Title: **DFN Science-to-Science: Peer-to-Peer Scientific Research**

Author: R. Wertlen
Affiliations: neofonie gmbH, Robert-Koch-Platz 4, 10115 Berlin, Germany,
rrrw@neofonie.de

Keywords : Peer-to-peer, information retrieval, user application

Extended Abstract:

The DFN [1] - the german research network - is a non-profit association providing the computer-based communication infrastructure for science, research and education in Germany. The project Science-To-Science is an attempt to harness peer-to-peer (p2p) technology to serve the needs of scientific research in order to promote the exchange of knowledge and information. The pilot phase of the project caters to interested scientific institutions connected to the G-WiN, the scientific Gigabit network run by the DFN.

The project is being managed and implemented by neofonie GmbH [2] a company based in Berlin and specializing in search and content management in distributed environments. neofonie has been developing search engines since 1998. It also developed the distributed content management system in use by AOL Germany. neofonie enjoys close ties with the research community and continues to be active in this field.

The main aim of the project is to unlock the wealth of knowledge hidden on machines connected to the G-WiN but not accessible to search engines. This information may come in the form of web pages, database contents, regular document files (sound, audio, video) on ftp or mail servers. The information is indexed as structured or unstructured full-text depending on the source. The indexes may be searched by anyone in the DFN S2S network.

On the one hand, the motor driving this heightened accessibility will be researchers and scientific institutions themselves. The p2p paradigm means autonomous control of one's own resources: scientists will choose which resource to make available and in which manner. The main reasons we have identified why researchers will want to make use of such an alternative publishing source are: the ultimate goal of research in an abstract sense is to make the results publicly known; until now the publication source has seldom if ever been under control of the publishing scientist; the limited amount of publication space available to researchers in traditional science and peer reviewed publications means that papers published often represent only a fraction of the material the scientist wishes to publish.[3]

On the other hand, the innovative fusion of neofonie's high speed, flexible XML-based software together with the XML-based p2p platform of the JXTA project [4] will provide the technology required to realise these aims. The project software will require no expert knowledge to be installed and operated. It will allow the user to select what sort of peer they wish to install: searching (termed "consumer") or indexing (termed "provider"), or a combination of these. Being created at the institution which produces and controls the data, the indexes are closer to the data source than regular search engines. Also they are responsible for a smaller portion of information. Thus we expect that information retrieved will be fresh - a consumer will always find current information. To increase utility of data, files may also be downloaded through the DFN S2S network. Scientists may also disable this feature if they do not want anonymous access to their shared information.

The peer computers in the network communicate using the JXTA Search open protocol [5]. Using an open protocol ensures that in the future it will be easy to extend the network by adding new data sources, as well as making the information in the network available in innovative ways. An important feature of the network is the presence of centraslistic components termed "hubs" which are responsible for optimising and monitoring the flow of information in the network in an automated fashion. Investigations show that messages can be exchanged much more efficiently in scale-free networks which are organised around super-peers (in this case the hubs) and exhibit the "small-world" phenomenon (short routes from a peer to any other peer), whereas p2p networks without topological consideration are more prone to bandwidth wastage [6,7].

The neofonie search suite of indexing, gathering and processing software provides indexing and information collection facilities in the project. The fast XML and full-text indexing capabilities as well as advanced relevance ranking ensure that researchers have precise access to relevant information [8]. Metadata expressible in XML is also able to be indexed preserving the structural information. The structural information can then be used to hone a search. This is analogous to "field searches" in retrieval technology terms. Additionally a gathering and processing component allows one to link ftp and database sources, while it also enriches documents with standardised metadata according to the Dublin Core schema. We do not expect that researchers will have the time to enrich data manually, nor that they would be willing to share this additional effort [3]. Thus the focus of the project is on automatic methods only.

The indexing software also plays a major role in the hubs of the network. Since JXTA is based on XML, indexes of messages including queries and responses can be maintained. These indexes will be used to match searcher with provider in an adaptive fashion. Specifically, responses are indexed at the hub and this information is used when deciding to which provider a query should be routed.

Some of the things we shall be observing in the pilot phase at the beginning of 2004 are:

- There will be greater exposure for scientific information not available by other means.
- Some scientific community building will take place via the software tools.
- Researchers will not invest any significant amount of time in preparing data for the network, rather they will make use of the automated tools only.
- 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. Solutions to this problem are being discussed and will most likely include some form of filtering at the hubs. One readily apparent mechanism is to attempt to construct profiles of providers from their responses in order to report or simply block suspect transactions.

During the pilot phase neofonie will provide a central service in matters such as inappropriate materials. The intention is to refine the network so that automatic mechanisms allow the network to regulate itself. If the pilot phase is a success and use of DFN S2S continues, an alternative body will have to be set up to monitor issues which the network cannot regulate.

Evaluation of the project results will rest mainly on the feedback of users as well as usage statistics and patterns. While the number of documents in the network could also be used as a measure of success (we are expecting several millions), ultimately the satisfaction of the researcher using the system is the only meaningful measure. Thus, the project strives for contact with the scientific community as users of the application. Without the user, the network cannot exist.

The project beta testing phase begins in Mid-2003. Progress can be monitored at the following web-site:
http://www.neofonie.de/profil/forschung_und_entwicklung/s2s.jsp

Acknowledgements:

Thanks to all those who reviewed and made suggestions to this extended abstract particularly Gerti Foest at the DFN.

References:
1. http://www.dfn.de/
2. http://www.neofonie.de/
3. R. Wertlen, "What users want: Position Paper for the International P2P Workshop 2003", unpublished paper from November 2002. http://www.neofonie.de/profil/forschung_und_entwicklung/IPTPS03_neofonie.pdf
4. http://www.jxta.org/
5. S. Waterhouse, D.Doolin, G. Kan, Y. Fabishenko " Distributed Search in Peer-to-Peer networks", IEEE Internet Computing vol. 6, no. 1 January/February 2002, pp. 2 - 6.
6. A.-L. Barabasi and R. Albert, Science 286, p. 509 (1999).
7. B. Yang and H. Garcia-Molina. Designing a super-peer network. http://dbpubs.stanford.edu:8090/pub/2002-13, 2002.
8. neofonie search: Guide to indexing and gathering, http://www.neofonie.de/download.jsp, May 2002.

Ronald Wertlen	Infonie GmbH	Tel: +49.30.24627-211
Projektleitung	Robert-Koch-Platz 4	FAX: +49.30.24627-120
rw@infonie.de	D-10115 Berlin	Web: www.infonie.de