neofonie

THE SPECIALIST FOR YOUR
INFORMATION ARCHITECTURE

DFN S2S: Peer-to-Peer Scientific Research

20/5/2003@terena.hr

(for the TERENA conference, Zagreb, 2003)
presented by Ronald Wertlen
rrrw@neofonie.de

Contents

- neofonie GmbH: Who we are and what we do.
- Definitions, goals and benefits of project **DFN Science-to-Science** (S2S)
- What can one do using S2S?
- Underlying technology
- Expectations for the use of the network
- Comparison with other projects, project roadmap
- Feedback

neofonie GmbH

- A dot com that survived by consistently producing innovative software solutions to today's pressing problems.
- Software products
  - neofonie:search suite – free trial online (neofonie:search-express)
  - neofonie:content suite
- Public services
  - www.fireball.de Germany's first and still most popular local search engine
  - www.paperball.de index of German news articles, updated daily
- Professional services
  - From consulting to development to maintenance
  - Our customers include
    - AOL.de, Bertelsmann AG, T-Nova, more

neofonie GmbH

- Application Services Provision
  - Affordable search for everyone
- :suchexpress = “express search”
- Constructs an index of your web pages at a central location – no hassles or hidden expenses.
- Small web sites gain a search engine at no cost.
- Easy integration (4 simple steps)

DFN S2S Definitions

- DFN
  - Deutsches Forschungsnetz
- Peer
  - A peer is a computer with the DFN S2S software installed
- Peer-to-peer (P2P)
  - Peers in the network can answer to and initiate requests, the roles of client and server are combined in the peer. User computers become active participants.
- Search
  - Search in the full text and in fields (if discernible) of textual resources

Goals of DFN S2S

- **To improve research capabilities** by implementing the task of **indexing** the Deep Web and other hidden content using a **peer-to-peer** approach, limited to materials interesting to **scientific research**.
- Focus is on search. The P2P substrate is handled by JXTA.
- Focus is on a working network which anyone belonging to the scientific community in Germany (i.e. connected to the G-WiN) can join.
- How?
  - **By installing a simple software package on their computer**

What is S2S?

- **S2S is a network** of peers which support search in local document sets.
- S2S is, from a user perspective, a piece of **software** which one can install to index one's own data, and
- it is an **online search service** allowing browser access to data in the network.

- Next slides:

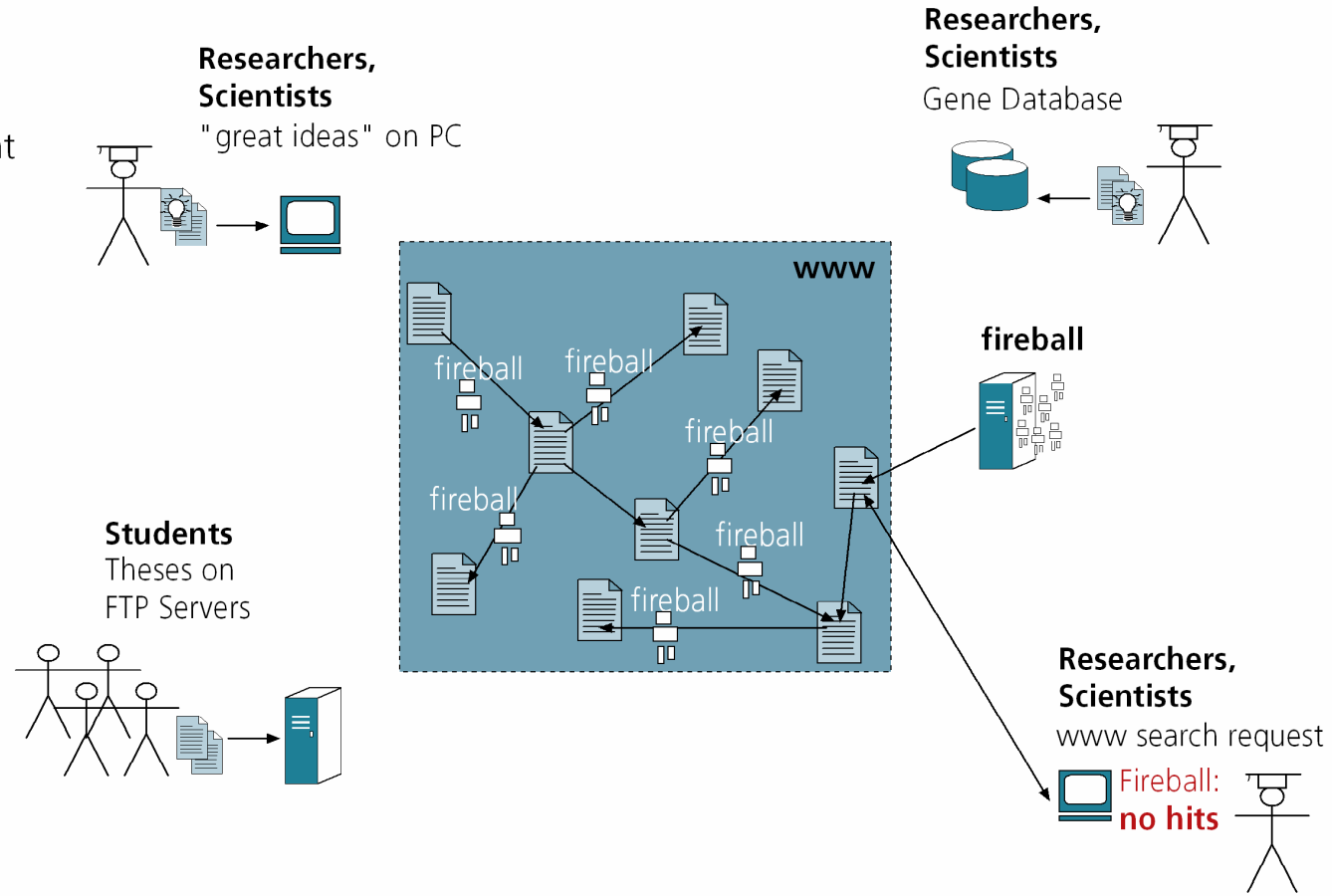
  comparison of S2S with conventional search engines

Fireball: Web Search

Centralised gathering,
Indexing aund Search

Disadvantages:
- doesn´t scale
- information not current
- source: only www
- centralised

Advantages:
- Availability controlled

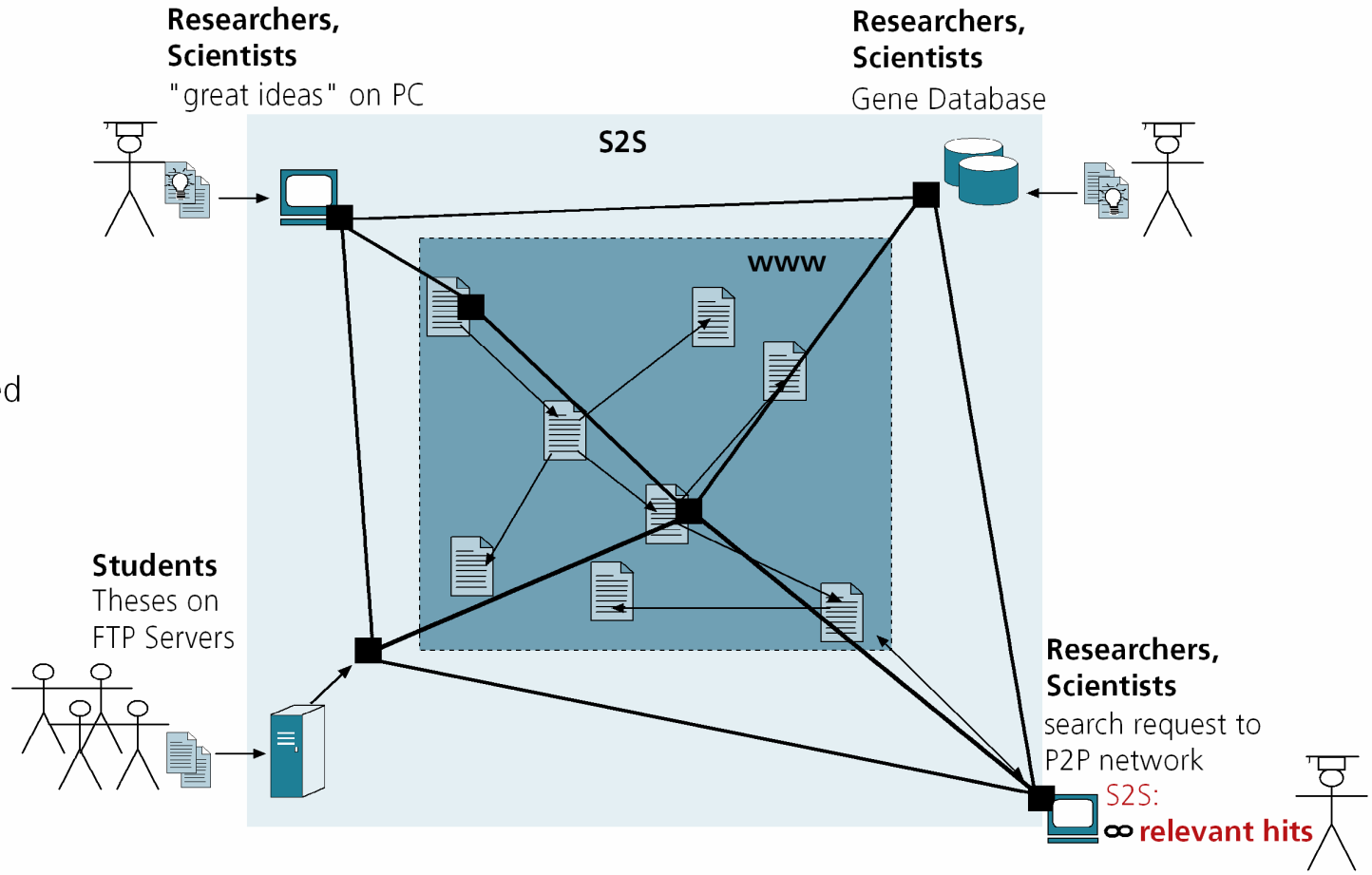


S2S: Deep Search

Distributed Indexing,
Search

Advantages:
- scales well
- current
- Information from entire Internet
- other yet to be researched

Disadvantages:
- yet to be researched (availability)



Comparison of S2S and Centralised Search

- The contents vary.

Criterion	Centralised	DFN S2S
Collection Method	Initial link may be submitted by users or editors, robots find links from initial point.	Providers specify an initial link or place contents into their shared directory.
Searchable Data Set	Dependent on the throughput of the software robots.	Dependent on the number of participants. The more that participate the more complete the data set.
Currentness	Depends on the scope of the searchable data set. The larger the data set, the less current the search results.	Each provider is responsible for a small piece of the data set and is able to keep indexes up-to-date.

Comparison of S2S and Centralised Search (contd.)

- Reaction times vary.

Criterion	Centralised	DFN S2S
Parallelism	Parallelism is internal to the single central node, e.g. the node consists of a cluster of computers, each of which may have multiple CPUs, etc.	P2P-style parallelism, in that the resources of each new node add to the power of the network
Availability	Depends on access to the server.	Not all nodes need be available all the time for the network to function. Possibility of commercial provision of data.
Search Throughput	Depends on the resources committed to serving searches. The greater the popularity the worse the preformance.	Increased throughput with increased popularity / number of nodes. Indirect communication (firewalls) overhead offset by caching.

Why Use S2S?

- **To search**: it is undisputed that current search methods used by researchers have grave inadequacies which are getting worse as the amount of information increases.
  - **Alternative** searching model – community based
  - Previously invisible contents (**deep web**)
  - **Up-to-date** nature of information – essential for researchers
- To **download**: Having found an interesting document the user may download it.

Why Share Data with S2S?

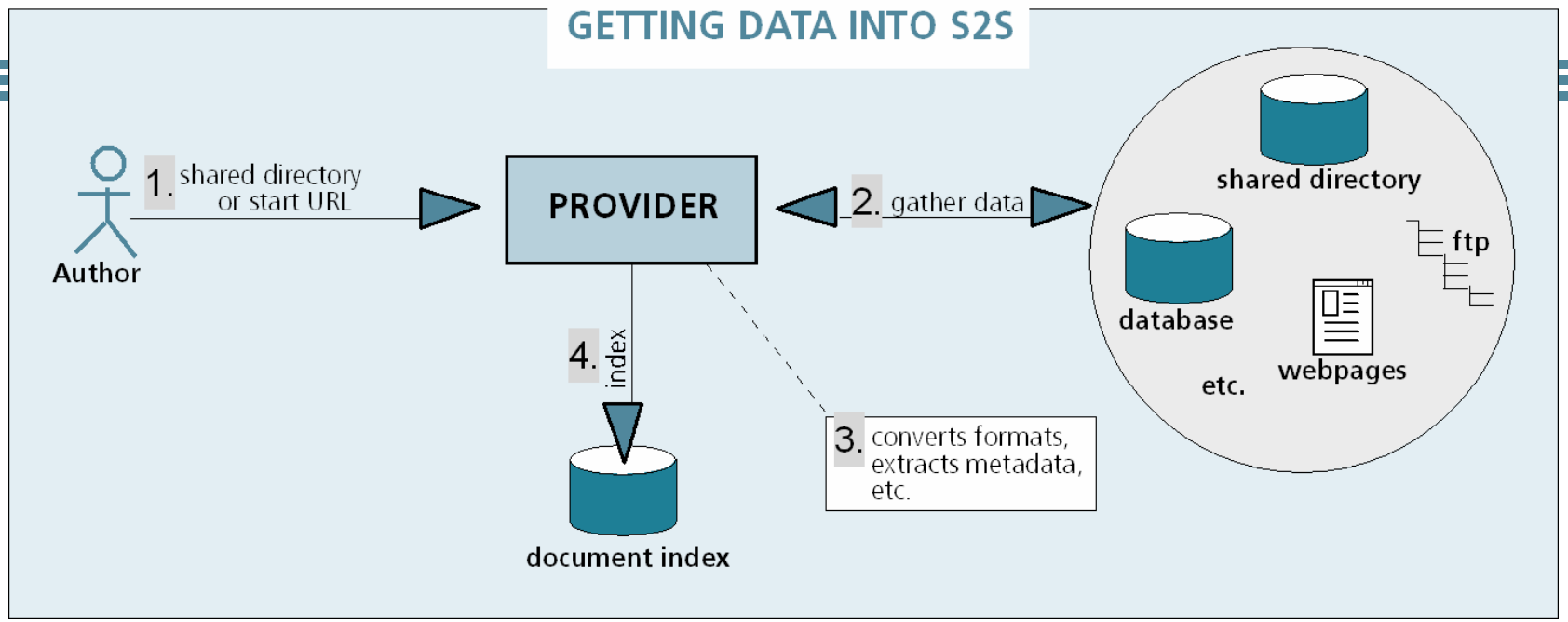
- There are several reasons why a researcher might be interested in sharing data with other researchers:
  - **Create a community** to share data
  - **Publication space limited** in conventional media – journals.
  - **Ease of publication**
  - Actually interested in indexing own material to make it searchable for herself
  - Uses the software to make an own search engine (similarly to **:suchexpress**)

Using S2S

- Recap: Basic functions of S2S are thus:
  - **Publication** of documents
  - **Building communities** through the publication process
  - **Search** for information
  - **Download** of documents
- These functions are **supported by the peers** in the network
- An overview of how they work is explained in the following diagrams…

Publication

- **“Providing data”** in JXTA Search / S2S terminology
- During installation: Data provider is prompted to **join** one or more communities and enter a **profile** (nickname, email, etc.)
- Publishing documents is a **one-step** process (see next slide)
- User can **check contents** of her index
- Process is **simple and safe** otherwise no-one would be prepared to share data



- 1. Set up shared directory

  AND/OR

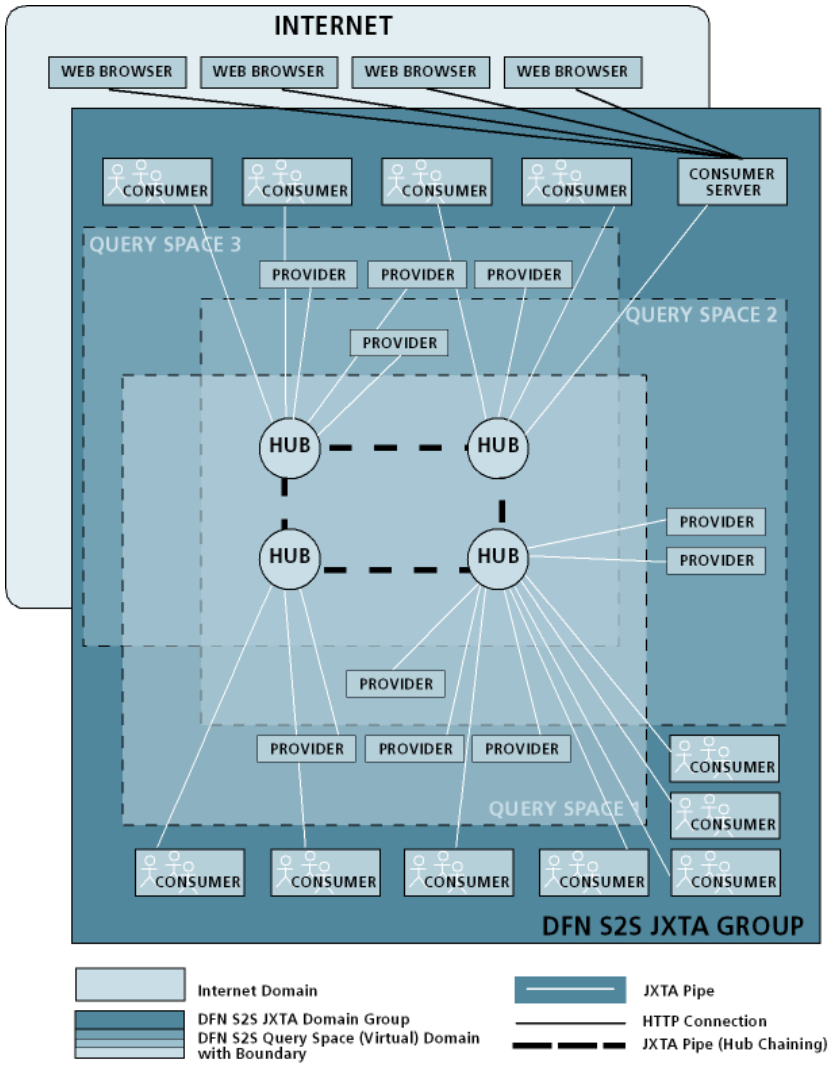
- 1. Choose start URL's – can be FTP, FILE or HTTP URL's.

- 2. S2S gathers the given documents from wherever they reside

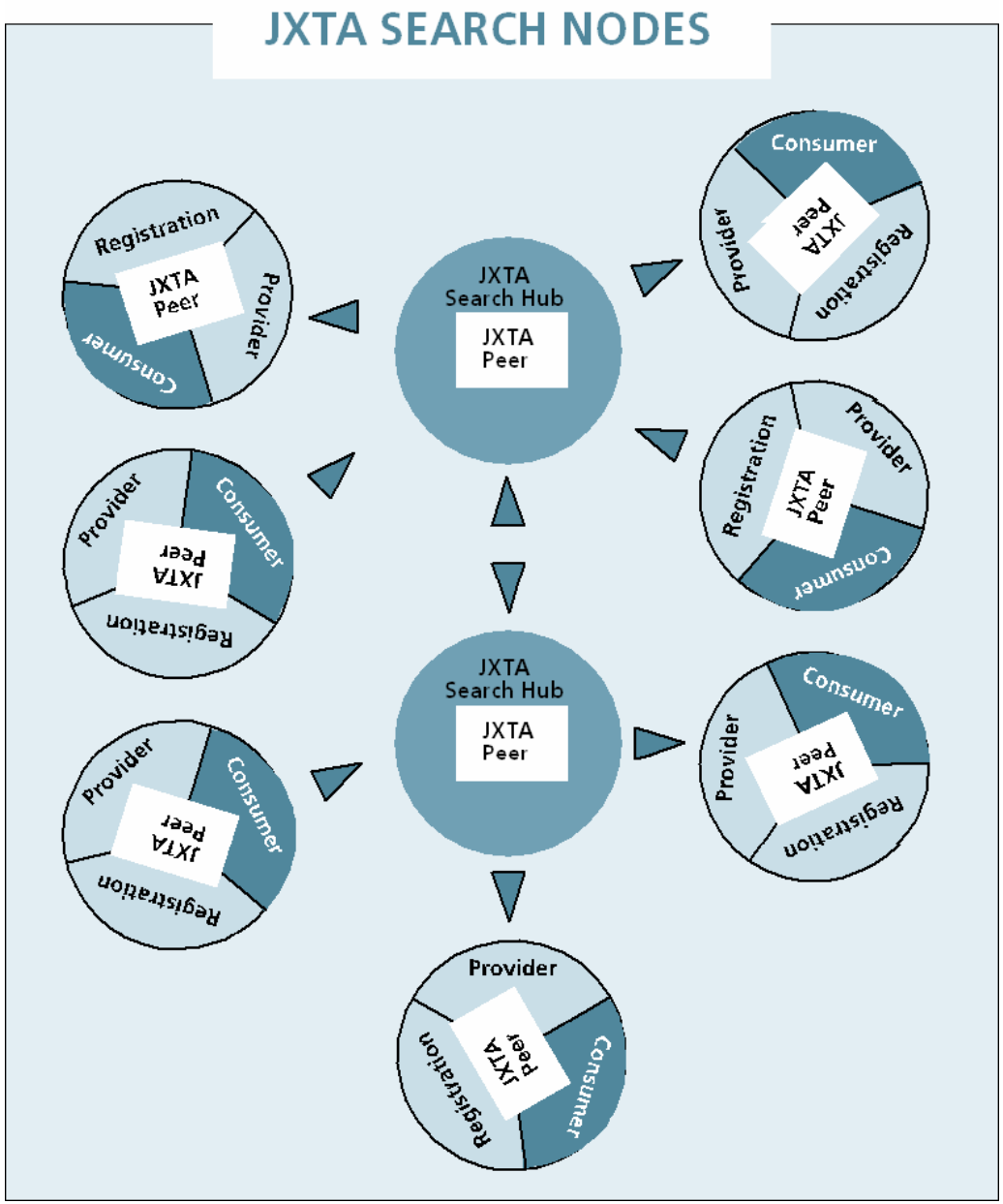
- 3. Metadata is extracted according to the format: HTML, Word, PDF, LaTex…

- 4. The enhanced document is indexed.

Building Communities through Queryspaces



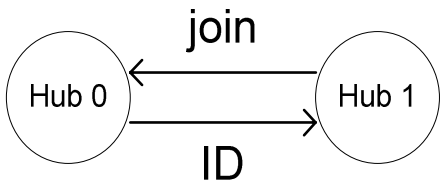
- S2S Communities are supported by the **queryspace** construct in JXTA Search (next slide).
- Queryspaces **direct searches** to particular providers belonging to a particular community.
- Users can **create** their own queryspaces or **join** predefined ones – this is as easy as selecting a name.
- Communities are **open** – cross community searching is allowed
- Grouping in communities means better results for directed searches.

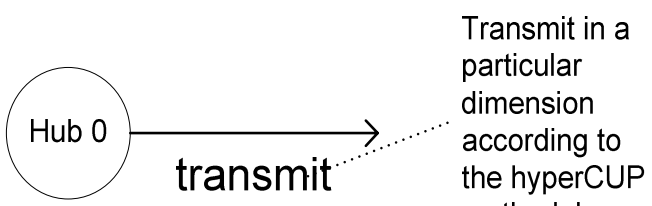


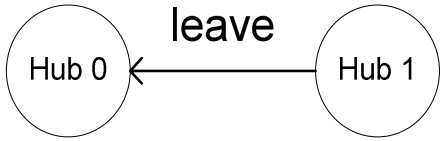
- Peers in the S2S network communicate using the **JXTA Search protocol,** which in turn bases on the JXTA protocol.
- **JXTA Search** also determines the structure of the network
- **JXTA Search** nodes are either regular peers (Consumer/Provider) or super-peers (Hub)
- In S2S hubs also **optimise relevance** ranking, **spam exclusion**, etc.

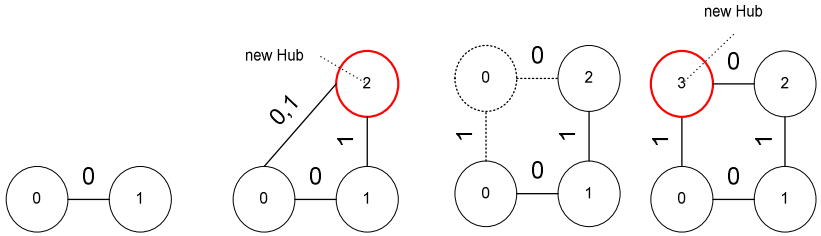
Hub chaining

extra







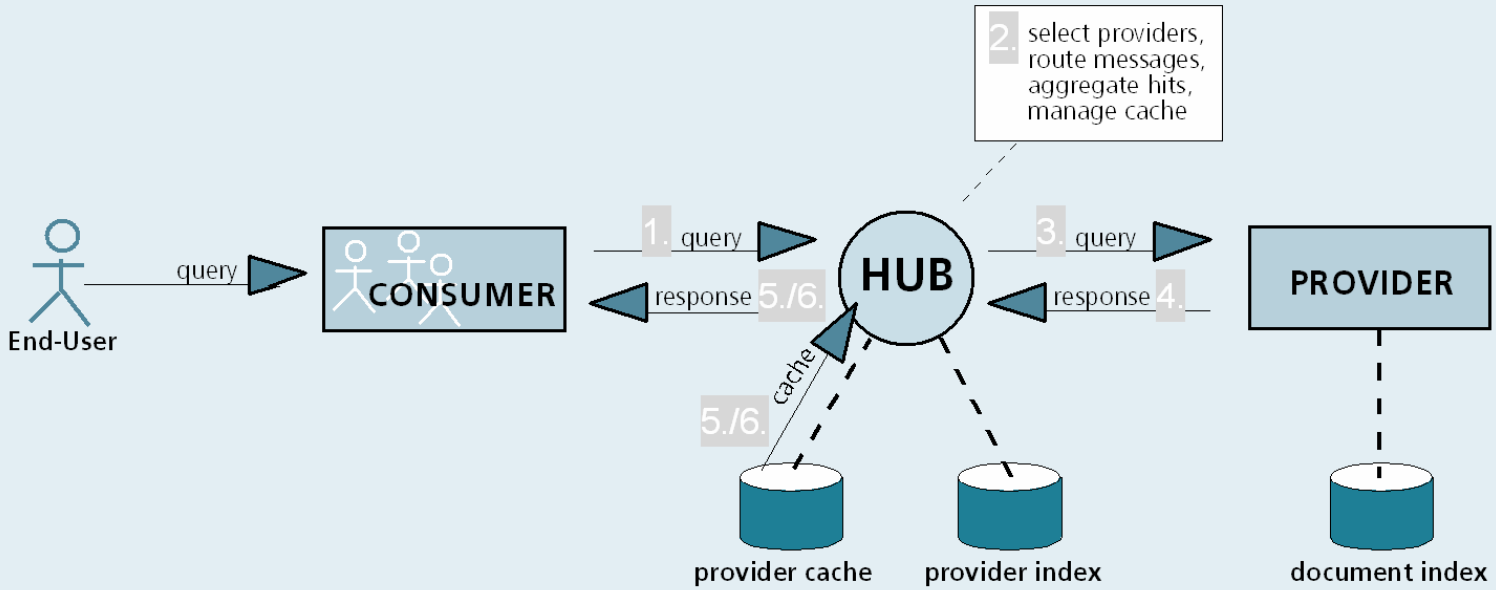


- A single Hub does following:
  - routes queries
  - maintains log of providers
  - maintains log of consumer “voting”
  - maintains cache
- Hubs need to communicate in order to share:
  - queries
  - registrations
  - index terms and document counts
  - “voting” & spamming information
- Hubs can be chained using the HyperCUP method to optimise communication.

Search and Download

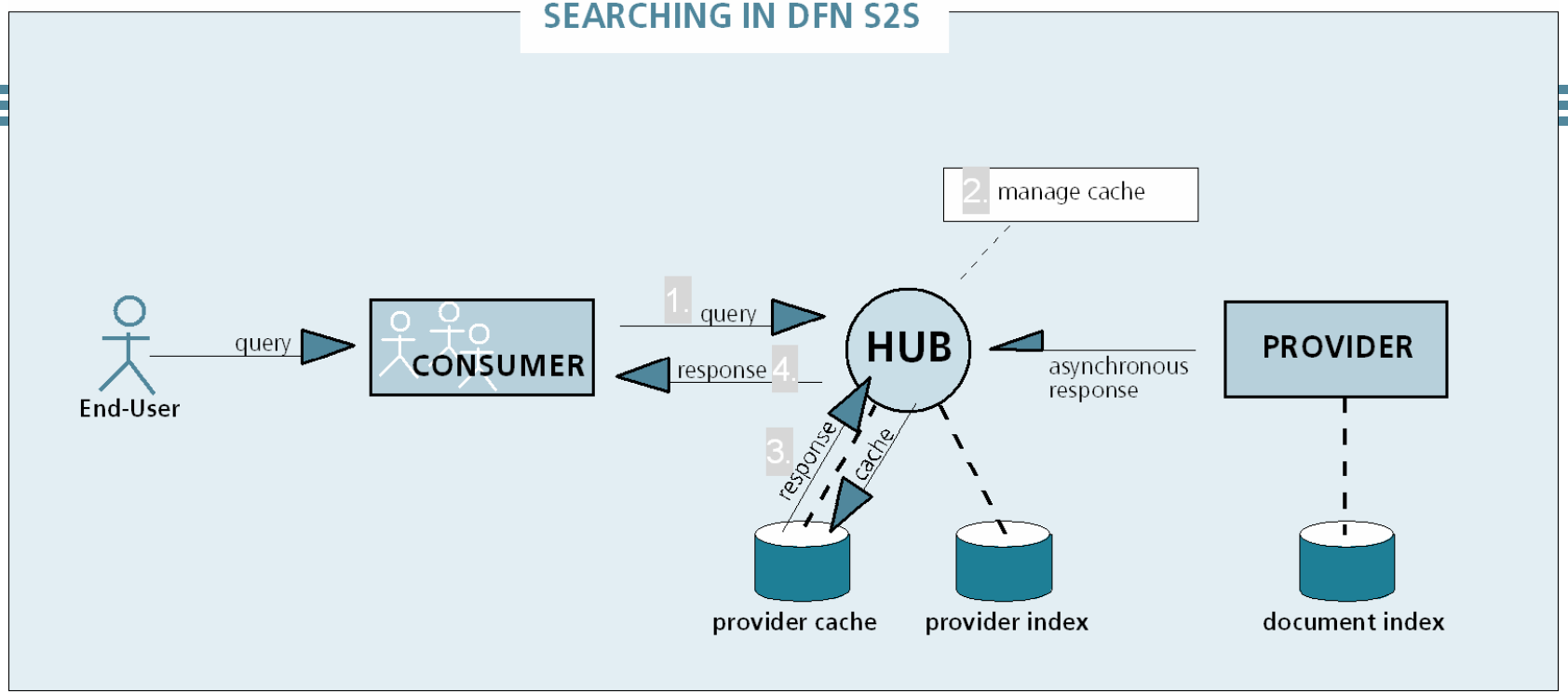
- **Access** to the S2S network is available via:
  - Graphical User Interface (**GUI**) – **Java application** automatically installed with the peer
  - **Web Browser** – to allow the contents to reach a wider audience
  - **Mobile Device** – if the researcher needs a quick reminder
- Search in **full-text** and **fields**, using **advanced operators** (+, -, " ", range, proximity, etc.)
- **Sort** results by relevance or date

SEARCHING IN DFN S2S



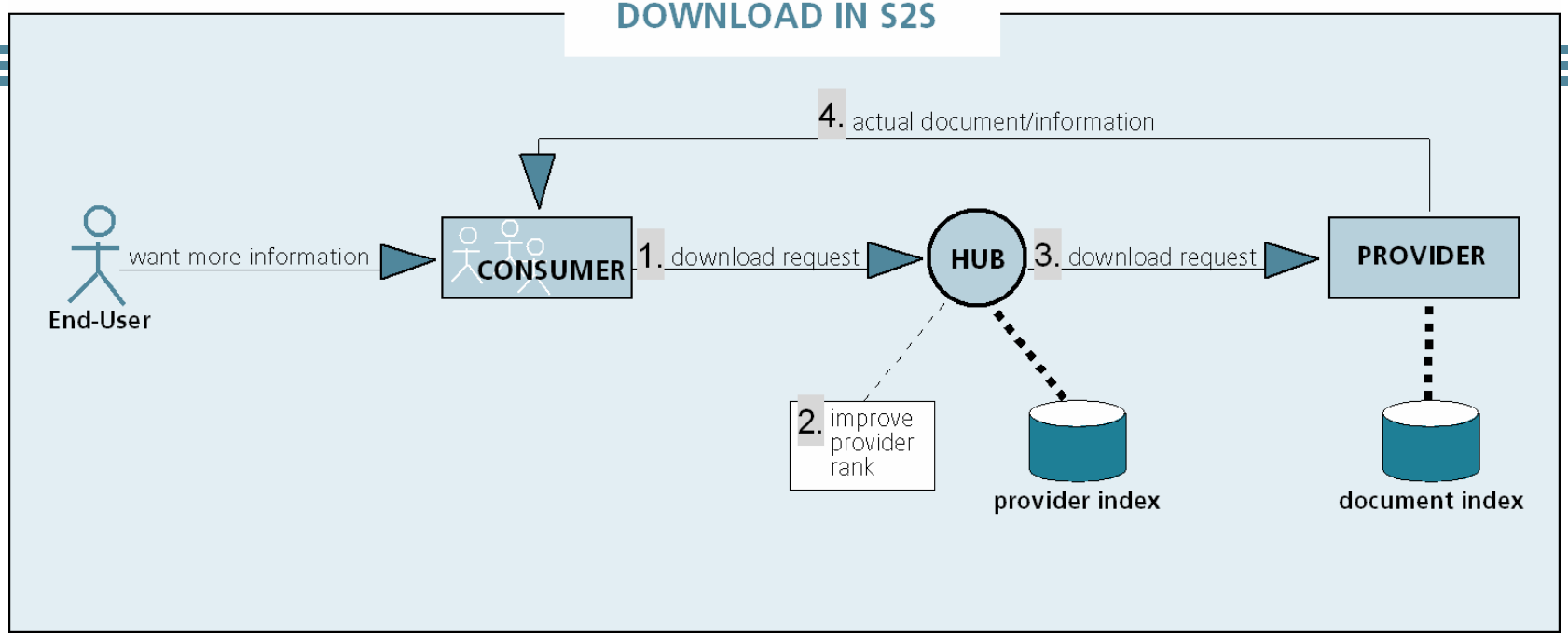
- **Basic Search Process** in S2S illustrating semantic routing and caching.
- 1. Consumer makes a query
- 2. Hub selects fitting provider i.t.o. registrations
- 3. The query is sent to the provider, who
- 4. Returns a response
- 5/6. Which is cached and if it is on time also passed back to the waiting consumer.

SEARCHING IN DFN S2S



- **Caching** is an important search optimisation in S2S
- Responses from **slower providers** can be added asynchronously, i.e. **resubmitting queries** allows the user to see responses which arrived in the meantime.
- If two users submit the **same query**, the network does not need to be consulted again – this is much more **efficient**

DOWNLOAD IN S2S



- The consumer requests a document from the Hub in order to keep track of **provider popularity**
- The provider sends the document **directly to the consumer**.
- A simple **extension** of S2S allows the provider to **protect documents** by allowing only certain consumers to download them.

Summary of Information Available

- **Searching** in S2S the user **finds**:
  - Information about a **document**
  - Information about the **provider peer** (measures of time in network, number of queries answered, etc.)
  - Information about the **peer network** (measures of numbers of peers / hubs and activity)
  - Information about **queryspaces** in the network: available only in the GUI to encourage users to install the software.

Summary of Control Possibilities

- The **hub** as the **centralised element** presents many possibilities for the control of network usage:
  - Bandwidth control to slower peers or in general
  - Semantic routing not flooding
  - Spam filtering (e.g. check response against query)
- The **provider** can **prevent the download** of detailed information or documents, thus improving security and decreasing bandwidth usage.
- **Consumer** can only vote for/against a provider or content (explicitly and implicitly)

Underlying Technology

- **JXTA**
  - Peer-to-peer platform providing the basic P2P communication functionality
  - Can tunnel through firewalls using simple polling - indistinguishable from browsing (by packet inspection, identification would require traffic analysis).
- **neofonie search**
  - Optimised search and retrieval technology
  - Provides S2S with a rich feature set (sorting, ranking, Boolean and range operators, field searching adaptability to providers' needs)

JXTA P2P Software

- JXTA is an **open source community Project** based on an Initiative of Sun Microsystems.
- Aim is an IETF (Internet Engineering Task Force) standard **protocol**
- Also a **software platform** implementing these protocols which:
  - Demonstrates that the ideas work
  - Is a freely available infrastructure on top of which others can build interoperable P2P networks.
- A maturing technology (**version 2.0**)

neofonie search

- Innovatively engineered software – **powers large applications** like the AOL website.
- 2 components are used in DFN S2S, **:robot** and **:engine**.
- **:robot** – multi-purpose spider and document processor,
  - converting formats, enriching with metadata
- **:engine** – an XML-indexer and repository with native full-text and XML search
  - patented relevance ranking technology
  - light-weight and fast

Expected Results

- System usage depends on offer of content.
  - The more content there is, the greater the usage, which in turn increases the amount of usage.
  - This results in a **cornucopia of the commons** (Dan Bricklin in Andy Oram, 2002)
  - Will S2S achieve this critical mass?

- There will be **greater exposure for scientific information** not available by other means.

- Some **scientific community building** will take place via the software tools. By placing the content within a specific queryspace and through the use of structured data, communities will be able to evolve.

Expected Results (2)

- **No significant resource investment** by researchers into preparing data for the network, rather they will make use of the automated tools only.
- **Secure operation** – viruses cannot spread automatically.
- **Legal problems** may arise because of copyright issues. These will have to be handled in an ad hoc manner.
- **Inappropriate material** may be included in the network.
- **Network administrators** at participating institutions, should have **no problems** with bandwidth use or having to support users

Expected Results (3)

- **Commercial demand** for applications based on the S2S model.
- No competition

	Share documents	Full-text and field search	Ad-hoc communities	Provider can choose tools	Access docs via file, http, or ftp
S2S	X	X	X	X	X
Grub		X			Cannot choose URL
Edutella	X	X	X	Metadata must be RDF	
Groove	X	X	X		

Roadmap

- **User**:
  - April / May 2003: call for Beta-testers
  - July 2003: first installations of software and building of prototypical S2S network
  - Entire 1st Quarter 2004: pilot phase, S2S for public consumption
- **Development**:
  - May 2003: Milestone 2, basic network functionality ready
  - December 2003: Milestone 3, easy install and advanced features (peer administration, full hub functionality) ready
  - March 2004: Milestone 4, network maintenance and perhaps link with other networks e.g. ELENA

Beta-Testing Program

- Try the software out for yourself.
- See the advantages, index your information, search other contents.
- **http://s2s.neofonie.de**

Thank You!

- Hopefully there is some time for feedback!