A Qualitative and Quantitative Evaluation of Adaptive Authoring of Adaptive Hypermedia

Maurice Hendrix and Alexandra Cristea

Department of Computer Science, The University of Warwick, Coventry CV4 7AL,
United Kingdom
{maurice,acristea}@dcs.warwick.ac.uk

Abstract. Currently, large amounts of research exist into the design and implementation of adaptive systems. The complex task of authoring of such systems, or their evaluation, is addressed less. We have looked into the causes of this complexity. Manual annotation is a serious bottleneck for authoring of adaptive hypermedia. All means for supporting this authoring process by reusing automatically generated metadata would therefore be helpful. Previously, we proposed the integration of a generic Adaptive Hypermedia authoring environment, MOT, into a semantic desktop environment, indexed by Beagle++. Based upon this approach, a prototype was constructed. The approach in general, as well as the prototype in particular, where evaluated through both qualitative and quantitative experiments. This paper is a synthesis of our work so far, describing theoretical findings and hypotheses, their implementation in short, and finally, the combined results of the evaluations.

Keywords: Authoring; Adaptive Educational Hypermedia, CAF; Evaluation, Metadata, RDF, Semantic Desktop, Semi-automatic adding, MOT, Beagle++.

1 Introduction

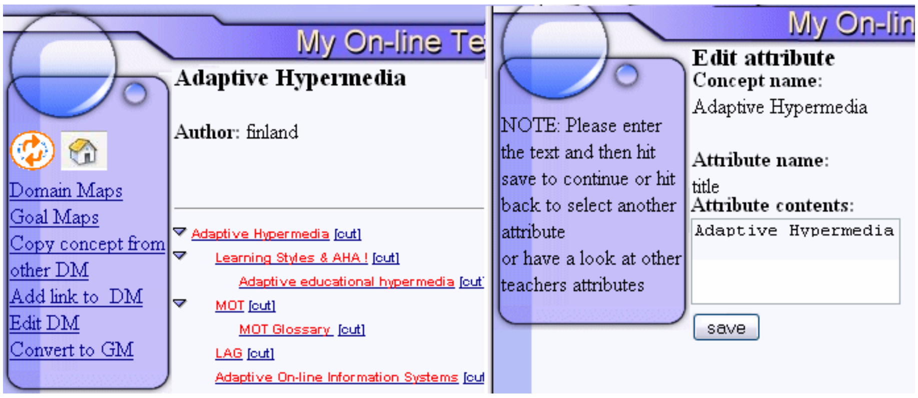
Authoring Adaptive Hypermedia can generate valuable personalized (learning) experiences [6], but it is known to be a difficult and time-consuming task [7]. A possible solution to this problem is to use automatically generated authoring as much as possible, and there has already been research into how to automatize authoring in different ways [3, 13]. A good basis is to use already annotated resources (such as provided by the Semantic Desktop [9,22]), which can be automatically retrieved when necessary, as dictated by the authoring process. In a Semantic Desktop, resources are categorized by rich ontologies, and semantic links can express various kinds of semantic relationships between the resources. For a file representing a paper, for example, the Semantic Desktop stores not only a filename, but also information about where it was published, when, by whom, etc. These metadata are generated automatically, and stored in the user's personal data store in RDF format. This rich set of metadata makes it easier for the user to semi-automatically retrieve appropriate material for different contexts, for example, when a teacher wants to select materials that fit a certain lecture. In this context, an author has to create some basic lesson material, which then serves as a framework for the final lesson to be created.

E. Duval, R. Klamma, and M. Wolpers (Eds.): EC-TEL 2007, LNCS 4753, pp. 71–85, 2007.

In [17] and [16] and [18] we previously described the interaction and exchange of data between the adaptive hypermedia authoring environment MOT [10,19], and the Beagle++ environment [1,2,8]. Here we are going to review only the essential parts allowing comprehension of the evaluation work.

MOT [10] is a *concept*-based adaptive educational hypermedia authoring environment with adaptive authoring support. It is an advanced system for authoring personalized e-courses based upon the LAOS [11] authoring framework and offers a web forms interface for the authoring process. The main parts of the LAOS framework it offers are the Goal & Constraints Model and the Domain Model. Elementary blocks of content are represented in the Domain Map, and in the Goal & Constraints Map blocks from domain maps are brought together. This forms an initial version of what end-users (students taking a course) will see before any adaptation is applied to it.

Currently, there are two versions of MOT. Fig. 1 shows a snapshot of the interface for authoring Domain Maps in the new MOT version.


My On-line Te
Domain Maps
Goal Maps
Copy concept from other DM
Add link to DM
Edit DM
Convert to GM
Adaptive Hypermedia [cut]
Learning Styles & AHA ! [cut]
Adaptive educational hypermedia [cut
MOT [cut]
MOT Glossary [cut]
LAG [cut]
Adaptive On-line Information Systems [cut
My On-lin
save

Adaptive Hypermedia

Author: finland

NOTE: Please enter the text and then hit save to continue or hit back to select another attribute or have a look at other teachers attributes

Edit attribute

Concept name:
Adaptive Hypermedia

Attribute name:
title

Attribute contents:
Adaptive Hypermedia

Fig. 1. Domain authoring in the new MOT

Beagle++ is an advanced search and indexing engine for the semantic desktop. It is an extension to the Beagle [1,8] search tool which generates and utilizes metadata information, and keeps a metadata index of all files. Initially, extraction tools are used to populate this metadata index.

Our approach uses a standalone Java application called the *Enricher* (or *Sesame2MOT conversion*) to implement the link between the MOT and Beagle++ systems described above. As is introduced in more detail in section 3, the Sesame2MOT conversion works by reading current courses from MOT in an XML format called CAF (Common Adaptation Format) and querying the Metadata index kept by Beagle++.

In this paper we evaluate our approach to conversion in general, and the Enricher prototype [18], which has been constructed, in particular. As MOT has a new version, we also evaluate whether this new version is indeed preferable, and whether we should base development of our prototype on this version of MOT.

The remainder of this paper is organized as follows. Based on [18], section 2 and 3 give a short illustrative scenario and a brief introduction of the system setup. The evaluation consists of both quantitative experiments [18], and qualitative experiments. The quantitative experiments consist of a SUS-questionnaire [5] testing the system usability [15], and of a focussed questionnaire [14]. The evaluation methodology used for these experiments is described in section 4, and in section 5 the results of the evaluation are presented. Finally, in section 6 we discuss what these results mean for our approach in general and for the prototype in particular.

2 Motivational Scenario

We use a scenario for adaptive authoring that builds upon a combination of automatically and manually generated metadata, as introduced in [18].
Prof. Jones is a hypothetical lecturer who is preparing a new course. His university allocates a limited amount of time to this. He uses MOT,

- because he considers it useful to be able to extend the course in the future with more alternative paths guided by adaptivity, and
- because he wants to benefit from automatic help during authoring.

This takes slightly more time than static course creation, as the course has to be divided into conceptual entities with explicit, independent semantics and semantic labelling.

The advantage is that the adaptive authoring system can afterwards automatically enrich the course based on pedagogical strategies. For example, the version created by Jones can be considered as the version for beginner students. For advanced students, such as those wishing to pass the course with high marks, the adaptive authoring system can use the Semantic Desktop search to automatically find on Jones' desktop any existing scientific papers that are relevant to the current course. These papers could then be used as alternative or additional material to the main storyline of the static course. This mechanism builds upon the following assumptions.

- Since Jones is a specialist in the subject he is teaching, he both publishes and reads papers of interest on the subject, which are likely to be stored on his computer.
- His collection of papers can be considered as useful extra resources for the current course, and can therefore be reused in this context.
- The storing process has taken place over time, and Jones may not know exactly where on his computer each article relevant to the current course is.
- Jones has been using the Beagle++ Semantic Desktop System to store both papers and all relevant metadata automatically in RDF format.

This situation can be exploited by the authoring tool; a search will find some of Jones' own papers on the course topic, as well as some papers written by his colleagues on the same topic. He may have saved these papers by himself, received them by e-mail from a colleague, or may have bookmarked them using his browser. In order for these retrieved resources to be relevant to the course, two conditions have to be fulfilled:

- the domain concept in the course where each resource is most relevant has to be found (*the right information*), and next,
- the resource must be bound to that particular concept (*in the right place*).

How can Jones find the right resource and store it in the right place? The automatic search can take place via the keywords labelling both the course components created by Jones and the matching keywords labelling the papers and resources on his desktop. How Jones can enrich his course automatically, without much extra work, as well as keep at all times overall control and a coherent overview, is described in more detail in [18]. The following sections evaluate this specific approach, as well as a prototype for it which has been implemented, and discuss possible improvements.

3 The Approach and System Setup

In this section we shortly review our method and system setup. As can be seen in Fig. 2, Beagle++, the Semantic Desktop Environment used in our prototype, stores all metadata in the Sesame RDF database [21]. All Beagle++ components that generate metadata (for example, the email, publication, web cache and file metadata generators) add the metadata to this database. All Beagle++ components which use metadata (for example, the search and indexing module, the ranking module or the browsing modules) retrieve their data from this repository, and, in some cases, write back new data (such as the PageRank [8] value for documents or other resources).

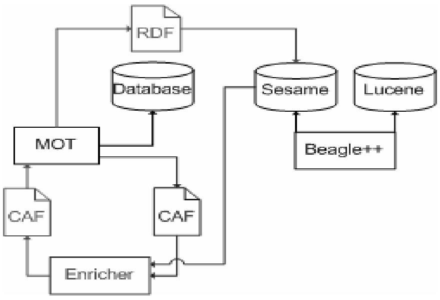


Fig. 2. System overview

It is easy to accommodate additional modules in this environment by writing appropriate interface components, which read and write from this repository. This is what we have done for MOT [10,19]. We have focused on the semi-automatic addition of articles stored on the user's desktop to a MOT lesson [10]). This represents an instantiation of the concept of adaptive authoring: authoring that adapts to the author's needs. In MOT, the addition is done to an existing lesson. Based on pedagogic goals, the author can then process the data, by adding more information on the article after the conversion. These additions can then be fed back into the RDF store, if necessary. We use CAF [12], a system-independent XML exchange format, to simplify the transformation process from RDF to the MOT MySQL storage format.

3.1 Enrichment of the Lesson and Domain Model

MOT is mainly a tool for authoring educational (adaptive) material, thus the internal information structures are based on strict hierarchies. When enriching the domain maps and lessons, we therefore aim at getting the right information in the right place in this hierarchy. To achieve this, the program first queries the Sesame database, using as search terms *title* and *keywords* of each domain concept found in the current existing lesson. The basic RDF query in the SeRQL [4] language looks as follows:

```
   SELECT x FROM x {p} y WHERE y LIKE “*keyword*” IGNORE CASE
```

Some alternative retrieval methods have been studied, implemented and evaluated, as follows.

3.1.1 Concept-Oriented Versus Article-Oriented

For computing the mutual relevance between an article and a concept, in order to decide the appropriate place of articles in the concept hierarchy, we previously [16], [17] have developed two slightly different theoretical alternatives, as follows.

3.1.2 Concept-Oriented Relevance Ranking Method

This method computes relevance of an article for a given concept as follows:

$$rank(a,c) = \frac{|k(c) \cap k(a)|}{|k(a)|} \tag{1}$$

where:
rank(a,c) is the rank of article a with respect to the current domain concept *c*;
k(c) is the set of keywords belonging to the current domain concept *c*;
k(a) is the set of keywords belonging to the current article *a*;
|*S*| = the cardinality of the set *S*, for a given set *S*.

This formula is *concept-oriented*, in the sense that articles ‘battle’ for the same concept: a given article is placed in the appropriate place in the hierarchy by it.

3.1.3 Article-Oriented Relevance Ranking Method

This method computes the relevance of a concept to a given article as follows (notations same as above):

$$rank(a,c) = \frac{|k(c) \cap k(a)|}{|k(c)|} \tag{2}$$

The equation shows how many of the keywords (shared by the article and the concept) are present in the concept, relative to the number of keywords present in the concept. As an extreme example, if two concepts share the same number of keywords with a given article, but one concept has less keywords than the other, ,the former concept will have a higher rank and ‘win’ the article, as it is more focussed on the shared keywords than the latter one.

3.1.4 Sets Versus Multisets

Next, once the formula is chosen, there is another possible distinction to be made: the cardinality of the intersection can take two forms; one set-based (with intersection operation on sets, as defined above), and one with *multisets* or bags (and the respective intersection operation on bags). The reason to use sometimes bags instead of sets is that the number of times keywords appear in certain texts can be relevant in itself (not just which keywords). A text containing a greater number of occurrences of a specific keyword could be a better match for that keyword than a text with only one occurrence of the respective keyword. The author can choose between the two.

3.1.5 Duplicates Handling for Sibling Concepts

The same resource may be relevant in more then one place within the hierarchy. In that case, the resource will be added to the place where it has the highest relevance, by default. If there are more places in the hierarchy with a value equal to the highest relevance, the current implementation yields the one with the higher position in the tree to win. For siblings with the same position in the tree, and with the same (highest) relevance, a decision has to be made: either to allow duplicates, or to select randomly one of the candidate sibling concepts and allocate the resource to it. This decision depends on how the option of concepts pointing to the same resources looks like from the point of view of the current pedagogic strategy. Therefore, the current implementation allows the author to decide, via a switch called '*allow duplicates*'.

3.1.6 Adding Meta-Data as Separate Concepts or as Attributes

The retrieved metadata also has a structure. For example, a retrieved paper might have a location it was presented at and a year it was presented in. This metadata can be added either as attributes of the new article-concept in MOT, or as a set of new sub-concepts, with their own attributes. The author can switch between these two possibilities in the Enricher program.

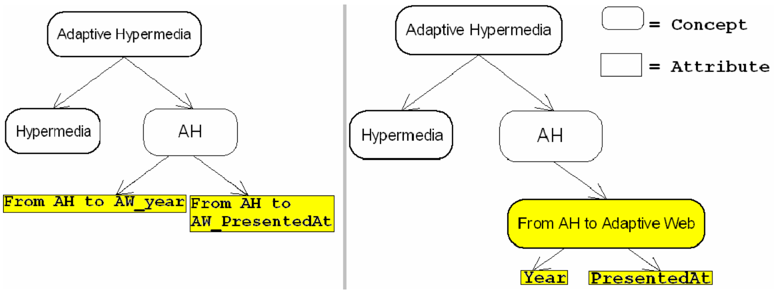


Fig. 3. result without; right, result with the 'add medatada as separate concepts' option

4 Evaluation

The evaluation of the conversion process, Enricher, and new MOT system has taken place in two steps so far.

4.1 First Evaluation Step

The first step was a small-scale qualitative experiment with 4 PhD students of the IMPDET course organized by the University of Joensuu in Finland, based on the think-aloud method [20]. As can be found in [18], the system was mainly *understood*, but respondents were unable to provide feedback on the method itself. Some *shortcomings of the user interface* of the prototype were identified and corrected as a result of this first evaluation step [18].

4.2 Second Evaluation Step

The second evaluation was of a much larger scale and conducted at the Politehnica University of Bucharest in January of 2007, and took place within an intensive two-week course on “Adaptive Hypermedia and The Semantic Web”, which was delivered as an alternative track to the regular “Intelligent Systems” course. The students were 4th year undergraduates in Engineering studies and 2nd year Masters Students in Computer Science, all from the English-language stream. Firstly, basic knowledge on Adaptive Hypermedia and Semantic Web was addressed – the first course week was dedicated to theory, and finished with a theoretical exam. Out of the initial 61 students, only the students with a satisfactory knowledge of the theory were selected to continue with the practical part. The 33 students that passed the theory exam worked with the two versions of MOT (*old* versus *new*) and the Sesame2MOT (Enricher) conversion, the prototype constructed for the automatic authoring approach [16]. After these experiments, they were requested to submit questionnaires, to answer both generic and specific issues regarding the automatic generation of adaptivity and personalization. The questionnaires consisted of five parts; first a *SUS* [5] questionnaire for each of the three systems, and then two more specific questionnaires, for the *Sesame2MOT conversion* and for the *comparison of the new version of MOT with the previous version*. Here, we mainly focus on the *usability aspect* targeted in the evaluation process.

4.3 Hypotheses

We based our evaluation firstly on a number of generic, high level hypotheses, as follows:

1. The respondents *enjoyed* working as authors in the system.
2. The respondents *understood* the system.
3. The respondents considered that *theory and practice match* (for Sesame2MOT).
4. The respondents considered the *general idea of Adaptive Authoring useful*
5. The respondents have acquired *more knowledge* than they initially had with the help of the theoretical course (explanation) part.
6. The *new MOT has a better usability then the old version*; hence we should base further developments on this version of MOT.
7. The respondents’ overall *preference* (from a usability perspective) is as follows, in increasing order: old MOT, new MOT, Sesame2MOT.
8. The *user interface* of both version of MOT is *sufficient*.
9. The *upload functionality* in the new version of MOT is a *necessary* improvement.

We refined these into more specific, lower granularity hypotheses (see Table 1), which ultimately generated our questions for the questionnaires. To explain the construction of the sub-hypotheses, let's take, for instance, hypothesis 3. There, we check the matching between theory and practice, i.e., between theory and the implementation. For the Enricher application, from a theoretical point of view, we have defined different ranking methods and other options, such as allowing duplicates or not between the imported articles, etc. These have been implemented as options for the user to select, and therefore, in this particular case, matching theory and practice means that these methods render *different results*, firstly, and secondly, that *these different results should be as the theory has predicted*. Therefore, sub-hypothesis 3.4, and its sub-hypotheses, 3.4.1, 3.4.2 and 3.4.3 emerged. As said, the hypotheses and sub-hypotheses feed into and determine the question.

Respondents where given the option to comment on their preferences, in order to also gain qualitative feedback. We also directly asked for comments on all three systems, as well as the approach in general.

4.4 The Questionnaires and Their Rationale

As said, we used two main types of questionnaires to estimate the truth value of our hypotheses. One type is standard questionnaires, such as SUS (System Usability questionnaire, [5]), and the other type is questionnaires built by ourselves, targeting specific aspects of the main hypotheses.

Moreover, we used two types of questions: one was numerical or multiple choice question (the latter also mapable on a numerical scale), and the other one was open-ended questions trying to extract respondents' opinions as well as possible aspects we have missed in the numerical questions.

Finally, out of the questionnaires we built ourselves, questions targeted two main issues: the *theory* behind the system and separately, the *system* itself.

4.4.1 SUS

As we were in fact evaluating three systems (old MOT, new MOT and the Sesame2MOT conversion) we applied the SUS questionnaire three times, once for each of them. SUS stands for System Usability scale [5] and gives a measure for comparing the usability of different systems. A SUS questionnaire consists of 10 questions with a 1 to 5 scale of agreement. By alternating positive and negative questions, respondents are forced to think about their agreement to each question. For the positive questions the score is the *agreement level-1*, for the negative questions the score is *1-the agreement level*. This now yields the *SUS score*, which is ideal for comparing the generic usability of different systems. This scores give us however little insight into where exactly the problems lie; therefore it is advisable to design more specific, targeted questions to extract these answers.

4.4.2 Sesame2MOT

We therefore also constructed a more specific questionnaire for the Sesame2MOT conversion. Here, we directly asked respondents whether they consider the general idea of semi-automatic authoring useful, as well as enjoyable. We also asked whether they understood each of the ranking methods and selection options, whether these did what was expected and whether their selection choice had any visible influence on the conversion process.

4.4.3 OLD Versus NEW MOT

For the comparison of the OLD and NEW MOT we constructed a more specific questionnaire as well. To gain more insight into specific issues than a SUS questionnaire can provide, we asked respondents directly:

- whether they consider the general idea of authoring support of adaptive hypermedia useful, and enjoyable;
- whether they enjoyed working with both MOT versions;
- whether they understood how both MOT versions work;
- whether they thought the user interfaces were sufficient;
- whether they thought both MOT versions where easy to use;
- whether they thought both MOT versions make Adaptive Hypermedia creation easier.

We also directly asked them their overall preference and their preference for Domain Model and Goal Model editing [11].

5 Evaluation Results and Discussion

As said in section 4, for testing our hypotheses we used two different types of questionnaires for all three systems, a SUS questionnaire to gain insight in the systems' usability and more specific questionnaires, to target our hypotheses more directly. In this section we will discuss the results obtained from both types of questionnaires, as well as discuss the qualitative feedback obtained both from the IMPDET experiment as well as from the experiment in Bucharest.

For testing our hypotheses against the quantitative feedback obtained, we used numerical averages, and tested their significances with the help of a T-test. We have used the parametric test based on the assumption of equidistant points of measurement of the interval scale. We assumed a confidence of 95% would be reasonable. The T-test establishes whether the difference between a value and the average of a sample or the averages of two samples is significant. For a hypothesis to be confirmed the difference needs to be significant and be in the direction the hypothesis suggests. For example, if we test the difference between pre-test and post-test exam and have as hypothesis that respondents did better in the post exam, the average from the post exam must be higher and the difference between pre- and post exam needs to be significant. The difference between two samples or a sample and a value is considered *significant* if the probability P that the difference arose by chance is $P<0.05$.

5.1 Questionnaire Feedback

In order to obtain numerical averages for testing our hypotheses, we mapped the multiple-choice answers of the questionnaires as follows: '*Yes*' was mapped to 1, '*no*' to -1 and '*mostly*' to 0. Hence the average was always 0 and the T-test was applied by comparing against the neutral result of 0. Below we present a table with each hypothesis, T-test results (*T* value, degrees of freedom *Df*, Mean *M*, probability *P*) and whether the results show that it was confirmed or not. The main hypotheses are shown in bold. Their result is obtained by combining the results of the sub-hypotheses.

Table 1. Sesame2MOT Conversion hypotheses results

Nr.	Hypotheses	T	Df	M	P	Confirmed (M>0; P<0.05)
1	**The respondents enjoyed working as authors in all systems**					**Not confirmed**
	Sesame2MOT	2.709	31	0.438	0.011	Confirmed
	OLD MOT	1.161	32	0.121	0.254	*Not confirmed*
	NEW MOT	3.546	32	0.333	0.001	Confirmed
2	**The respondents understood all systems.**					**Confirmed**
	Respondents understood Sesame2MOT					Confirmed
	Respondents understand the option: Concept oriented	4.458	31	0.625	0.000	Confirmed
	Article oriented	3.788	31	0.563	0.001	Confirmed
	Allow duplicates	10.063	31	0.875	0.000	Confirmed
	Compute resources as set	5.271	31	0.688	0.000	Confirmed
	Add meta-data as separate concepts	6.313	31	0.750	0.000	Confirmed
	Respondents understood OLD MOT	5.899	32	0.576	0.000	Confirmed
	Respondents understood NEW MOT	6.197	32	0.546	0.000	Confirmed
3	**The respondents considered that theory and practice match (for Sesame2MOT).**					**Confirmed**
	The two ranking methods (concept-, article-oriented) do deliver different results.					Confirmed
	Concept Oriented	6.313	31	0.750	0.508	Confirmed
	Article Oriented	6.313	31	0.750	0.508	Confirmed
	Ranking methods (concept-, article-oriented) in line with the theory.		31			Confirmed
	Concept Oriented	2.252	31	0.375	0.032	Confirmed
	Article Oriented	2.709	31	0.438	0.011	Confirmed
	The different options influence the result		31			Confirmed
	Allow duplicates	7.760	31	0.813	0.000	Confirmed
	Compute resources as set	3.215	31	0.500	0.032	Confirmed
	Add meta-data as separate concepts	6.313	31	0.750	0.000	Confirmed
	The results of the conversion are in line with the theory					Confirmed
	The two ranking methods					Confirmed (see above)
	Allow duplicates	7.760	31	0.813	0.000	Confirmed
	Compute resources as set	2.252	31	0.375	0.032	Confirmed
	Add meta-data as separate concepts	4.458	31	0.625	0.000	Confirmed
4	**General idea useful**	**15.000**	**31**	**0.938**	**0.000**	**Confirmed**
5	**The respondents have acquired more knowledge than they initially had**	**25.59**	**57**	**5.75**[1]	**0.000**	**5.75 out of 10; p=0.00<0.05; t=25.59),**

[1] An average increase in grades occurred of 5.75, out of a possible 1-10 with 10 being the best.

Table 1. (*continued*)

6	**The new MOT is more usable then the old version; we should base further developments on this version of MOT.**					**Confirmed**
	Respondents preferred NEW MOT (over OLD MOT) for authoring	5.600	32	0.636	0.000	Confirmed
	Respondents would choose NEW MOT (over OLD MOT) for DM authoring in general	9.339	32	0.788	0.000	Confirmed
	Respondents would choose NEW MOT (over OLD MOT) for GM authoring in general	9.238	32	0.727	0.000	Confirmed
	Respondents prefer all editing sub functions of NEW MOT (over OLD MOT) for DM authoring					Confirmed
	Adding/modifying DM sub-concepts	3.213	32	0.394	0.003	Confirmed
	Deleting DM sub-concepts	5.555	32	0.545	0.000	Confirmed
	Adding/modifying DM attributes	3.922	32	0.455	0.000	Confirmed
	Deleting DM attributes	5.899	32	0.576	0.000	Confirmed
	Respondents prefer all editing sub functions of NEW MOT (over OLD MOT) for GM authoring					Confirmed
	Conversion from GM	10.902	32	0.788	0.000	Confirmed
	Adding/modifying GM labels	10.000	32	0.758	0.000	Confirmed
	Deleting GM labels	8.579	32	0.697	0.000	Confirmed
	Adding/modifying GM weights	9.238	32	0.728	0.000	Confirmed
7	**The user interface of both version of MOT is sufficient.**					**Not confirmed**
	OLD MOT user interface is sufficient	-1.715	32	-0.152	0.096	Not confirmed
	NEW MOT user interface is sufficient	1.971	32	0.057	0.057	Not confirmed
8	**The upload functionality in the new version of MOT is a necessary improvement.**	9.339	32	0.788	0.000	**Confirmed**

As we have seen, most hypotheses have been confirmed based on the current data. The Sesame2MOT conversion is indeed considered useful and in line with the theory. Its options are understood. Respondents agreed strongly with most of our hypotheses, with all means above zero. Looking at the ones with lower scores, such as concept-oriented and article-oriented method, as well as computation of resource as set, they were less sure in their statements. This is probably due to the fact that they did not work with these options enough. This shows that more targeted evaluations may be necessary to establish without a doubt the acceptance rate of these features. Also respondents where not clear about enjoying working with the old version of MOT. This might be related to either the formal setting of the course or to the difference in user interface.

5.2 SUS Usability Feedback

With a SUS score the usability of systems can be compared. The average score can be contrasted and visual graphs can be constructed to identify specific problem points. The questions (which are alternately positive and negative) are plotted on a circle using a scale from 1 (strongly disagree) to 5(strongly agree), with 1 in the centre and 5 at the border. If the results for the questions are placed on the scale, the ideal system should show a perfect star shape, as positive and negative questions alternate. In Fig. 4 below, the SUS scores for the different systems are shown in such a SUS graph. The figure shows that the systems have relatively similar scores. Visible differences are that Sesame2MOT seems to have a higher perceived learning threshold, whereas the old MOT is considered more inconsistent and more cumbersome.

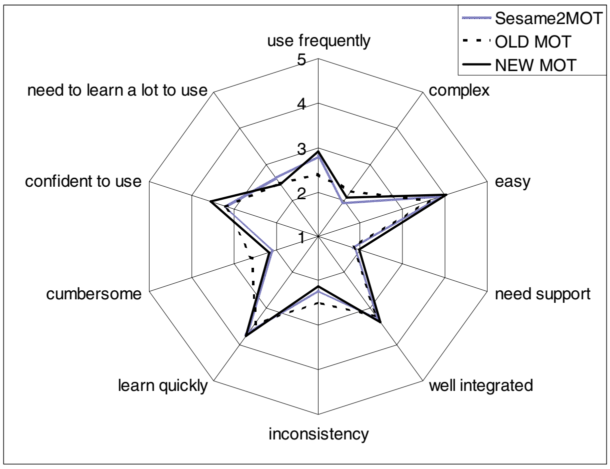


Fig. 4. SUS score for the three systems

Normalized responses range from 0 to 4, see [5]. Thus we applied a T-test comparing the normalized results against the average neutral value of 2. A paired T-test was used, since we compared answers of the same sample (group of students). Moreover, the main hypotheses were further broken into sub-hypotheses.

5.2.1 General Hypotheses

Below we list some of the main hypotheses, which are related to the SUS questionnaires, and comment on how much they are supported by the SUS results.

1. *The respondents enjoyed working as authors in the three systems from a usability perspective.*
   The results for the old MOT (mean 2.39 (expected >2); p=0.519>0.05; t=0.65) and Sesame2MOT (mean 2.78 (exp >2); p=0.095>0.05; t=1.72) on enjoyment were not significant. The respondents did significantly enjoy working with the new MOT (mean 2.97 (exp >2); p=0.01<0.05; t=2.66). The

hypothesis as a whole cannot be supported. This is possibly due to the formal setting of the course.

5. *The respondents' overall preference, from a usability perspective, is as follows, in increasing order: old MOT, new MOT, Sesame2MOT.*
   The results on learning preferences, and the preference for Sesame2MOT over the new MOT (difference -0.07 (>0 exp.); p=0.18<0.05; t=-1.44) were not significant. The hypothesis cannot be supported. Preference for the new over the old MOT (diff 0.26 (>0 expected); p=0.00<0.05; t=4.16) was confirmed.
6. *The new MOT is more usable, hence we should base further developments on this version of MOT.*
   For all different parts, as well as overall SUS score (see hypothesis 3), the new version of MOT is preferred over the old version. Thus we should indeed focus further development on the new version. The hypothesis is supported.

In general SUS questionnaires cover more issues than just the main hypotheses. For instance, none of the hypotheses related to learning threshold showed any significant difference between the three systems. This is possibly due to the fact that systems respondents had to learn all the theory before working with the three systems, or that both MOTs are very similar from a theoretical point of view.

We computed the correlation between the SUS scores for the 3 different systems. This showed that the respondents' answers to all three systems' SUS questionnaires are significantly correlated. This seems to be due to one of the following two reasons:

- respondents were not quite aware for which systems they were filling in the SUS questionnaire (suspicion based on some questions from students)
- or the students perceived the three systems as variants or parts of the same system.

Moreover, we also found that the correlation between the scores for the new MOT and for the Sesame2MOT conversion is highest. This could indicate that a substantial number of respondents viewed the Sesame2MOT conversion and the new MOT as one system, since Sesame2MOT is currently integrated into the new MOT.

5.3 Qualitative Feedback

As discussed in section 4, the prototype was also qualitatively evaluated both in a small scale experiment in cooperation with the University of Joensuu as well as in the larger scale experiment in Bucharest. The IMPDET experiment showed that the system was mainly understood, but respondents were unable to provide feedback on the method itself. Some shortcomings of the user interface of the prototype were identified. The qualitative feedback gathered from the Bucharest experiment showed a few issues. First of all, the user interface needs to be improved. Tool tip help functionality, a better interface for weight/label setting that allows changing of individual weights/labels and extra information, like ranking of the article is needed.

6 Conclusions

In this paper we have reviewed an authoring environment for personalized courses, as well as an Enricher mechanism and prototype based on Semantic Desktop technology. The paper briefly sketches the theoretical considerations for the implementation of the Enricher, and then, in parallel, the evaluation of these considerations, as well as of the prototype. From the two evaluation steps performed, the general result is that, to the extent it was understood, the theoretical concept of Adaptive Authoring of Adaptive Hypermedia was perceived as useful. We have also gained some important feedback into possible improvements to the Enricher application itself. Respondents in our experiments pointed out that the integration is currently not optimal and the user interface has to be improved. We plan to integrate the Enricher further into MOT by making a web based version and enhance the usability of the selection options. As especially the qualitative feedback showed the user interface of Sesame2MOT clearly needs to be improved.

Beside the hypotheses analysed here, we are also looked into the correlation between the students' responses and their comprehension of the theory on adaptive hypermedia and authoring thereof; students with higher grades in general responded more positively to the direct questions about liking the systems. For the SUS questionnaire we couldn't find any correlation. We performed comparisons of the students' preferences, however we didn't find any significant results.

Concluding, these tests have shown us that automatic generation and linking of material for adaptive presentation is possible, and that students with only one week of introduction into the whole concept of adaptive hypermedia as well as to the systems implementing it were able to work with our prototypes. For educational researchers, such as in the Joensuu test, this was possible after only one session. These tests however point to the fact that preliminary schooling is necessary for authors to be able to correctly use these concepts and apply personalization to their content presentation.

Acknowledgments. We acknowledge the Socrates mobilities exchange program, the ALS Minerva Socrates Project and Prof. Dr. Paul Cristea for giving us the opportunity to conduct our evaluation experiments.

References

1. Beagle website, viewed 21 March 21, 2007, http://beagle-project.org/Main_Page
2. Beagle++ website, viewed March 21, 2007, http://beagle.kbs.uni-hannover.de/
3. Brailsford, T.J., Ashman, H.L., Stewart, C.D., Zakaria, M.R., Moore, A.: User Control of Adaptation in an Automated Web-Based Learning Environment. In: First International Conference on Information Technology & Applications (ICITA 2002). Bathurst, Australia (2002)
4. Broekstra, J., Kampman, A.: SeRQL: An RDF Query and Transformation, viewed March 21, 2007, http://wwwis.win.tue.nl/ jbroekst/papers/SeRQL.pdf

5. Brooke, J.: SUS: a quick and dirty usability scale. In: Jordan, P.W., Thomas, B., Weerdmeester, B.A., McClelland, A.L. (eds.) Usability Evaluation in Industry, Taylor and Francis, London (1996)
6. Brusilovsky, P.: Adaptive hypermedia, User Modelling and User Adapted Interaction. In: Kobsa, A. (ed.) Ten Year Anniversary vol. 11 (1/2), pp. 87–110 (2001)
7. Celik, I., Stewart, C., Ashman, H.: Interoperability as an Aid to Authoring: Accessing User Models in Multiple AEH Systems. In: Proceedings of A3H: 1st International Workshop on Authoring of Adaptive and Adaptable Hypermedia (2006)
8. Chirita, P.-A., Costache, S., Nejdl, W., Paiu, R.: Beagle++ Semantically Enhanced Searching and Ranking on the Desktop. In: Proceedings of the 3rd European Semantic Web Conference, Budva, Montenegro (2006)
9. Chirita, P.-A., Gavriloaie, R., Ghita, S., Nejdl, W., Paiu, R.: Activity-Based Metadata for Semantic Desktop Search. In: Proceedings of the 2nd European Semantic Web Conference, Heraklion, Crete (2005)
10. Cristea, A.I., De Mooij, A.: Adaptive Course Authoring: My Online Teacher. In: Proceedings of ICT'03, Papeete, French Polynesia (2003)
11. Cristea, A.I., De Mooij, A.: LAOS: Layered WWW AHS Authoring Model and their corresponding Algebraic Operators. In: WWW03 (The Twelfth International World Wide Web Conference), Alternate Track on Education,Budapest,Hungary (2003)
12. Cristea, A.I., Smits, D., De Bra, P.: Writing MOT, Reading AHA! - converting between an authoring and a delivery system for adaptive educational hypermedia. In: A3EH Workshop, AIED'05, The Netherlands, Amsterdam (2005)
13. Cristea, A.I., Stewart, C.: Automatic Authoring of Adaptive Educational Hypermedia. In: Web-Based Intelligent e-Learning Systems: Technologies and Applications, IDEA Publishing group, Zongmin Ma (2005)
14. Hendrix, M., Cristea, A.I.: Evaluating Adaptive authoring of Adaptive Hypermedia. In: A3EH: 5th International Workshop on Authoring of Adaptive & Adaptable Educational Hypermedia, the 11th Internat. Conf. on User Modeling, Corfu Greece (2007)
15. Hendrix, M., Cristea, A.I., Joy, M.S.: Evaluating the automatic and manual creation process of adaptive lessons. In: The 7th IEEE International Conference on Advanced Learning Technologies (ICALT), Niigata, Japan (2007)
16. Hendrix, M., Cristea, A.I., Nejdl, W.: Authoring Adaptive Learning Material on the Semantic Desktop. In: 1st International Workshop on Authoring of Adaptive and Adaptable Hypermedia, Adaptive Hypermedia Dublin Ireland (2006)
17. Hendrix, M., Cristea, A.I., Nejdl, W.: Automatic and Manual Annotation Using Flexible Schemas for Adaptation on the Semantic Desktop. In: Nejdl, W., Tochtermann, K. (eds.) EC-TEL 2006. LNCS, vol. 4227, Springer, Heidelberg (2006)
18. Hendrix, M., Cristea, A.I., Nejdl, W.: Authoring Adaptive Educational Hypermedia on the Semantic Desktop. International Journal of Learning Technologies (IJLT) (to appear, 2007)
19. MOT homepage, viewed March 21, 2007, http://prolearn.dcs.warwick.ac.uk/mot.html
20. Nielsen, J.: Usability Engineering. Academic Press, Boston (1993)
21. Schlieder, T., Naumann, F.: Approximate tree embedding for querying cml data. In: ACM SIGIR Workshop on CML and Information Retrieval
22. Semantic Desktop, viewed March 21, 2007, http://www.semanticdesktop.org/