PUBLISHED AS:

H.A. Proper. Pronir proposal. Technical report, Nijmegen Institute for Information and Computing Sciences, University of Nijmegen, Nijmegen, The Netherlands, EU, 2002. NWO Project Proposal.

1 Project

Title	**P**rofile based **R**etrieval **O**f **N**etworked **I**nformation **R**esources
Acronym	PRONIR
Principal investigator	Dr. H.A Proper

2 Summary

The knowledge and information resources we need to conduct our activities in daily life, be it at work or at home, are increasingly available in some electronic form by way of the Internet. Some examples of such resources are: *documents*, *people* (by their e-mail or chat addresses), *document collections*, *objects and facts in databases* and even entire *applications*. People in search of knowledge turn to the Internet with the aim of finding relevant knowledge or information that will support them in executing their tasks.

This research project aims to develop a theory, and demonstrate its validity by means of a prototype system, for profile based retrieval of heterogeneous networked resources. The proposed research involves two key focus areas:

Profile based retrieval This area focuses on the ability of an information retrieval system to tune the set of retrieved resources to the specific information needs of a searcher.

Uniform resource access The focus of this area is on a uniform way of modelling and characterisation of heterogeneous knowledge resources.

3 Classification

The research projects fits within the DISH (Electronische Snelweg) program. The following (sub)disciplines of the "*Verkenningscommissie Informatica*" are applicable to the project: *2.3 Information Retrieval*.

4 Composition of the Research Group

The project involves researchers from three institutes:

> *The University of Nijmegen, The University of Tilburg, and The Ordina Institute for Research and Innovation.*

Researcher	Dr. P. van Bommel	Dr. J. Hoppenbrouwers	Prof. Dr. M. Papazoglou	Dr. H.A. Proper[1]
Institute	University of Nijmegen Sub-faculty of Informatics	University of Tilburg Infolab	University of Tilburg Infolab	University of Nijmegen Sub-faculty of Informatics and Ordina Institute
Expertise	Conceptual modelling, database optimisation, information retrieval	Digital libraries	Distributed databases, E-Commerce, digital libraties	Information architecture, conceptual modelling, information retrieval
Involvement	1 day/week	1 day/week	1 day/week	1 day/week

Researcher	Dr.ir Th.P. van der Weide	Vacancy 1	Vacancy 2	
Institute	University of Nijmegen Sub-faculty of Informatics	University of Tilburg Infolab	University of Nijmegen Sub-faculty of Informatics	
Expertise	Conceptual modelling, information retrieval	Business modelling	Information modelling	
Promotor		Prof. Dr. M. Papazoglou	Dr.Ir. Th.P. van der Weide	
Involvement	1 day/week	Full-time	Full-time	

[1] Dr. Proper also works as Principal Scientist at the Ordina Institute for Research and Innovation and is a Partner of the Ordina Group of Companies.

5 Research Schools

Institution	School
University of Nijmegen Sub-faculty of Informatics	NICI (Nijmegen Institute for Cognition and Information) IPA (Institute for Programming research and Algorithmic)
University of Tilburg	SIKS (School for Information and Knowledge Systems)

6 Description of the Proposed Research

6.1 Socio-economic relevance

The internet has become the virtual reality of mankind - a world that we shape without many of the imperfections of reality. We can jump to literally every place in no time, live and die many times, change our identity at will, and reach every resource anywhere anytime. In particular this last promise of *information at your fingertips* is under siege. The growing complexity of information space overwhelms the wired consumer and the vast increase in information is outpacing the improvement of retrieval tools.

The knowledge and information resources we need to conduct our activities in daily life, be it at work or at home, are increasingly available in some electronic form by way of the Internet. Some examples of such resources are: *documents*, *people* (by their e-mail or chat addresses), *document collections*, *objects and facts in databases* and even entire *applications*. People in search of knowledge turn to the Internet with the aim of finding relevant knowledge or information that will support them in executing their tasks.

6.2 Research goals

This research project aims to develop a theory, and demonstrate its validity by means of a prototype system, for profile based retrieval of heterogeneous networked resources. The proposed research involves two key focus areas:

Profile based retrieval This area focuses on the ability of an information retrieval system to tune the set of retrieved resources to the specific information needs of a searcher.

Uniform resource access The focus of this area is on a uniform way of modelling and characterisation of heterogeneous knowledge resources.

6.2.1 Profile based retrieval

Profile based retrieval refers to the ability of an information portal to aid users in effectively finding relevant resources, while taking their specific interests, defaults and needs into account [3, 2, 14].

A pivotal role is played by what the searcher's profile. It is this profile that enabled an information retrieval system to better tune its behaviour to the needs of the searcher. Some examples of aspects that could be part of such a profile are:

- Defaults searchers may harbour with respect to the use of search terms. For example, a user who, referring to surfing usually refers to wave surfing as opposed to wind surfing or internet surfing.
- The searcher's aim for retrieving the resources. For example, reference purposes, orientation on the subject, in-depth study of the subject, etc.
- The searcher's cognitive style of consuming information (*depth-first*, *breadth-first*, *berry picking*, ...). Based on these latter characterisations, a retrieval system may quite well suggest an effective order in which a set of information objects should be read to best satisfy their information need.

For most individuals, profiles are likely to differ depending on a particular role or task at hand. This implies that we should go further than just a personal profile, but should also distinguish role or task based elements. An ensuing requirement is that there must be some mechanism to combine several profiles. For example, a searcher's private profile and a profile that is relevant for the task she is currently performing.

To develop theory for profile based information retrieval, a fundamental understanding of the information retrieval problem and its relation to the searcher needs is required. The research, as conducted by one of the project proposers, as reported in [11] provides a semantic base to start from. This semantic base has been developed in terms of the concept of infons as used in Situation Theory [1] and logic. As reported in e.g. [8, 13], logic is playing an increasingly important role to more fundamentally define the reasoning mechanisms driving information retrieval.

For a profiling mechanism to be useful, it is imperative that the definition of an actual profile does not put an extra burden on the searcher. To this end, strategies and algorithms are needed to define and/or derive profiles. A user specific profile may, for example, be based on the searcher's implicit or explicit feedback when using the information retrieval system [2, 14]. A role, or task, based profile may be defined in conjunction with the definition of the workflow or business process to which the role/task is associated. Alternatively, using data mining techniques, a role/task profile may be derived by observing different searchers performing the same role/task.

Specific research goals, with respect to the profiling focus, are:

- The definition (both syntax and semantics) of a profile specification language.
- Development of strategies and algorithms to define user, task and role specific profiles.
- Development of, logic based, information retrieval mechanism, which includes the searcher's profile.

The language, algorithms and mechanisms developed, should be tested in practice by the development of a prototype profile based retrieval engine.

6.2.2 Uniform resource access

The availability of numerous information resources via the Internet, brings about a natural expectation of being able to search and access these resources in a uniform way [10]. The present situation, however, is quite different. Users need to use different systems to find the resources they need. This requires the introduction of a mechanism by which highly heterogenous information resources of different origin can be presented to searchers as one integrated search space.

In this project we will focus mainly on providing a uniform view of different properties, such as meta-data of information resources and relationships between information resources. This project is not concerned with standardisation of transfer protocols or storage formats.

Currently available meta-data standards, standards for defining the structure of information resources and resources description frameworks, such as [6, 4, 12], provide a good starting point. However, in this research project we would like to go beyond these frameworks, aiming to provide more richer functionality with respect to dealing with heterogeneity of information resources and their semantics.

- Supporting searchers in formulating their needs, require information retrieval systems to have a good understanding of the structure and semantics of the domains used to characterise the information resources.

  For example, different domains may have been used to characterise what information the resource provides. Examples are (freely selected) keywords, keywords from a restricted set, noun phrases, or even entire conceptual graphs, etc.

  A more simple example would be the characterisation domain used to express the price that must be paid to obtain an information resource. What makes this domain an interesting example, is that in the context of the Internet we will have to cater for the use of different currencies. How will an information retrieval system be able to identify the cheapest information resource when prices are stated in different currencies?

An example of a more complex characterisation domain are the relationships that one information resource may have to other information resources. For instance: *associated to*, *author of*, *part of*, *abstract of*, *works for*, etc.

- It is not advisable to "hard-wire" the definitions of characterisation domains into an information retrieval system, as variations and evolution of these domains are likely to occur. An information retrieval system should ideally be able to interpret the definition of a characterisation domain and act accordingly.
- In addition to the definition of characterisation domains, it will be necessary to provide definitions of translations between these domains.

  For example, between prices in Euro and prices in USD, or from noun-phrase based descriptions to keywords. In certain cases, it will even be necessary to refer to the underlying resource itself. For example, when only a keyword based characterisation of the resources is available when a noun-phrase based one is more preferred, the keywords could be used to derive apt noun-phrases from the original resource (presuming we are dealing with a textual information resource).
- Within a collection of resources (a corpus) certain rules may be applicable with respect to the relevance of these resources to a searcher's needs.

  For example, in the context of a corpus on management of software development projects, a user who considers a resource on ”unstable software requirements” relevant, is likely to find a book on evolutionary software development more relevant than a book on linear software development.

In the proposed research, we will use knowledge representation languages and frameworks such as KQML [7], OIL [9] and DAML [5] from the DARPA as a base to develop a definition language for characterisation domains. The very existence of projects such projects also underlines the fact that RDF is not rich enough (yet) to deal with these richer forms of semantics.

Specific research goals:

- The definition of a (machine interpretable) language to define characterisation domains.
- A transformation mechanism to provide translations between different characterisation domains (when applicable).

Again, the languages, algorithms and mechanisms developed, should be tested in practice by the development of a prototype system.

6.3 Research approach

Information retrieval research requires a balance between theoretical research and evaluation of theoretical results in terms of a research prototype of a search engine. The theoretical research as well as the evaluation of the prototype systems are the key responsibilities of the two OiO's for which we seek funding. Each OiO will be responsible for one of the defined focus areas. The development of the prototype system is key the responsibility of the scientific programmer for which we also seek funding.

The project aims to develop the theory and prototype system in an incremental way, using three iterations, each iteration taking just over one year to complete (see the *work program* section below). After each incremental step, the resulting prototype is evaluated emperically, while the evaluation results are used as input to the next increment.

The prototype system will be developed as open source. This is a conscious choice. We strongly believe that developing the prototype system as Open Source may have the following additional benefits:

- It forces the use of a well-thought out software architecture, enabling third party involvement. Be it MSc students from the institutes involved, programmers from interested software companies, or be it the open source community in general.

- We aim to develop the prototype system in such a way that it will also serve as a base for additional information retrieval experiments beyond the project's own scope. Both during the duration of the project and beyond its duration.
- When a wider audience used the prototype system for further experimentation, this is likely to benefit the quality of the theories developed by the original project team.

6.4 Relationship to current activities

Two of the research projects that are currently being carried out by the research-line IRIS of the University of Nijmegen provide research results that will contribute to the proposed project:

1. The *Profile project*, which is a collaboration with the "Nijmegen Institute for Cognition and Information (NICI)", aims to develop mechanisms for pro-active information filtering. This project investigates a deeper representation for the user profile, namely at the level of goals and interests that gave rise to the information need in the first place.

   The research results that follow from the Profile project will therefore contribute to the aims of the proposed project, in particular where it concerns the precise definition of a user's information need and purpose of this need.

2. The *Concept Lattices project*, which is conducted in collaboration with the faculty of arts (linguistics), aims to develop mechanisms to extract more meaning from text documents. In this project, concept lattices are used as a representation language for this meaning. The experiences with this project may be directly applied to the characterization of information objects in terms of a semantically richer characterization mechanism.

Two of the research projects that are currently being carried out at the Infolab of the University of Tilburg, will have research results that can be applied to the proposed project:

- The *Decomate-II project* (LB-5672/B-DECOMATE II), which is partially funded by the European Commission DG XIII Telematics for Libraries programme, aims to develop an end-user service which provides access to heterogeneous information resources distributed over different libraries in Europe using a uniform interface, leading to a working demonstrator of the European Digital Library for Economics.

  The results of the Decomate-II project will be used in the proposed project, where the results of Decomate-II will be generalised further.

- The *TREVI project* (Text Retrieval and Enrichment for Vital Information), which is also partially funded by the European Commission, aims at offering a solution to the problem of information overload, i.e. the difficulty experienced by both small and large companies in extracting useful information from large amounts of data coming from the numerous electronic textual information services available at a local or global level (Internet, proprietary networks, subscription services, World Wide Web, etc.).

  The experiences with this project may be directly applied to the characterization of information objects in terms of a semantically richer characterization mechanism.

- *MACS* (Multilingual ACcess to Subjects) is a project of the Conference of European National Librarians (CENL). Project goal is to develop a management system to join various European Subject Heading Languages (thesauri) into one, virtual search space. The MACS Project is related to work in the Cobra+ Project (COmputerised Bibliographic Record Actions).

  Key issue of the project is to maintain the independence of the various contributing Subject Heading Languages. In order to achieve this, we are building a federated SHL/thesaurus management system. It does not physically or logically combine the contributing SHLs into one merged, consolidated system. Instead, it acknowledges the independence of the contributing SHLs and uses explicit links between them which are maintained outside of the SHL data bases. The maintenance of the Link Database is decentralized and done by the participating authorities in a federated organization.

- *MEMO: MEdiating & MOnitoring electronic commerce* The goal of this project is to prototype, demonstrate and evaluate a lightweight Electronic Commerce environment (Electronic Commerce-broker) for the facilitation of smaller-scale and more diverse Electronic Commerce applications.

  The EC-broker provides an added value in the phases previous to the execution of the transaction where EDI traditionally provides support. With support in phases of partner search, negotiation and contracting the EC-broker in this project aims to fill a gap in currently available support to Electronic Commerce.

7 Work programme

OiO 1	OiO 2	Scientific programmer	Duration
Orientation	Orientation	Orientation	14 Weeks
Prioritisation	Prioritisation	Prioritisation	1 Weeks
Development of theory	Development of theory	Development of infrastructure	28 Weeks
Prototype design	Prototype design	Prototype development	15 Weeks
Prototype evaluation	Prototype evaluation	Prototype evaluation	12 Weeks
Prioritisation	Prioritisation	Prioritisation	1 Weeks
Development of theory	Development of theory	Prototype consolidation	28 Weeks
Prototype design	Prototype design	Prototype development	15 Weeks
Prototype evaluation	Prototype evaluation	Prototype evaluation	12 Weeks
Prioritisation	Prioritisation	Prioritisation	1 Weeks
Development of theory	Development of theory	Prototype consolidation	28 Weeks
Prototype design	Prototype design	Prototype development	15 Weeks
Prototype evaluation	Prototype evaluation	Prototype evaluation	12 Weeks
Finalisation of PhD	Finalisation of PhD	Prototype consolidation	26 Weeks

The total duration of the project is 4 years. In the above planning, the following steps have been identified:

Orientation In this step, the project team will establish itself. Furthermore, relevant literature will be surveyed.

Prioritisation At the beginning of each increment, the project team will set the ambition for that theory/prototype increment.

Theory development Development of theories conform the ambitions set out for the specific increment.

Development of infrastructure While the theory development of the first increment progresses, the scientific programmer will set up a technological infrastructure and preliminary software architecture for the prototype system. The latter architecture is particularly needed to make the prototype into a truly Open System, allowing for third party involvement.

Furthermore, during this step, the use of already available Open Source components to fill-in parts of the system should be evaluated as well. There are, for example, Open Source information retrieval systems available that may be used as abase.

Prototype design and prototype development During the development of the prototype system, the two OiO's are mainly responsible for the design of the functionality of the system, based on the theory developed by them, while the programmer is responsible for developing the (integrated) prototype.

Prototype evaluation In this step, the prototype will be evaluated.

Prototype consolidation While the two OiO's work on new theory, the scientific programmer will be able to further consolidate the existing prototype, which will serve as a base for further extensions and refinement. The consolidation steps are also needed to further enable third party involvement.

Finalisation of PhD Time is allocated for the two OiO's to write-up their PhD's.

8 Literature

In the reference section at the end of this document, the publications referenced in this document can be found. The five key publications of the project proposers, that are relevant to this proposal, are:

1. A.H.M. ter Hofstede, H.A. Proper and Th.P. van der Weide. Query formulation as an information retrieval problem. *The Computer Journal*, 39(4):255–274, 1996.

2. M.P. Papazoglou and S. Milliner, Subject-based Organization of the Information Space in Multi-database Networks. *Proceedings of the Tenth International Conference CAiSE'98 on Advanced Information Systems Engineering*, 251–272, Lecture Notes in Computer Science, Springer-Verlag, Pisa, Italy, 1998.

3. Th.P. van der Weide, T.W.C. Huibers, and P. van Bommel. The Incremental Searcher Satisfaction Model for Information Retrieval. *The Computer Journal*, 41(5):311–318, 1998.

4. F.C. Berger and P. van Bommel, Th.P. van der Weide. Ranking strategies for navigation-based query formulation. *Journal of intelligent information systems*, 12(1), 1999.

5. M.P. Papazoglou, H.A. Proper and J. Yang. Landscaping the information space of large multi-database networks. *Data and Knowledge Engineering*, 36(3):251–281, 2001.

9 Requested budget

Funding is requested for two OiO positions and a scientific programmer, for a period of four years, starting from January 2001. This yields the following budget request:

Item	Number	Amount/item	Sub-total
OiO Salary costs	2	Kf. 228,00	Kf. 456,00
OiO Travel allowance	2	Kf. 7,35	Kf. 14,70
Scientific programmer	1 x 4 Yrs	Kf. 90,00	Kf. 360,00
Total:			Kf. 830,70

References

[1] Jon Barwise and John Etchemendy. Information, infons, and inference. In Robin Cooper, Kuniaki Mukai, and John Perry, editors, *Situation theory and its applications*, volume 1 of *CSLI Lecture Note Series*, pages 33–78. Center for the study of language and information, CSLI, 1990.

[2] F.C. Berger, P. van Bommel, and Th.P. van der Weide. Ranking strategies for navigation-based query formulation. *Journal of intelligent information systems*, 12(1), 1999.

[3] F.C. Berger and T.W.C. Huibers. A framework based on situation theory for searching in a thesaurus. *The New Review of Document and Text Management*, 1:253—276, 1995.

[4] Tim Bray, Jean Paoli, C. M. Sperberg-McQueen, and Eva Maler. Extensible markup language (xml) 1.0 (second edition). Technical report, World Wide Web Consortium, October 2000.

[5] The DARPA Agent Mark-up Language, 2001.

[6] Dublin Core Metadata Initiative, 1999.

[7] Knowledge Query and Manipulation Language, 2001.

[8] M. Lalmas. Logical models in information retrieval: Introduction and overview. *Information Processing & Management*, 34(1):19–33, 1998.

[9] Ontology inference layer, 2001.

[10] M.P. Papazoglou, H.A. Proper, and J. Yang. Landscaping the information space of large multi-database networks. *Data & Knowledge Engineering*, 36(3):251–281, 2001.

[11] H.A. Proper and P.D. Bruza. What is Information Discovery About? *Journal of the American Society for Information Science*, 50(9):737–750, July 1999.

[12] Resource Description Framework (RDF) Model and Syntax Specification, 2000.

[13] F. Sebastiani. On the role of logic in infromation retrieval. *Information Processing & Management*, 34(1):1–18, 1998.

[14] Th.P. van der Weide, T.W.C. Huibers, and P. van Bommel. The Incremental Searcher Satisfaction Model for Information Retrieval. *The Computer Journal*, 41(5):311–318, 1998.