Slicepedia: Towards Long Tail Resource Production through Open Corpus Reuse

Killian Levacher, Seamus Lawless, Vincent Wade

Trinity College Dublin

Abstract. The production of resources supporting the needs of Adaptive Hypermedia Systems (AHS) is labor-intensive (e.g. in areas such as educational, news media etc...). As a result, content production is focused upon meeting the needs of resources with higher demand, which limits the extent upon which long tail content requirement niches of AHS can be met. Open corpus slicing attempts to convert the wealth of information available on the World Wide Web, into customizable information objects. This approach could provide the basis of an open corpus supply service meeting long tail content requirements of AHS. This paper takes a case study approach, focusing on an educational sector of adaptive hypermedia, to test out the effect of using Slicepedia, a service which enables the discovery, reuse and customization of open corpus resources. An architecture and implementation of the system is presented along with a user-trial evaluation suggesting slicing techniques could represent a valid candidate for long tail content production supply of AHS.

1 Introduction

Adaptive Hypermedia Systems (AHS) have traditionally attempted to respond to the demand for personalized interactive learning experiences through the support of adaptivity, which sequences re-composable pieces of information into personalized presentations for individual users. While their effectiveness and benefits have been proven in numerous studies [1], the ability of AHS to reach the mainstream audience has been limited [2]. For example, in educational hypermedia systems, this has been in part due to their reliance upon large volumes of one-size-fits-all educational resources available at high production costs [3].

Although extensively studied, solutions proposed so far (section 2) do not address the fundamental problem directly which is the labor-intensive manual production of such resources. As a result, content creation is naturally focused upon addressing the needs of targeted resources in higher demand (area 1 in figure 1). AHS content requirements however, naturally follow a long tail distribution. They require large varieties of unique niche content supplies needed for once-off usages only (area 2), which traditional content production approaches desist due to prohibitive costs. In parallel to these developments, the field of Open Adaptive Hypermedia (OAH) has attempted to leverage the wealth of information, which has now become accessible on the WWW as open corpus information. Open Corpus Slicing (OCS) techniques [4] in particular aim at automatically converting native open corpus resources into customizable content objects meeting various specific AHS content requirement needs (topic covered, style, granularity, delivery format, annotations).

We believe that, in order to serve AHS long tail content requirements, the cost intensive, manual production and/or adaptation of educational resources

must be augmented (and if possible replaced) by the automated re-purposing of open corpus content into such resources. A fully-automated, on-demand, content production system based upon an OCS approach could thus theoretically address the long tail content supply paradigm described previously. This paper builds upon previous research in slicing techniques and attempts to provide a first step towards this goal. Educational Hypermedia Systems (EHS) are cited as being the most successful but also most expensive systems to develop content for [5], this paper hence takes a case study approach investigating how educational AHS in particular can be supported by an automated content production service based upon OCS techniques.

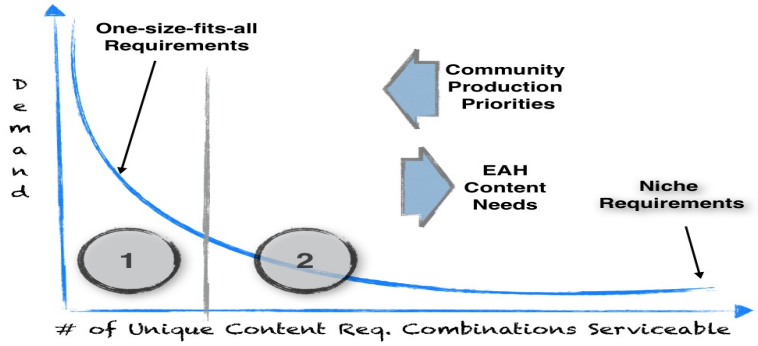


Fig. 1. The Long Tail of EAH Content Supply

Contribution: The rest of this paper presents Slicepedia, a content supply service, based upon slicing techniques, that leverages open corpus resources [1] to produce large volumes of right-fitted information objects. An i) architecture and implementation of the system is presented, followed by ii) a user-trial evaluation, applied in an authentic educational scenario, comparing the suitability of such a content supply service with respect to a manual approach.

2 Background and Related Work

The reliance of EHS upon the ability to access large volumes of diverse educational resources has been a problem investigated by many researchers. The reuse and sharing of existing resources, through encapsulation standards such as Learning Objects Metadata(LOM)[2], followed a natural need by the community to reuse previously produced resources, which subsequently led to the creation of many learning object repositories (LOR)[3]. As these repositories grew in size, research focusing upon improving the search and discovery of existing resources [6] naturally emerged as consequence. Tools improving production efficiency [7], by re-purposing existing material already available in LORs into new learning resources, ultimately emerged as both the size and search capabilities of LOR improved. Although these solutions certainly do reduce the production time and costs of learning resources, none of them directly address the fundamental issue

[1] In order to deal with the significant challenges of right sizing and reuse, copyright and digital rights management issues are deemed beyond the scope of this paper
[2] www.ieeeltsc.org
[3] NDLR: www.ndlr.ie, MERLOT: www.merlot.org

which is the initial labor-intensive production of such content. When production efficiency is taken into account, improvements through re-purposing are still carried out manually [5]. This results in a costly content production paradigm with limitations delimited in terms of volume of resources available.

OAH attempts to surpass these volume supply limitations through the incorporation of open corpus resources through either manual incorporation [8] techniques, automated [9] and community-based [10] linkage or Information Retrieval (IR) approaches [11]. In contrast with previous solutions, this paradigm offers a much cheaper alternative supply of content, with limitations, in terms of volume of resources available, significantly lower. However even when relevant open web resources are retrieved, IR techniques suffer because they only provide untailored, document level, delivery of results, with limited control over topics, granularity, content format or associated meta-data. Open corpus resources, originally produced for a pre-define purpose, are generally incorporated in their native form, as "one-size-fits-all" documents. This represents a very inadequate match for long tail content niche requirements. As pointed out by Lawless [11], the reuse potential of such resources (i.e: a news article for instance), complete with original menus, advertisements and user comments, are far less adequate than the reuse of selected parts of the article, de-contextualised from their original setting, at various levels of granularity (from a paragraph on a specific topic to an entire article), with associated meta-data and in a delivery format of choice.

Open EHS Supply Requirements: In order to clarify what we mean by a long tail content supply scenario, lets consider the following open and user-driven EHS use case scenario. Suppose Alice wishes to improve her grammar skills in Portuguese and decides to use a portal specifically built for this purpose. The system provides e-assessments consisting of traditional gap filling exercises for a given piece of native language text (figure 3a). It provides Alice with a list of various languages Λ, grammar skills Γ and reading level difficulty ϱ to choose from which she selects accordingly to her training needs. So as to sustain learner motivation, the portal additionally provides the ability to select topics of interest, among a large list Θ, which training exercises should also cover. Whenever Alice starts her training, the system searches for resources on the web fulfilling the combined requirements $\Sigma\{\Lambda, \Gamma, \Theta, \varrho\}$ and converts these into grammar e-assessments. The system continuously records the sets of mistakes μ performed by Alice and includes this additional variable to its content requirement combination Σ in order to address the required subset of grammar points of immediate importance. As Alice progresses, the set of requirements Σ evolves and so does the content supplied. The portal can supply Alice with as much content as needed for her training.

As pointed out by Steichen et al. [12], the production of educational resources a-priori of any learner interaction, generally assumes that the type and quantity of resources needed for a particular EHS is known in advance of system deployment. In the use case presented above however, the number of content requests is unknown in advance and the content requirement combination possibilities for Σ are very numerous. For this reason, only a handful of deliveries will ever occur for most individual content requests possibilities. This situation portrays a typical long tail distribution (figure 1) scenario which could only be sustained by an

automated content production service guaranteeing the on-demand provision of i) large volumes of content, ii) at low production costs, iii) suitable for a large range of potential activities [5].

OCS techniques [4, 13] aim at automatically right-fitting open corpus resources to various content requirements. Open corpus material in its native form is very heterogeneous. It comes in various formats, languages, is generally very coarse-grained and contains unnecessary noise such as navigation bars, advertisements etc. OCS provides the ability to harvest, fragment, annotate and combine relevant open corpus fragments and meta-data in real time. Hence, although the quality of content supplied (with respect to relevance) is important of course, this aspect is addressed by traditional retrieval systems. OCS techniques instead aim at improving the quality, in terms of appropriateness for further AHS consumption, of open corpus resources identified as relevant. It is still unclear, however how the suitability of content generated automatically by such techniques would compare with content manually hand crafted within an educational use case scenario. If such a technique is to be used as a basis for long tail content supply chain services, the suitability of the content produced (requirement iii) from a user's point of view as well as it's production cost (requirement ii) must be examined. The authors are fully aware that, full automation of educational resource production might not be an adequate solution for the entire set of possible EHS use cases. However, a selected subset of those could clearly benefit from automation or even an increase in production efficiency via semi-automated alternatives. An experiment in progress, investigating whether such a technique can be applied within a high school science text book curriculum use case is currently in progress. The following sections describe the architecture and implementation of such a system called Slicepedia.

3 Slicepedia Anatomy

As depicted in figure 2, a slicer is designed as a pipeline of successive modules, analysing and appending specific layers of meta-data to each document. Large volumes of resources openly available on the WWW, in multiple languages and with unknown page structure, are gathered and then transformed, on demand, into reusable content objects called slices. EHS thus use the slicer as a pluggable content provider service, producing slices which match specific unique content requirements (topic, granularity, annotations, format etc.). The aim of this section is to present the overall architecture and detailed implementation of the slicer used within this experiment. For a more complete description of an OCS pipeline architecture, the reader is referred to the original papers [4, 13].

Harvester: The first component of a slicer pipeline acquires and locally caches open corpus resources from the web in their native form. Although standard IR or focused crawling techniques [11] would generally be selected for this phase, this experiment required a tighter control over resources automatically harvested (to allow direct comparisons with manual repurposing approaches see section 4). Thus a simple URL list-harvesting feature was used instead for this component.

Fragmentation: Once open corpus resources are acquired, each individual document is fragmented into structurally coherent atomic pieces (such as menus, advertisements, main article). The Kohlschutter et. al [14] densitometric approach

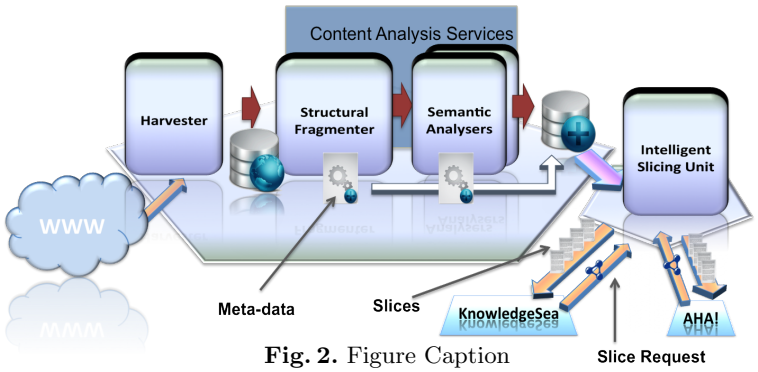


Fig. 2. Figure Caption

to fragmentation was selected for the purpose of this experiment as it can process virtually any xml-based document at very high speed.

Semantic Analyser: Each fragment is then analysed by a set of annotators. The intention is to produce sufficient discerning meta-data to support the identification and selection of adequate meta-data/fragments combinations matching each EHS content request. Such meta-data might include, writing style, topic covered or the level requirement of content. The set of annotators used for the purpose of this experiment consisted of: i) the AlchemyApi concept tagging service[4], which identifies and associates concepts mentioned within each fragment with Dbpedia instances[5], ii) the open source Flesh annotator[6] determining reading-level difficulties of resources as Flesh Reading scores. iii) Part of speech[7], iv) noun and verb phrases[8], were also identified within fragments and annotated with their relevant linguistic attributes. Finally, fragments were annotated using v) a boilerplate detection algorithm[9], determining to what degree individual page parts are reusable or not. All annotations and fragments were stored as rdf data within a Sesame triple store[10] and available as linked-data.

Slice Creation: Once individual fragments and meta-data annotations are available, a slicer is ready to receive slice requests. For each request, a slicing unit combines atomic fragments together, along with relevant meta-data, into customized slices. A slice is defined as: *Customized content generated on-demand, consisting of fragment(s) (originating from pre-existing document(s)) assembled, combined with appropriate meta-data and right-fitted to the specific content requirements of a slice consumer (with various application-specific content-reuse intentions).* Within this implementation, slice requests were converted into SPARQL queries and submitted to the triple store in order to identify any matching fragment/annotation combinations. Fragments identified were then appended to each other and annotations inserted in the resulting compounded fragment. The array of possible adjustments (such as the extent of control over granularity, formats

[4] http://www.alchemyapi.com/api/
[5] http://dbpedia.org/About/
[6] http://flesh.sourceforge.net
[7] Modified version of the Brill Tagger in ANNIE http://gate.ac.uk/
[8] Verb group and noun phrase chunkier in http://gate.ac.uk/
[9] http://code.google.com/p/boilerpipe/
[10] www.openrdf.org

and annotation) a slicer can offer upon an open corpus resource, is referred to as its Content Adaptation Spectrum (CAS). The CAS provided by this slicer therefore consists of 3 right-fitting dimensions (Content Style, Granularity and Annotation Type) including 10 adaptation variables (content style, topics covered, reading difficulty, verb chunk tense, number of annotations, paragraph/word number, topic focus, annotation focus, original sources, delivery format) that can be arbitrarily combined to suit various content requirements over any relevant open corpus resources identified. A slice request could hence consist of the following: *slices which originate from only a specified list of websites and have a granularity ranging from 3 sentences up to 3 paragraphs. These slices should cover the topics of whale migration, atlantic ocean or hunting, should not contain any tables or bullet point lists, should have a Flesh reading score ranging from 45 to 80, and contain between 7 and 15 annotations consisting of verbs conjugated at the past perfect continuous.*

4 Evaluation & Results

Aim and Hypothesis: As discussed in section 2, in order to supply long tail niche requirements, a content production system service should guarantee the provision of i) large volumes of content, iii) at low production costs, iii) suitable for arbitrary activities performed. Although the first condition is necessarily achieved through the selection of an open corpus reuse strategy, the performance of a slicer with respect to the two other conditions is yet to be examined. For this reason, this evaluation focuses, as a first step, upon investigating the suitability and cost of content produced automatically by a slicing system with respect to content manually produced for the same activity. This initial experiment is performed upon a sample of independently selected niche requirements and open corpus resources. Assuming positive results were measured, this would indicate such a technique could also scale for any niche requirement using any open corpus content. Any issues detected at this early stage would only be amplified within a large scale deployment making any scaling pointless. Additionally, since production cost is dependent upon local prices of manual labor as well as individual server specifications, an estimation of production time was considered instead a better proxy for production cost differences. For this reason, the time required to produce content batches was measured for both the educators and slicer. The hypotheses tested within this evaluation are therefore as follows:

- H1: Content manually & automatically produced achieves similar suitability results, from a users point of view.
- H2: Automated slicing offers a reasonable production cost solution for large volumes of content, in contrast with a manual production which is unsustainable

Evaluation Design: Since content consumed by EHS is ultimately presented to people, any content production measurement should consider user experience as critical. Furthermore, as the aim consists in evaluating content produced by two different methods, slices should be assessed individually, with interface complexity (figure 3a) and re-composition kept to a minimum in order to avoid any possible interference with the content being evaluated. For this reason, a simplified version of the language e-assessment use-case application presented in section 2 was built

specifically for the purpose of this experiment. Within the context of this paper, the purpose of this educational application is to be used only as a "content reuse vehicle" for evaluating the slicer (i.e. not discussing educational aspects).This application represents a well known approach to grammatical teaching, via a Computer Adaptive Testing (CAT) tool, similar in nature to the item response assessment system introduced by Lee et al. [15]. For this reason, it provides a good example of the kind of educational application which could avail of 3rd party content. In this application, users are presented with series of individual open corpus resources (involving no re-composition), repurposed (manually or automatically) as traditional gap filler exercises. Verb chunks items are removed and replaced by gaps, which users must fill according to particular infinitives and tenses specified for each gap. Answers provided are then compared to the original verb chunks and users are assigned a score for each specific grammar point. Slicing open corpus content is known to affect the reading flow and annotation quality of content produced [4]. An evaluation of content performed within the context of a language learning task (by individuals with varying levels of competency in the English language) should hence make participants very sensitive to reading flow properties of content. Moreover, since grammar e-assessment units are created according to verb chunk annotations, any annotation quality issue would provoke major assessment errors, making any training over such content pointless and time consuming to users. Suitability performance, within the context of this use case, hence refers to the ability to correctly assess an individual.

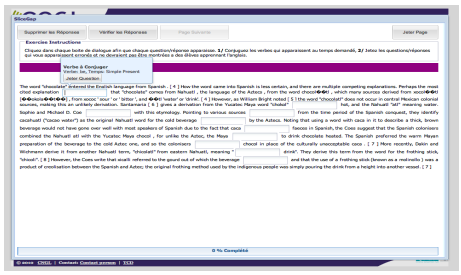
Content Batch Creation: The manual content production activity selected for this experiment deliberately involved only the repurposing of existing content (in the spirit of educator repurposing activities described in section). This decision aspired to replicate a manual content production scenario (used as a baseline) with minimal production time requirements. Our assumption was that any content authoring activities would always depict higher time requirements. Hence, in order to select a truly random set of open corpus pages, a group of five independent English teachers, were asked to arbitrarily select a combined total of 45 pages of their choice from the web. The pages could be selected from any source, according to various combinations of requirements (topics covered, tenses...). They were then asked to select fragments of pages harvested, which they felt were adequate for grammar exercises, and manually annotate tenses encountered within these extract to produce content batch CBM.

Fragments could consist of any granularity as long as content, which was not about a specified topic, was discarded. The collection of arbitrarily selected pages, was then harvested from the web in their original form by the slicer. CAS characteristics (including topics) of resources manually produced, were identified and fed into the slicer as independent niche content requirement parameters. The entire set of open corpus content harvested was then sliced with these parameters to produce content batch CBA. Content produced in both batches were subsequently converted into grammar e-assessment pages.

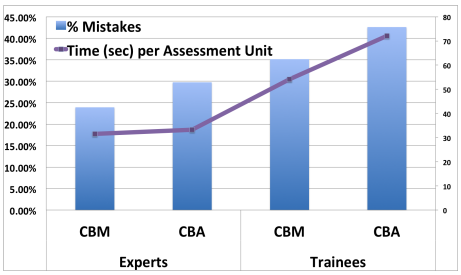
Evaluation Scenario: The entire experiment was available online to the public with the interface and questionnaire available in English, Spanish and French. Native and non-native speakers were invited to perform a set of English grammar training exercises using resources randomly selected from each content

batch using a latin square design distribution. A unique color was assigned to each content batch and users were unaware which was being presented to them at each task. Users were asked to fill in any blanks encountered (10 gaps on average per page) with the appropriate verb and tense specified for each case (Figure 3a). Following these exercises, they were subsequently asked questions directly, answered using a 10 point Likert scale. Finally, they were asked to order colors, corresponding to each content batch presented, based on their perceived quality.

Results: The rest of this section presents a summary[11] of the findings observed throughout this experiment in relation to each hypothesis. A total of 41 users, divided into two groups (Experts (63%) and Trainees (37%)), performed the experiment, most using the English interface (en=66%, non-en=34%).



(a) Slicegap Screenshot



(b) User Performance Results

Fig. 3. Content Supply Evaluation

H1: Statistical t-test analysis, of both the number of mistakes performed by users as well as the time required to perform e-assessments, revealed that any differences measured were statistically insignificant ($p > 0.100$). When trainees were asked whether, for content batch CBA, *"the number of erroneous assessment units presented was tolerable"*, a mean score of 7 out of ten was measured. When asked whether *"Overall, I felt this content was adequate to perform a grammar training exercises"* both content achieved very similar scores (CBM=8.33, CBA=8.57, p=0.536) with t-tests again suggesting any difference observed was insignificant. However, when plotting this data on a graph(Figure 3b.), a pattern can be observed for both the trainee as well as expert group. In both cases, the number of mistakes and time taken to perform e-assessments, upon content created automatically, appears to be higher than for the content produced manually. This would suggest users do make more errors when using content from the CBA batch. Although automated verb chunk annotation recall accuracies measured outperformed those produced manually (Recall A=0.92, M=0.68), manual annotations precision accuracies were slightly higher than automated ones (Precision M=0.89, A=0.86), which could explain this minor increase in errors observed. Finally, although the difference in errors between content batches was slightly higher for trainees in comparison to experts group, an independent t-test[12] indicated this difference was insignificant (mean dif. E=5.80%, T=7.48%, p=0.891). This indicates that users from the expert group didn't appear to use their language skills to compensate content suitability differences between both batches.

[11] Results on 14 variables at: https://www.scss.tcd.ie/∼levachk/tcdwebsite/data.html'
[12] Equal variances assumed

Overall, these results suggest that although a pattern of slightly lower performances on assessments automatically generated was observed, this difference was insignificant and didn't appear to affect trainees more than the experts group of users, nor did it appear to decrease the perceived usefulness of the content for the assessment task performed.

H2: Since the set of pages used as preliminary resources was purposely manually harvested from the web (section 4), no automated harvesting was performed by the slicer during this experiment. Nevertheless, in order to provide a fair comparison with a manual production approach, an estimation of time required by the slicer to perform this task was necessary. Teachers were asked to only harvest open corpus resources matching specific criteria combinations (such as topics covered, tenses etc...). Hence, an automated harvesting time estimation based upon traditional IR services would have created an unfair advantage towards the slicer, since these techniques only provide keyword searches with little guarantee that harvested resources satisfy these criteria. For these reasons, the OCCS focused crawler [11] was considered a fairer option since it provides the means to specify content to be harvested based upon a wider range of constraints (including topics covered) and also guarantees resources harvested, if any, meet these sets of constraints. Time measurements obtained for this experiment reveal that when no particular content requirement was specified, teachers took an average of 3.75 minutes to harvest open corpus resources and extract arbitrary fragments suitable for grammar exercises. Requesting resources to be on a specific topic, only slightly increased the average time measured (4 min) whereas requesting resources to possess verbs conjugated at particular tenses nearly tripled the time needed (10.5 min). These results follow common sense, since the ability of humans to identify topics covered within resources is much more straight forward than for machines, however the reverse is also true when dealing with specific fine grained requirements such as verb tenses. Teachers on average took 4.25 minutes to annotate fragments (189 words in average, 14 annotations per fragments) leading to a total time ranging from 8 min to 14.75 min to produce these resources. This would be the equivalent of between 1.5 to nearly 3 years of manual labor necessary to produce a hundred thousand of such resources. According to Lawless et al. [11], a time performance of 149,993 valid resources harvested in 43h was measured for the OCCS system (without any cpu parallelization). This is equivalent to $17.2 * 10^{-3}$ minutes of harvesting time necessary per page. Summing extraction, annotation and slice creation time performed on a 2.8GHz machine [13] leads to a total of $5.4 * 10^{-1}$ minutes necessary to produce each page. Assuming no parallelization was used during the slicing process (section 3), this already represents a difference of up to 96% production time increase with respect to it's manual production equivalent. Although automated and manual production time are clearly not directly comparable, one can assume in most cases, server costs per time unit to be much lower than labor costs. Considering the low server production time measured in comparison to the manual tasks, automated content production cost can be inferred to be also much lower than a manual approach and hence more adequate for large number of resources produced.

[13] CPU: 2.8Ghz Intel Core 2 Duo, Mem: 4GB 1067 MHz DDR3

5 Conclusion

Although differences were observed between content automatically produced and manually hand crafted, results presented in this paper indicate that any differences were statistically insignificant. However, when taking into account content batches production costs, automatically generated resources significantly outweighed those manually produced. Hence, in the context of a high speed, low cost production environment, one could easily assume any content produced with unsatisfactory suitability to be discarded and rapidly replaced, which could compensate any decrease in quality. The ability of automated open corpus slicing techniques to produce large volumes of content on-demand, at very low costs and with a suitability comparable to manually produced resources, would thus appear to represent a promising candidate approach to consider for long tail content supply services. As this initial experiment only took into account specific aspects of content quality (i.e. reading flow, annotations...) within a chosen educational content reuse scenario, further research should investigate this approach within various use cases. A user trial incorporating this approach within a high school science text-book content supply use case is currently underway with a major publisher in the USA.

References

1. Lin, Y.l., Brusilovsky, P.: Towards Open Corpus Adaptive Hypermedia : A Study of Novelty Detection Approaches. In: UMAP'11: Proc. of the 19th Conf. on User Modeling, Adaptation, and Personalization. (2011) 353–358
2. Armani, J.: A Visual Authoring Tool for Adaptive Websites Tailored to Non-Programmer Teachers Jacopo Armani. Educational Technology & Society **8** (2005)
3. Jednoralski, D., Melis, E., Sosnovsky, S., Ullrich, C.: Gap Detection in Web-Based Adaptive. In: ICWL'10: Proc. of the 9th int. conf. on Web-based Learning. (2010)
4. Levacher, K., Wade, V.: Providing Customized Reuse of Open-Web Resources for Adaptive Hypermedia. In: 23rd Conf. on Hypertext and Social Media. (2012)
5. Meyer, M., Rensing, C., Steinmetz, R.: Multigranularity reuse of learning resources. Transactions on Multimedia Computing, Communications, and Applications (2011)
6. Diaz-aviles, Nejdl, W.: Unsupervised Auto-tagging for Learning Object Enrichment. In: Proc. of the 6th int. conf. Euro. Conf. on Tech. Enhanced Learning. (2011)
7. Madjarov, I., Boucelma, O.: Learning Content Adaptation for m-Learning Systems : A Multimodality Approach. In: ICWL'10. (2010) 190–199
8. Henze, N., Nejdl, W.: Adaptation in Open Corpus Hypermedia. IJAIED'01: Int. Journal of Artificial Intellligence in Education (2001) 325 – 350
9. Zhou, D., Goulding, J.: Automatic Hypertext Generation Utilizing Language Models. In: Proc. of the 18th int. conf. on Hypertext and hypermedia. (2007)
10. Brusilovsky, P., Farzan, R.: Social adaptive navigation support for open corpus electronic textbooks. In: Adaptive Hypermedia and Web Based Systems. (2004)
11. Lawless, S.: Leveraging Content from Open Corpus Sources for Technology Enhanced Learning. PhD thesis, Trinity College Dublin (2009)
12. Steichen, B., Wade, V.: Providing Personalisation across Semantic, Social and Open-Web Resources. In: conf. on Hypertext and Hypermedia. (2011)
13. Levacher, K., Wade, V.: A Framework for Content Preparation to Support Open-Corpus Adaptive Hypermedia. In: 20th Conf. on Hypertext and Hypermedia. (2009)
14. Kohlschütter, C., Nejdl, W.: A Densitometric Approach to Web Page Segmentation. In: int. conf. on Information and knowledge management. (2008)
15. Lee, Y., Choi, B.u.: A Personalized Assessment System Based on Item Response Theory. In: ICWL'10 int. conf. on Web-based Learning. (2010)