

INFORMATION SOCIETY TECHNOLOGIES
(IST)
PROGRAMME

Project IST-2001-37057 MKMNet

**Deliverable 2.5:
Evaluation of D2.4**

J.A.Padget (University of Bath)

Project Acronym: MKMNet

Proposal/Contract no.: IST-2001-37057 MKMNet

This document is a placeholder to make the deliverable package complete.

There is no deliverable D2.5 as described in the Technical Annex of the project. An explanation for its absence appears below along with some context for WP2 and the problems that have apparently arisen.

The project reviewers remarked:

- **Point 7 (P7)**

The Work Package 2 (WP2) had as one of its main objectives the “development of stochastic models and corresponding estimators for identification clouds of both formulas and key phrases” (deliverable D2.2). D2.2 presents a model for the growth of indexes, then proceeds with the introduction of the concept of identification clouds and its possible applications, ending with “Probably the most crucial issue to be addressed at this stage is the formulation of a good probabilistic model of ID clouds complete with statistical estimators”.

So, adequate modeling of identification clouds for formulas and key phrases could be an issue here (incidentally, D2.2 contains at the end of Section 13 the sentence

“Some preliminary work on formula recognition using identification clouds is planned in the EC project [8].”

where ref. [8] is the MKMNET project itself:

8. J Davenport, a.o., MKMNET. Mathematical knowledge management network, Project IST-2001-37057. September 2002 - December 2003. 2001.)

Michiel Hazewinkel replies:

Adequate modelling of identification clouds is indeed an issue and a far harder one than was originally thought (by me). The project has not been abandoned however. Originally the plan was to carry out further work within the framework of the successor to MKMnet. Unfortunately that project was not retained for funding and so is now irrelevant. This is also the cause of the mistaken reference to [8]; it should have been to the planned successor of [8]. Research on the stochastic modelling of identification clouds is now under way at the Inst. of Mathematics, Lithuanian Academy of Sciences, Vilnius, Lithuania in the form of a PhD thesis project. The persons involved are Prof. Rimantas Rudzkis (thesis adviser (together with myself) Vaidas Balys (PhD student).

- **Point 8 (P8)**

Deliverable D2.3 was supposed to produce “Software tools for the (semi-)automatic generation of identification clouds (given an adequate list of key phrases to be used)”. D2.3 seems instead to contain a program (fizkf) to find keyphrases in a collection of abstracts.

Michiel Hazewinkel replies:

The keyphrase program is also the tool for the semiautomatic generation of identification clouds (given an adequate list of key words that are suitable for that purpose). All one has to do is to feed fizkf with the abstracts (chunks of texts) that contain a given key phrase and the list of key words from which the identification clouds should come.

- **Point 9 (P9)**

Deliverable D2.4 had to include “Identifications clouds for a substantial research level part of mathematics”. However, the current D2.4 seems to include a list of key phrases (incidentally, with duplicates as well).

Michiel Hazewinkel replies:

The intention was to do this for the subfield 'Combinatorics' of Mathematics. And in fact there is a key phrase list for that part of mathematics (both published and electronic) and a corresponding list of words from which identification cloud members can be taken. The Vilnius group is working with these data now. Both are freely available. What has not been done is to actually run fizkf a few thousand times to actually generate identification clouds for the available key phrases from combinatorics. That can be done of course (and automated). In the absence of an adequate stochastic model, notably concerning the distribution (distances) of identification cloud members to (present or not) key phrases this is a fairly empty exercise.

- **Point 10 (P10)**

The output of Task 2.4 (Validation), i.e., D2.5 and D2.6, seems missing.

Michiel Hazewinkel replies:

I wrote about this in my report. What we have learned from MKMnet is that the matter of identification clouds is far trickier than one would naively suppose (see also my article 'Dynamic stochastic models for indexes and thesauri, identification clouds, and information retrieval and storage, In: H Gzyl (ed.), Applied probability, to be published by KAP later this year. I still firmly believe that identification clouds will turn out to be a most valuable tool. But getting the tool to work and implemented is a far far larger job than I thought some years ago.