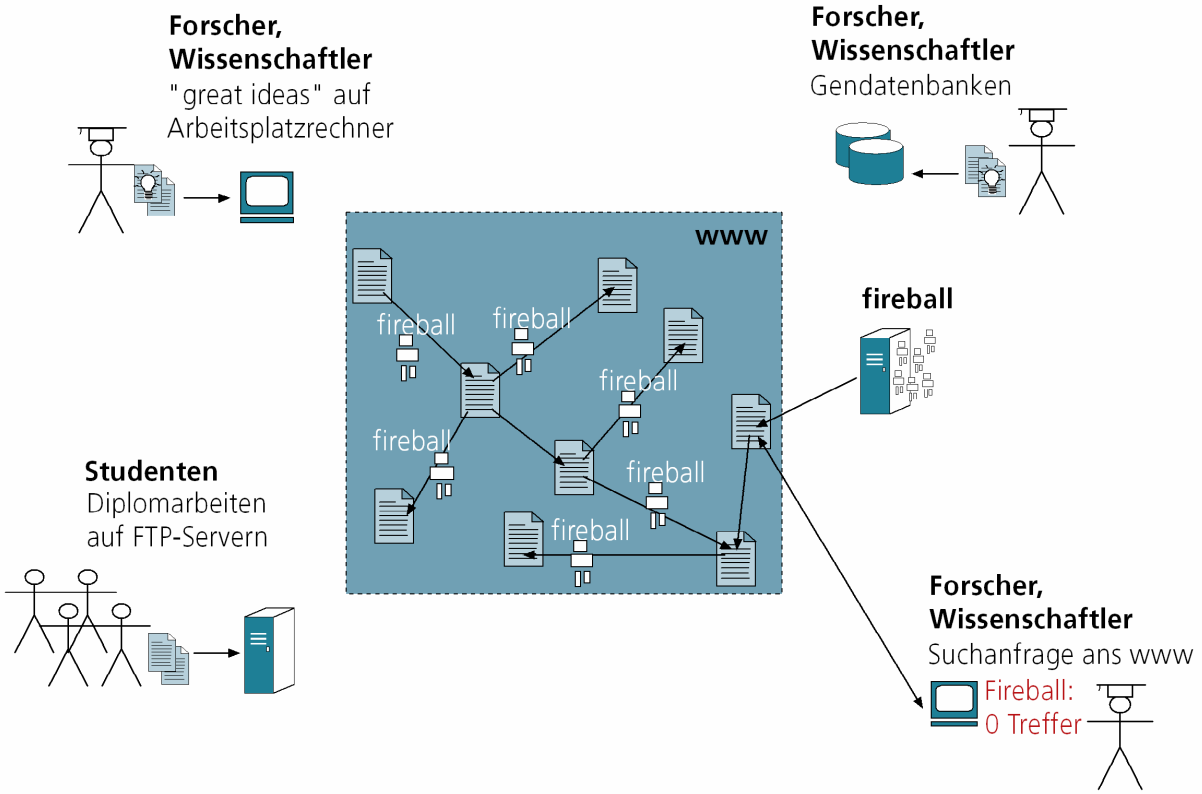
Fireball: Web Search

DFN Science-To-Science

Zentrale Sammlung,
Indexierung, Suche

Nachteile:
- skaliert schlecht
- Aktualität mangelhaft
- Bestand: nur WWW
- zentralisiert

Vorteile:
- Verfügbarkeit



S2S Recherche im „Deep Web“

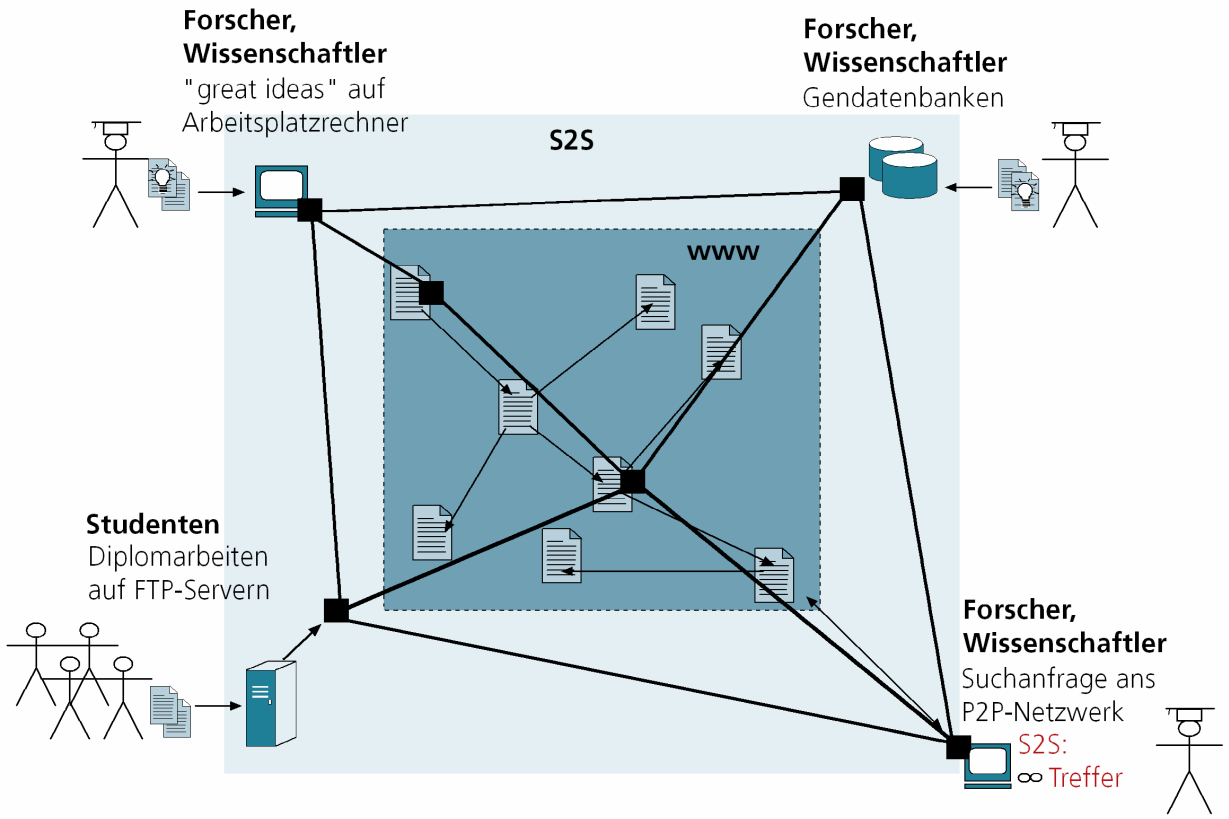
S2S: Deep Search

DFN Science-To-Science

Verteilte Indexierung, Suche

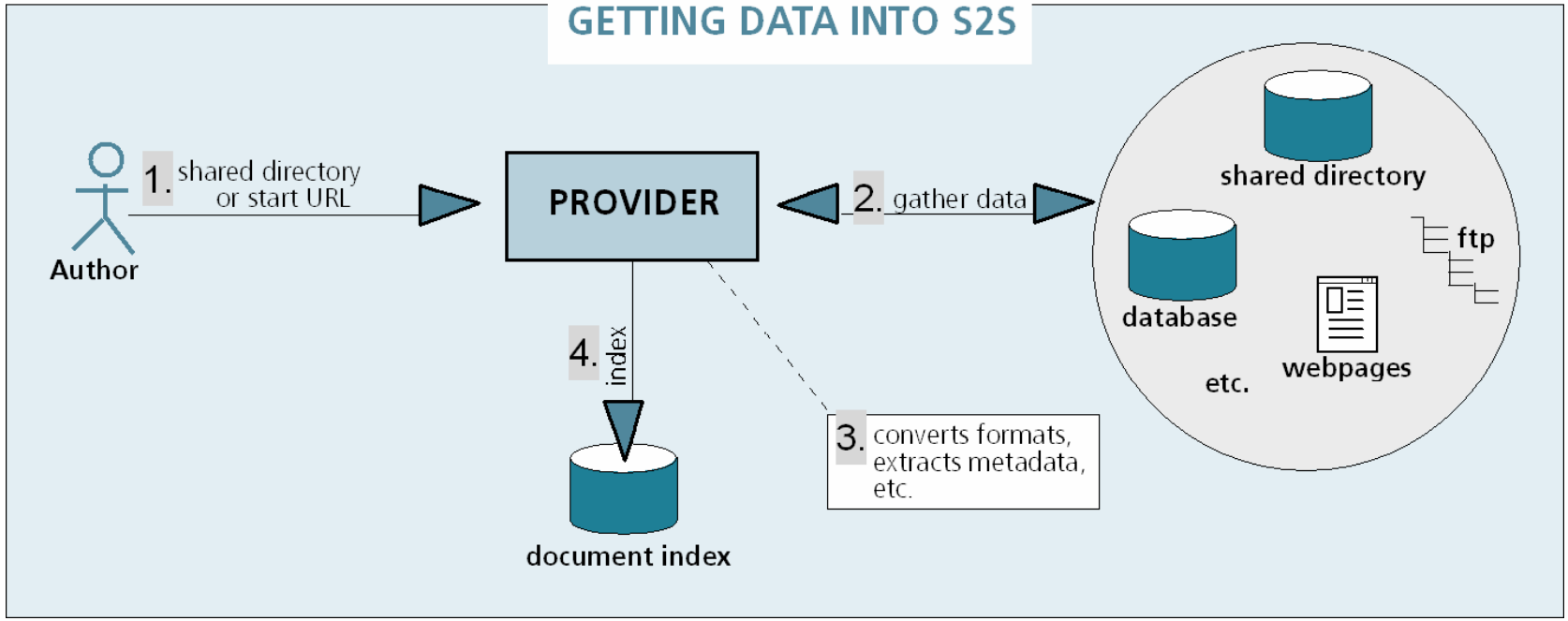
Vorteile:
- skaliert
- aktuell
- Bestand im gesamten Internet
- noch unerforscht (weiteres)

Nachteile:
- noch unerforscht (vermutl. Verfügbarkeit)



Eigene Dokumente Durchsuchen

- So einfach geht das…
- 1. User gibt eine URL an… S2S macht den Rest



Queryspaces

- Mit der Auswahl eines Queryspace oder mit der Definition eines neuen Queryspace können Wissenschaftler sich in Gruppen zusammenschliessen.
- Das geht dadurch, dass Suchanfragen gezielt an bestimmte Queryspace gestellt werden können. So bekommt der Nutzer nur Treffer aus einer bestimmten Gruppe von Providern.
- Queryspaces werden durch einen Namen (URI) gekennzeichnet und sie werden inhaltlich von ihrer Registrierung definiert (ein grundsätzlicher Unterschied zum Edutella Ansatz)
- Es ist nicht geplant (obwohl technisch einfach) geschlossene Gruppen in S2S zu erlauben. Eine geschlossene Gruppe unterscheidet sich dadurch, dass nur Suchanfragen, die gezielt an den Queryspace dieser Gruppe gerichtet wurden, beantwortet werden. Zur Zeit werden auch Inhaltlich passende Fragen beantwortet.

Grundlegende Technik

- Die S2S Software besteht aus drei funktionalen Teilen:

1. Die JXTA Plattform wird benutzt, um Peers auf unkomplizierter Art zu verbinden. Da die Plattform schon unterschiedliche P2P Probleme löst und Modelle für die Funktionsweise vorgibt, passt es zu diesem Projekt, da wir uns auf unser Problemgebiet - die Suche - konzentrieren können.

2. Die von neofonie entwickelte Suchsoftware „neofonie search“, die Hauptfunktionen des Providers implementiert (Datensammlung, Metadaten Anreicherung und Indexierung). Das Indexieren von XML spielt auch im Hub eine Schlüsselrolle.

3. Die JXTA Search Spezifikation stellt wie wir schon gesehen haben, den Grundriss eines P2P-Netzwerkes dar. Die S2S Software implementiert diese Ideen und verfeinert sie.

JXTA

- Protokoll Spezifikation mit Referenzimplementierung, in einem Open Source Gemeinschaftsprojekt organisiert, welches dennoch von SUN Microsystems unterstützt wird. Siehe http://www.jxta.org

- Protokolle sind speziell für P2P entwickelt worden, sie sind asynchron und unzuverlässig.
  - In einem P2P Netzwerk kann man Antwortzeiten nicht garantieren, es kann auch sein, dass keine Antwort kommt
  - Antworten dürfen länger dauern aber sind dafür immer aktuell

- Erlaubt die Bildung von Gruppen, um Nachrichtenflüsse zu beschränken

- Erlaubt die Entdeckung von aktuellen Ressourcen im Netzwerk anhand von Werbungen (Advertisements)

Warum JXTA für DFN S2S?

- Es ist ein sehr aktivstes Open Source Projekt, wurde über 1 Million Mal heruntergeladen
- Stabile API
- Stellt die notwendige Funktionalität zur Verfügung
- Zielplattformen sind nicht festgelegt. Es soll unter diversen Hardware und Software Umgebungen funktionieren. (Implementiert in Java, C++ und C# / .NET auf TCP, HTTP und UDP kommt noch)
- XML – basiert bedeutet sehr leichte Integration mit neofonie:search …

neofonie search:engine

- Volltext-Suchmaschine mit nativer XML-Unterstützung und zum Patent angemeldeten Such- und Sortierverfahren: Berücksichtigung von textuellem Inhalt, Struktur, Verlinkung sowie Optimierung durch Auswertung des Nutzer-Verhaltens.

- Basis ist ein hocheffizient organisierter Index. Bewertungskriterien sind in der Suche einstellbar, Kombinationsmöglichkeiten für Suchoperatoren sind umfangreich.

- Hierarchische Felder werden berücksichtigt. Die Präsentationsschicht erlaubt maximale Freiheiten bei der Gestaltung und Vorverarbeitung von Suchanfragen. Die Nachbearbeitung von Suchergebnissen wird unterstützt.

- Wird eingesetzt als XML-DB in DFN S2S im Hub und im Provider

neofonie search:robot

- Informationen in allen Winkeln des Netzes werden gefunden, analysiert, der gewünschten Anwendung (z.B.: einer IR-Engine oder auch einem CMS) geliefert und nach einer konfigurierbaren Strategie aktualisiert.

- Dokumente im WWW, im Intranet und auch auf File-Servern werden betrachtet, ein hoher Durchsatz garantiert, externe Datenquellen durch eine adaptive Begrenzung geschont und selbstverständlich alle internationalen Vereinbarungen (robots.txt) berücksichtigt.

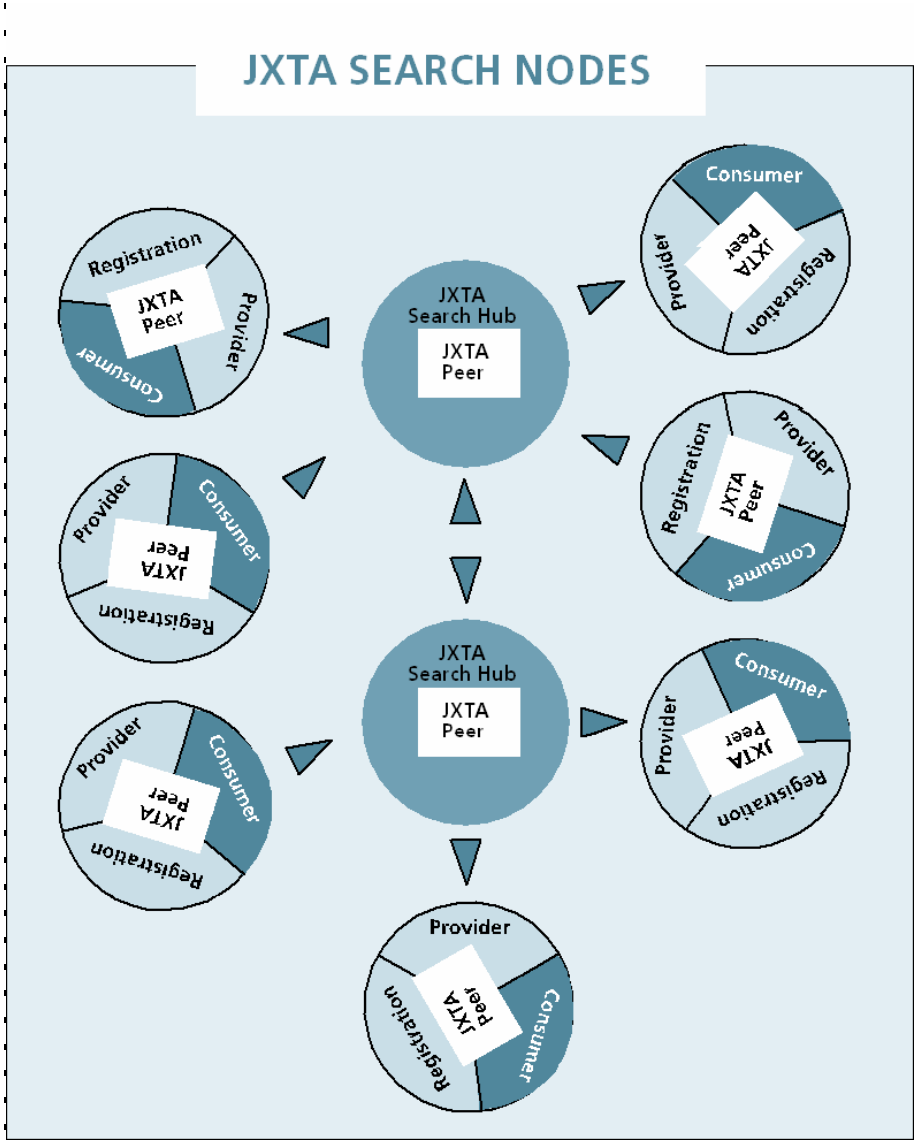
- Dokumentenprozessoren für die häufigsten Anwendungsszenarien sind bereits integriert. Automatische Format-Konvertierungen nach XML und eine musterbasierte Content-Extraktion sind möglich (siehe :purifier).

- Wird im DFN S2S Provider zum Herunterladen von Ressourcen und für die Vereinheitlichung der Formate eingesetzt.

S2S Software

- Implementiert:
  - Nachrichten-Formate nach JXTA Search.
  - Abläufe, die von JXTA Search spezifiziert werden in einem flexiblen Java Rahmenwerk welches Nachrichtenflüsse dynamisch (zur Laufzeit) anhand von Funktionalen Bausteinen (Processors, Processor Managers, und Dispatcher) aufbaut.
  - Web-Anwendung für Internetzugang
- Die Asynchronität eines P2P Netzwerkes wird automatisch vom Rahmenwerk behandelt
- Ermöglicht wird vom Rahmenwerk auch, dass je nach Konfiguration Hub, Provider oder Consumer mit Java Swing GUI enstehen.

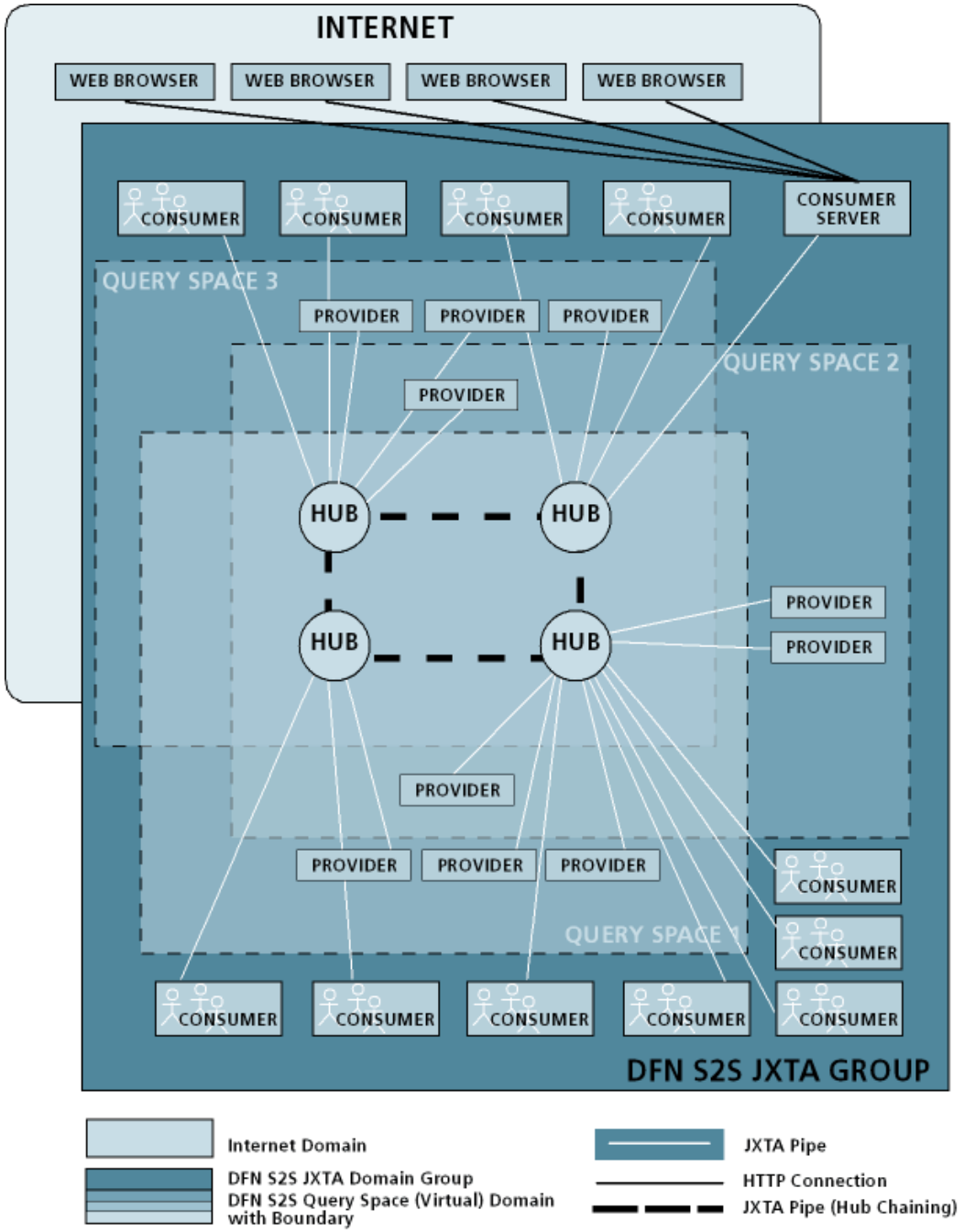
JXTA Search Architektur



- Provider bieten Daten an
- Provider müssen sich registrieren, um Suchanfragen zugeschickt zu bekommen.
- Hubs sind „super-peers“ im Netzwerk. Sie verwalten einen Cache, Registrierungen und andere Informationen über Provider und leiten Suchanfragen semantisch weiter.
- Consumer suchen nach Dokumenten und können sie auch herunterladen.

Architektur mit Queryspaces

- Grafik welche den S2S Inhaltlichen Raum dem Internet gegenüberstellt.
- Auch dargestellt ist die Aufteilung des S2S Netzwerkes in Gruppen durch Queryspaces.
- Consumer gehören keinem Queryspace an und dürfen ihre Suchanfragen an beliebige Queryspaces richten. Dafür muss die URI bekant sein.



Features

- Caching Mechanismus um langsame Peers auch in die Ergebnisse mit einzubeziehen.
- Provider Verhaltensdaten werden im Hub gesammelt und ermöglichen Relevanzbewertungen anhand von Anbindungsdauer und Verbesserung des semantischen Routing anhand von Suchergebnissen.
- Unterstützt automatische Query Prädikate Generierung mittels „globalem“ TFIDF
- MIDP Schnittstelle mit abgespicktem JXTA Search Format für mobilen Zugang
- Kollaboratives Filtern – Nutzer-Klicks werden gezählt und beeinflussen wie Treffer von diesem Provider eingeordnet werden. Dieses dient auch als „anti-spam“ Mittel.
- In Java implementiert.

XML in S2S (1)

- 1. Dient als die interne Repräsentation aller Dokumente im System.
- 2. Ist die Grundlage der JXTA Nachrichten die im Netzwerk fließen.
- 3. Dient zum Ausdruck folgender inhaltlichen Elemente, nach JXTA Search:
  - Suchanfragen (Requests), Registrierungen (Registrations), Suchergebnisse (Responses)
- 4. XSLT transformiert Suchanfragen in einem Format welches leicht mit Registrierungen verglichen werden kann, so dass Suchanfragen gezielt weitergeleitet werden können
- 5. XSLT transformiert Suchergebnisse in Registrierungen, die die eigentlichen Registrierungen unterstützen
- 6. XSLT transformiert Suchanfragen so, dass sie vom :engine (XML-Index) bearbeitet werden können.

XML in S2S (2)

- Vorteile:
  - Interoperabilität - hinzufügen von eigenen Knoten usw.
  - Erweiterbarkeit - Aussagekräftige, lesbare Nachrichten zirkulieren im System, die erlauben mehr Funktionalität einfach einzubauen.
  - Erlaubt das System strukturierte und unstrukturierte Inhalte einheitlich zu bearbeiten.
  - XML Datenbank als einheitliche Informations-Repository, auf Administrationsebene und auch auf Inhaltsebene. -> Komponentenkomplexität verringert.

- Nachteile:
  - Die relative Langsamkeit der internen Verarbeitung
  - Extra Sicherheitsschicht muss darauf aufgebaut werden - aber mindestens ist man sich klar darüber wie sicherheitsschwach die Nachrichten sind - etwas was einem bei Proprietäre Systemen nicht immer so klar ist.

Problemengebiete

- Wie geht man mit unerwünschten Inhalten („spam”) um?
- Wie verhindert man, dass geschützte Dokumente bzw. Dokumente zu welchem die Nutzer keine Rechte behalten, im Netzwerk landen?
- Gibt es genug Bedarf oder öffentliche Dokumente für solch ein vorgehen?
- Inwiefern skalieren JXTA Search und die S2S Erweiterungen?

Mögliche Erweiterungen

- Spam: Unterscheidung zwischen bekannten und unbekannten Providern
- Sicherheitsschicht: Verschlüsselung zwischen den Peers
- Unterstützung für bezahlte Inhalte (Pay-per-View)
- Erweiterung des Angebots, z.B. Anschluss zum ELENA Netzwerk
- Unterstützung für XML-Schema definierte Queryspaces und entsprechende Prüfung von Registrierungen
- Hub-Verkettung, um besser zu Skalieren
- „Voting“ zum Ausschluss von Providern

Zeitplan

- Zur Zeit werden Beta-Tester gesucht
- Erste Beta-Installationen sind im Juli geplant
- Ab Anfang Januar 2004 ist der öffentliche Pilotbetrieb geplant

Diskussion

- Fragen, Anregungen und Kritik sind erwünscht!
- Vielen Dank