

Master:	Information Sciences	Supervisor:	dr. Marieke van Erp
Student name:	ing. Youssef Ibrahimi	Second supervisor:	Tina Mioch, MSc
Student number:	1924761	Company:	TNO
Date:	2011/04/01	Document status:	<i>Final</i>

Context

In 2011, approximately 60 projects will be carried out within the Perceptual and Cognitive Systems department of TNO¹.

Within these projects, methodologies, approaches, problems, solutions, and obtained results can differ from each other. This content-specific project information can however be used again as a starting point for other (new) projects or as an information source for researchers. By content-specific project information; project information such as methodologies, obtained results, and approaches are meant. Project information such as budget and timeframe are out of scope of this study.

In the present situation, researchers can use a wide range of information systems to find *general* project related information. These systems include; (1) network shares with hundreds of documents; (2) a digital library containing primarily publications; (3) “TNO Spider/City”, an intranet portal focusing on general project information, such as project members, internal project numbers and the possibility to look up financial information regarding projects; (4) “Yammer”, a Facebook like internal communication system; and (5) “ScienceDirect/Scopus”, searchable databases with scientific content. The documents on the network shares can be final version documents (e.g. deliverables) or documents in progress (e.g. draft versions of work documents). However, if a researcher wants to easily and quickly find content-specific information about a project, he or she does not have an accessible way in doing this in the present situation. For example, questions such as - what are the obtained results and which approaches contributed to the results? Which projects are similar to my project? –are relevant for researchers.

Problem statement

Within TNO applied science is the core business, therefore each project milestone, idea, method, or intermediate result, can be valuable to the entire organisation. Insight in what has been done in a project is essential for other projects, for example so employees do not “invent the wheel again”.

The present situation lacks content-specific information access: researchers do not have a quick and easy way to access this kind of information. A lack of access to relevant previous work can result in developing the same method, approach or idea multiple times. In turn, developing the same method or approach multiple times can have a negative effect on time, effort and money spent within a project.

¹ <http://www.tno.nl/index.cfm?Taal=2>

Absence of a formal model to structure content-specific project information may pose the danger of developing the same method or approach multiple times. Therefore, we formulate the following challenge for this project:

Can we capture project information in such a way it becomes more accessible and thus reusable for researchers?

This study extends previous work on structuring research information within the Information Retrieval domain [1, 2, 3]. In particular, we will explore how research information can be structured within TNO, by using the Information Retrieval techniques: automatic document classification and – segmentation [4].

Research sub-questions

The derived sub-questions from the main problem statement are:

*1a. Can we automatically identify units of project information **among** documents?*

*1b. Can we automatically identify units of project information **within** documents?*

We consider units of project information on two levels. The first level is the document level, in which we consider documents as a whole. The second level is the content level, in which we define units of project information within one document. These two levels will present two different, but parallel phases within this master project.

2. Can we automatically map units of project information to an ontology?

We consider a map of project information units as a collection of core concepts. Identifying and defining these core concepts, and mapping them to an ontology is the aim of this sub-question.

Methodology

In this master project, a combination of multiple research methods will be used. The research methods we will be using are listed in table 1 with their corresponding goals.

Research method	Goal
Open-interview	- Collecting background information about the problem situation and HUMAN project.
Document-review and analysis	- Collecting background information and scientific literature about the problem situation.
Experiment	- Identifying units among and within documents - Mapping units to ontology concepts - Evaluating proposed model and demonstrator (user-based)

	- Testing proposed demonstrator (technical).
Demonstrating	- Building demonstrator

Table 1 – Research methods

Planning

This master project will be carried out from 25 February 2011 to 29 July 2011 (approx. five months). Table 2 describes the activities and amount of workweeks for completing each particular activity. Also, the most important milestones with their corresponding deadline are described.

*Note: In table 2, **writing** the paper, is not mentioned as a standalone activity, rather this activity will be carried out throughout the whole master project. Each section of the paper will be written when finishing a particular activity, which could contribute to the final paper.*

Activity	Approx. amount of workweeks	Deadline
Collecting background information about the problem situation.	1 week	
Collecting and reading scientific literature.	2 weeks	
Writing research proposal.	3 weeks	
Collecting and reading background information about the HUMAN project, creating snapshot of dataset.	1 week	
Milestone 1 - Research proposal (final)		31 March 2011
Milestone 2 – Snapshot of dataset		31 March 2011
Document classification	5 weeks	
Milestone 3 – Identified units among documents		06 May 2011
Document segmentation, mapping of units to an ontology.	5 weeks	
Milestone 4 – Identified units within documents and mapped to an ontology		10 June 2011
Building demonstrator, experiments	5 weeks	
Milestone 5 – Project information accessible through demonstrator		15 July 2011

Polishing up paper	2 weeks	
Milestone 6 – Final version paper		29 July 2011
Milestone 7 – Presenting results at TNO & VU		TBD

Table 2 – Planning

References

- [1] Bechhofer, S., De Roure, D., Gamble, M., Goble, C. & Buchan, I. (2010). “Research Objects: Towards Exchange and Reuse of Digital Knowledge.” In: *The Future of the Web for Collaborative Science (FWCS 2010), April 2010, Raleigh, NC, USA.*
- [2] Hyvönen, E., Saarela, S. & Viljanen, K. (2003). “Ontogator: Combining View- and Ontology-Based Search with Semantic Browsing.” In: *Proceedings of XML Finland 2003, Open Standards, XML, and the Public Sector; Kuopio, October 30-31, 2003.*
- [3] Soldatova, L. N. & King, R.D. (2006). “An Ontology of Scientific Experiments.” In: *Journal of The Royal Society Interface, Vol. 3, No. 11. (22 December 2006)*
- [4] Manning, C. D., Raghavan, P. & Schütze, H. (2009). “An Introduction to Information Retrieval.” *Cambridge University Press.*

dr. Marieke van Erp

Tina Mioch, MSc

ing. Youssef Ibrahimi
