Slicepedia: Providing Customized Reuse of Open-Web Resources for Adaptive Hypermedia

Killian Levacher, Séamus Lawless, Vincent Wade
Centre for Next Generation Localisation, Knowledge and Data Engineering Group
School of Computer Science and Statistics, Trinity College Dublin, Ireland
{Killian.Levacher, Seamus.Lawless, Vincent.Wade}@scss.tcd.ie

ABSTRACT

A key advantage of Adaptive Hypermedia Systems (AHS) is their ability to re-sequence and reintegrate content to satisfy particular user needs. However, this can require large volumes of content, with appropriate granularities and suitable meta-data descriptions. This represents a major impediment to the mainstream adoption of Adaptive Hypermedia. Open Adaptive Hypermedia systems have addressed this challenge by leveraging open corpus content available on the World Wide Web. However, the full reuse potential of such content is yet to be leveraged. Open corpus content is today still mainly available as only one-size-fits-all document-level information objects. Automatically customizing and right-fitting open corpus content with the aim of improving its amenability to reuse would enable AHS to more effectively utilise these resources.

This paper presents a novel architecture and service called Slicepedia, which processes open corpus resources for reuse within AHS. The aim of this service is to improve the reuse of open corpus content by right-fitting it to the specific content requirements of individual systems. Complementary techniques from Information Retrieval, Content Fragmentation, Information Extraction and Semantic Web are leveraged to convert the original resources into information objects called slices. The service has been applied in an authentic language elearning scenario to validate the quality of the slicing and reuse. A user trial, involving language learners, was also conducted. The evidence clearly shows that the reuse of open corpus content in AHS is improved by this approach, with minimal decrease in the quality of the original content harvested.

Categories and Subject Descriptors

H3.3 **[Information Search and Retrieval]**: Information Filtering; Retrieval Models; Selection Process;
H.5.4 **[Hypertext/Hypermedia]**: Architectures; User Issues;

Keywords

Customized Hypertext Content Generation, Open Corpus Content Processing, User Experience

1. INTRODUCTION

Adaptive Hypermedia Systems have traditionally attempted to deliver dynamically adapted and personalised presentations to users through the sequencing of pieces of information. While the effectiveness of such systems and the benefits of their use have been proven in numerous studies [9], the adaptivity offered by AHS, and thus their widespread adoption, have been restricted by the lack of sufficient content in terms of volume, granularity, style and meta-data. Such systems have traditionally relied upon the manual production [3] of bespoke proprietary content [1], resulting in potentially low volumes and high production costs. Open Adaptive Hypermedia (OAH) research addresses this challenge by leveraging open corpus information available on the Worldwide Web (WWW) and utilising it in AH presentations.

However, when open corpus reuse has been achieved, it is generally performed manually [5] or at best using traditional information retrieval (IR) approaches [1]. The usage of IR approaches has met with very limited success, even when retrieving relevant open web information. These techniques suffer because they only provide one-size-fits-all, untailored delivery of results, with limited control over granularity, content format or associated meta-data. This results in limited and restricted reuse of such resources in AHS. 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. Hence, there remains a significant barrier to automatically convert native open corpus content into right-fitted information objects meeting the specific content requirements (such as granularity, delivery format, annotations) of individual AHS.

Contribution: In this paper, a novel approach to open corpus reuse through right-fitting is proposed. This novel approach leverages complementary techniques from IR [11], Content Fragmentation [6], Information Extraction (IE) [10] and Semantic Web [4] to improve the reuse of open corpus resources by converting them into information objects called slices. By analysing, fragmenting and re-assembling previously published documents into customized objects, native open corpus resources can be right-fitted to meet a diverse range of content requirements from individual adaptive systems. An implementation of the system architecture, called Slicepedia, is presented, which has been applied in an authentic educational scenario. An evaluation (detailed in section 4) based upon a user-trial was conducted to validate the reuse and appropriateness of the right-sizing. Results provide evidence that this approach improves the reuse of open resources while minimizing any decrease in quality of the original open corpus content harvested. In order to deal with the significant technical challenges of right sizing and reuse, some specific aspects are deemed beyond the scope of this paper; namely copyright and digital rights management issues.

2. BACKGROUND

OAH research has attempted to resolve the closed corpora dependency of traditional AHS by leveraging the wealth of information available over the WWW. These techniques usually consist of incorporating open corpus resources using either i)

Conference'23, June 25-28, 2012, Milwaukee, WI, USA.

manual, ii) community-based, iii) automated linkage or iv) IR approaches. Manual incorporation techniques [5] allow users/designers to incorporate documents within an adaptive experience. However these techniques require a significant amount of time and effort due to the difficulty in identifying adequate content and annotating these external resources prior to incorporation within individual systems. Automated linkage [13] and community-based [2] approaches attempt to improve this situation by providing guidance with respect to the relevant content that is available. While the former automatically estimates the semantic relatedness between pages, community-based approaches analyse the quantity of users stepping between various resources in order to derive this information. IR approaches [11], on the other hand, aim to provide pluggable services for AHS by offering various search capabilities to support open corpus content identification and incorporation. The OCCS system [7] for instance, uses focused crawling techniques to harvest large amounts of web resources and identify those most relevant to specific contexts of use, based on arbitrarily pre-selected topic boundaries. All of these approaches focus on dealing with the information overload encountered while leveraging such large quantities of resources; either by improving the identification of suitable content or by connecting resources together. However resources identified by such techniques are still used in their native form, as one-size-fits-all documents, in their original format of delivery. As pointed out by Lawless [7], *"there is an inverse relationship between the potential reusability of content and its granularity"*. The reuse potential of a previously published news page, complete with original menus, advertisements and user comments, is far less than if the article could be reused alone, de-contextualised from its original setting, at various levels of granularity (from a paragraph on a specific topic to the entire article), with associated meta-data and in a delivery format of choice.

Although the field of Personalised Information Retrieval (PIR) has indeed focused on the issue of personalising information delivery, the techniques used by such systems [11] have typically focused on re-ranking the order of resources delivered as opposed to personalising the content itself. Furthermore, the use of such content delivery systems is mostly confined to human "consumption" (which by nature can easily distinguish areas of interest within documents), displayed as entire documents within traditional web browsers. Re-composable information objects, on the other hand, must possess a granularity, scope and discerning meta-data/annotations, specific to each use case, in a format chosen by each individual AHS. If the full reuse potential of open corpus content is to be leveraged, the reuse through right-fitting of these resources represents a necessary preparation step for AHS. This necessity emerges directly from the diversity of formats, page layouts, domains, styles and multilingual nature of open corpus content. A good example of what could be achieved, if open corpus resources were fully available as right-fitted content objects, is the Personal Multilingual Customer Care (PMCC) system developed by Steichen et al. [12]. This system leverages both corporate and user generated resources available in the wild by integrating and re-composing these resources into single coherent presentations to users. Fragments of interest within corporate content and targeted forums are extracted using content specific rule-based algorithms, assigned system specific meta-data and stored in a proprietary format. An ontology is derived from the structure of the corporate content, linked with the forum resources and used as a domain model.

Although this system does reuse these open content resources, it does so by applying manually crafted, content specific rule-based algorithms on targeted resources with a predefined structure. Furthermore, the re-composable content objects which are generated by this system are single purpose, with use case specific format and meta-data as well as pre-defined granularities. If the entire range of open resources available on the WWW is to be fully leveraged and delivered to a diverse range of AHS, a system converting open corpus resources into reusable intelligent content objects, must deal with: i) unknown page structure; ii) multiple languages; and iii) domains. It should also serve a variety of: iv) content requirements (meta-data, granularities); v) delivery formats; and vi) use cases in which content is consumed by an AHS. Additionally, as pointed out by Steichen et al [12], *"OAH systems typically focus on producing educational […] compositions on predefined needs, rather than […] informal user need[s] indicated by a query"*. Manually crafted, rule-based content analysis algorithms, for targeted resources, clearly do not scale if the entire open web is to be targeted as a potential resource. Hence the need for: vii) a fully automated solution to open corpus reuse.

As the range of domains available on the open web is unbounded, a content specific ontology provides an inadequate bounded domain model for the purpose of open content object re-composition. Even within a pre-defined domain, various AHS might interpret specific concepts differently. The rise of the semantic web and in particular linked data[1] now provides a wealth of open, shared and interconnected conceptual models, which could provide open content models the ability to anchor fundamental meta-data descriptions of their underlying resources upon linked data concepts within multiple repositories. This approach would also fully disentangle open content models from any particular domain model as well as providing AHS with various domain ontologies to interpret these resources. Semantic searches [4] could then make use of these anchors to manipulate and identify reusable content objects. Such semantic reasoning capabilities could be used for the adaptive selection of open content based on a chosen domain model, unknown by the open content model.

The opportunity in OAH hence lies in the ability to improve the reuse of existing open corpus resources by converting them, into intelligent content objects created on-demand for individual AHS. By focusing on providing a fully automated, independent and pluggable content delivery service, such an approach would enable OAH to fully leverage open resources by providing more control over the granularity, meta-data and delivery of these native resources to individual AHS. The following sections describe the architecture and implementation of such a system called Slicepedia, with an initial evaluation focusing on assessing the reuse improvements of content delivered over native open corpus content.

3. THE WEB DELIVERED AS SLICES

3.1 Slicepedia Architecture

As depicted in Figure 1, a slicer is designed as a pipeline of successive modules, each analysing open corpus resources and appending specific layers of meta-data to each document. Resources openly available on the WWW are gathered and then

[1] www.linkeddata.org

transformed, on demand, into reusable content objects called slices. AHS (or so-called slice consumers) thus use the slicer as a pluggable content provider service, producing slices matching specific unique content requirements (topic, granularity, annotations, format etc.). Although the architecture presented in this paper could of course be implemented over closed corpus resources, the ultimate aim is to provide a pluggable, fully automated service, which can process large volumes of open corpus resources, without any prior knowledge of structure, domain, or language used by each document and with the ability to support diverse AHS content requirements.

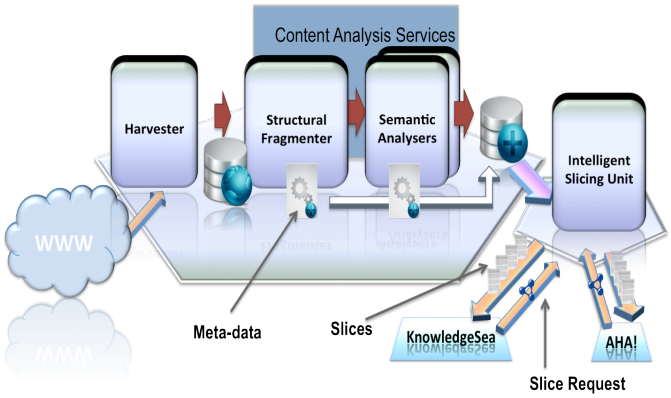

Figure 1 Slicepedia Architecture

Harvesting: The first component of a slicer pipeline aims to acquire open corpus resources, from the web, in their native form. Standard IR systems or focused crawling techniques [7] are used to gather relevant documents, which are then cached locally for further analysis.

Structural Fragmentation: Once resources have been identified, each individual document is analysed and fragmented into structurally coherent atomic pieces (such as menus, advertisements, main article). Structural meta-data, such as the location of each fragment within the original resource, is extracted and stored in the meta-data repository. This phase is critical since, as mentioned previously (section 2), maximising the reuse potential of a resource involves the ability to reuse selected parts of documents, which in turn, depends upon correctly identifying individual sections of pages to produce de-contextualised structurally coherent fragments. Any structural inconsistencies (such as parts of original menus erroneously broken in two and merged with paragraphs or various titles merged together) produced at this stage, will have a direct impact upon the quality of reuse, regardless of the performance of subsequent steps. As mentioned in the previous section, any algorithm used within this component must be fully automated and deal with unknown document structures in multiple languages.

Semantic Analyser: Once an initial set of fragments is produced, each is analysed by standard IE and Natural Language Processing (NLP) algorithms (such as entity extraction [10] and topic detection) with the intention of producing discerning meta-data, supporting the identification and selection of such content for reuse by individual slice consumers. Such meta-data might include, writing style, topic covered or the difficulty level of content. Since what constitutes "discerning" meta-data is highly dependant upon the reuse intentions of each slice consumer, an inappropriate provision of such data can lead to very low reuse scenarios across consumers. Hence great care must be taken during the selection of suitable annotators in order to support the broadest needs possible of targeted consumer use cases. Additionally, this phase is also essential with respect to the level of granularity control provided to slice consumers. The ability to "focus" (or constrain) a resource to only cover a chosen topic, is clearly dependent upon the capacity to accurately match selected parts of single resources with appropriate concepts. Moreover, as the targeted document space of a slicer is by definition open, the resulting open content model available to slice consumers shouldn't be restrained by any predefined subjective domain model. For this reason, Slicepedia disentangles domain modelling from its open content model and instead provides anchors to linked-data concepts as a foundation for any domain model chosen by individual slice consumers. Additionally, all fragments and annotations produced by Slicepedia are also available as linked-data. The intention is to reduce low reuse effects related to the subjective nature of discerning meta-data, and support collaboration with reusable annotations across institutions.

Slice Creation: Once individual fragments and meta-data annotations are available, the slicer is ready to combine these individual elements into slices. The array of possible adjustments (such as the extent of control over granularity, formats and annotation) a slicer can offer upon an open corpus resource, is referred to as its Content Adaptation Spectrum (CAS). Whenever slice requests are received, an intelligent slicing unit combines atomic fragments together, along with relevant meta-data, into customized slices. Since a slicer is to be used as a pluggable service within a variety of slice consumers, a range of delivery formats should be supported. A slice is therefore 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)"*. Slices can contain other slices and can be reused or re-composed with many slices.

3.2 Implementation

The Slicepedia architecture presented in this paper was implemented for the purpose of the experiment described in section 4, in such a way that allows alternative implementations of individual components to be substituted and exchanged with each other to support various measurements. Hence, two alternative implementations of this architecture were implemented. The first implementation supports closed corpus slicing, while the second targets open corpus resources. Both implementations share most components but differ mainly with respect to the fragmentation algorithm selected.

Although IR based harvesting modules (such as the OCCS [7] and Yahoo web search[2]) could easily be plugged into both versions of the slicer, these IR harvesting modules were ignored within the context of this experiment and a simple URL list-harvesting feature was used instead as it was necessary to isolate the components which aim at improving the reuse of these resources as opposed to their identification (see section 4.4). This offered tighter control over the resources supplied to both slicers simultaneously and allowed for a direct comparison of both slicers over a common subset of resources (see section 4.2). With respect to the fragmentation component, the closed corpus slicer used a rule-based approach to fragmentation (section 4.2). For the

[2] http://developer.yahoo.com/search/web/V1/webSearch.html

purpose of this experiment, Wikipedia pages were downloaded in xml format, converted and stored in a local mysql database. Rules built within the JWPL library[3], developed by Darmstadt University, were subsequently used to ignore any clutter within the pages and fragment each one at various granularities (paragraphs, sections etc…) with ideal accuracy. A densitometric approach [6] to fragmentation was selected for the open corpus version of the slicer. It's ability to fragment pages regardless of the meaning or structure of xml tags used and without the need for any rendering, allows it to process virtually any xml-based documents at very high speed. A prior detailed analysis of this algorithm also revealed that it could fragment parallel corpora in multiple languages with a predictable degree of accuracy. However, limitations of this fragmentation algorithm with respect to content type (such as forum or product pages) were also discovered during this analysis. For this reason, this experiment considered a subset of the WWW as an open corpus target, consisting of any news or encyclopaedia type pages. Although, this clearly represents a limitation with respect to the aim of improving the automated reuse of any open corpus resource available on the WWW, considering the wide diversity and quantity of such pages currently available, this subset was deemed sufficient for the purpose of this experiment. Both slicers used a common set of annotators consisting of the AlchemyApi concept tagging service[4], which identified and associated concepts mentioned within each fragment with Dbpedia instances[5]. The reading-level difficulty expressed as Flesh Reading scores was determined using the open source Flesh annotator[6]. Part of speech[7], noun phrase, and verb phrases[8], were also identified within fragments and annotated with their relevant linguistic attributes. All annotations and fragments were stored as rdf data within a Sesame triple store[9]. The closed corpus slicer additionally associated subject categories as well as content style (ie: bullet point, prose) attributes to individual fragments based on the structure of targeted closed corpus resources (see section 4.2). Finally, the open corpus slicer used a boilerplate detection algorithm[10], annotating to what degree fragments of a page were reusable or not. Slice requests were then converted to SPARQL queries by the slicing unit and submitted to the triple store in order to identify any matching fragment/annotation combinations. Fragments identified were then appended to each other (with an order consistent with their original position in pages) and annotations inserted in the resulting compounded fragment. Since, for the purpose of this experiment, no 3rd party slice consumer was involved, slices were output in xml format.

The result of these pipeline instances provide an open corpus CAS to slice consumers consisting 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:

Slice request example: slices should originate from only a specified list of urls and have a granularity ranging from 3 sentences up to 3 paragraphs. They should cover the topics of "whale migration", "atlantic ocean" or "hunting" and should have a Flesh reading score ranging from 45 to 80. They should not contain any tables or bullet points lists but be focused on the specified topics (ie: exclude content not on these topics). Slices should contain between 7 and 15 annotations consisting of verbs conjugated at the past perfect continuous and should be delivered as LOM objects.

4. EVALUATION & RESULTS

Although the reuse of open corpus material is ultimately aimed at large scale reuse and re-composition, this experiment focuses on evaluating the automated reuse of individual open corpus slices. The assumption is that, in order for open corpus content to be reused through re-composition with other slices, the quality of individual slices delivered must be guaranteed to the AHS consumer. Any quality issues arising within individual content objects would subsequently affect any re-composition produced that includes these objects. An experiment investigating the re-composition of slices in an independent third party AHS, as well as scaling performances, is currently in progress. The overall evaluation strategy adopted within this paper consisted in comparing various reuse metrics across open corpus reused i) in its native form, iv) using a manual approach ii) closed corpus approach, or iv) open corpus approach.

4.1 Language eAssessment

In order to evaluate the proposed approach to open corpus reuse along with the two slicer implementations (for open and closed corpora) presented in 3.2, a real life user-trial was performed in the application domain of language e-assessment. In this scenario, native and non-native English speakers assess their personal English grammatical skills using an online e-assessment application, built specifically for the purpose of this experiment, called Slicegap. This simple application presents users with different sets of open/closed corpus resources reused within the context of grammatical exercises. Each resource is sliced, using the various slicing techniques discussed above, and presented individually to users as traditional gap filler exercises. Verb chunks conjugated at specified tenses are removed and replaced by gaps, which users must fill according to particular infinitives and tenses specified for each gap. The answers provided are compared to the original verb chunks and users are assigned a score for this specific grammar point.

The slice consumer application represents an excellent evaluation platform for the purpose of this experiment for many reasons. As this experiment aims to evaluate individual slices as they are delivered to slice consumers (prior to any adaptation or re-composition), the level of re-composition or adaptation performed on these slices needed to be kept at a minimum. Traditional grammar exercises are by nature created using individual pages (or slices) and do not necessarily require any re-composition. Furthermore, the activity involved allows the interface complexity of the application to be kept to a minimum (Figure 2) to avoid any possible interference with the evaluation of the content.

[3] http://code.google.com/p/jwpl/

[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 chunker in http://gate.ac.uk/

[9] http://www.openrdf.org/

[10] http://code.google.com/p/boilerpipe/

Additionally, although component level evaluations are of course necessary, the content consumed by AHSs is ultimately presented to people. Hence, any open corpus reuse measurement should consider as critical, the user-experience aspects of such an evaluation. For this reason, it was necessary to select a "reuse vehicle" where the user needs are very sensitive to: (i) the accuracy of annotations; (ii) the visual layout; and (iii) the linguistic quality of the content presented. A grammatical e-assessment task requires verbal annotations to be precise and the content to be formatted correctly as well as easily readable by groups of users with varying linguistic competencies.

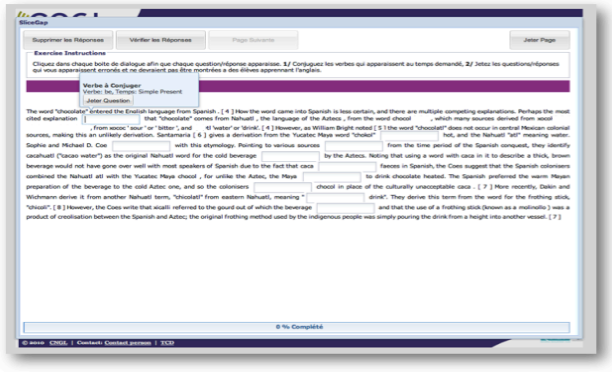

Figure 2 Slicegap Screenshot

4.2 Content Batches

In this experiment, 8 different content batches were presented to users:

1/Content Batch Native (CBN): Content CBN (used as a baseline) consisted of the entire set of open corpus resources used within this experiment in their original native form without any slicing.

2/Content Batch Closed Corpus (CBC): Content CBC consisted of a subset of these resources retrieved as closed corpus content (section 4.4) and sliced using the closed corpus slicer. With the aim of avoiding any confusion, closed corpus content, throughout this research, refers to resources previously known in advance of any slicing performed. Such content, as in the PMCC system described earlier [12], can therefore be sliced with very high precision using a set of manually crafted rule-set algorithms.

3/Content Batch Wikipedia (CBW): Content CBW consisted of the same set of pages available in batch CBC but downloaded in their original form as html web pages.

4/Content Batch Open Corpus (CBO): Content CBO consisted of content batch CBW augmented with a set of random arbitrarily user selected web pages. This content batch represents a true set of open corpus resources since (with the exception of sub-batch group CBW) pages source and structure were unknown in advance of slicing.

The following 4 content batches aimed at evaluating the ability of a slicer to focus native resources upon a topic or set of annotations. For this reason, 2 sets of content were produced both manually and automatically.

5/ Focused Topic Manual (FTM): Batch FTM consisted of pages arbitrarily selected by independent users (section 4.4) on a particular topic T, manually sliced and focused upon content within the original document covering only this specific topic T.

6/ Focused Topic Automated (FTA): Content FTA consisted of the same set of pages independently selected for content batch FTM, sliced and focused automatically upon the same topic T using the open corpus slicer.

7/ Focused Annotation Manual (FAM): Pages within content batch FAM were independently selected in the same way as for FTM but instead focused upon annotations created within this content.

8/ Focused Annotation Automated (FAA): Finally, content FAA consisted of the same set of pages selected for FAM, sliced, annotated and focused upon annotations created using the open corpus slicer.

Furthermore, these 8 content batches can be classified within 4 different content reuse strategies:

1/Manual reuse: As content batch FTM and FAM were produced manually by independent users. These two content batches represent a manual reuse scenario of open corpus content.

2/Native reuse: Since content batch CBN simply consists of unchanged open corpus resources, this content batch represents the equivalent of a traditional IR open corpus reuse strategy (section 2) with resources delivered, in their native form, as one-size-fits-all document objects.

3/Closed corpus reuse: Content batch CBC, on the other hand, represents semi-automated reuse scenarios of content. As part of this reuse strategy, systems (such as the PMCC example in section 2) have full knowledge of the origin and structure of content being reused prior to slicing. For this reason, this form of reuse constitutes the ideal large-scale slicing quality achievable using a form of automated process.

4/Open corpus reuse: Finally, since the origin and structure of pages used to produce content batches CBO, CBW, FTA and FAA was unknown prior to slicing, these content batches represent the equivalent of a large scale, fully automated, reuse of open corpus resources available on the web.

4.3 Aim & Hypothesis

The purpose of the evaluation presented below was to investigate whether the approach to open corpus content reuse, proposed in this paper via slicing, could (H1) improve the reuse of original open corpus resources in their native form (CBN) without (H2) reducing drastically the quality of the original content harvested. In the context of this research, quality refers to the readability and structural coherency of content (for example parts of original page aren't merged with menus from other parts of the page). Finally, assuming the process of automated reuse using closed corpus slicers does indeed improve the reuse of native resources, (H3) can the same performance be achieved using open corpus slicers on any open corpus resources?

The rest of this section lists the hypotheses and sub-hypotheses of the evaluation described previously, followed by the experimental set up and analysis presented in this article.

- H1: Slicing improved the reuse of existing resources
  - H1.1: The correct removal of original clutter present in the resources improved the reuse appropriateness of such content.
  - H1.2: Sliced content retrieve could be correctly focused upon a chosen topic/annotation

- H1.3: The slicer was able to annotate correctly the open corpus content for the purpose of the task being performed.

- H2: Sliced content didn't present any important decrease in quality with respect to the original content
  - H2.1: Sliced content didn't present any important decrease in readability with respect to the original content
  - H2.2: Sliced content didn't present any important decrease with respect to the structural coherence of the original content

- H3: Slicing performance over open corpus content doesn't present any major decrease in performance in comparison to closed corpus.

4.4 Experimental Design

In order to perform an evaluation of the slicer implementations presented in section 3.2, the goal of the following experimental design was to present users with different alternative content, within identical reuse scenarios. This would allow a direct comparison of the four reuse strategies described in section 4.4.1 and thus highlight any differences in content quality perceived by users. These perceived differences would relate to the slicing approach used. Hence, since the aim of this experiment was to focus specifically upon measuring the ability of a slicer to improve the reuse of open corpus resources, as opposed to its ability to identify specific resources, the components of the slicer pipeline built for the purpose of accomplishing such a task were isolated. For this reason, the IR harvesting component of both slicers was omitted from the trial and only the URL list-harvesting feature was used. This ultimately encapsulates the research question as follows: Assuming a correct open corpus resource was initially identified, can slicing improve its reuse?

4.4.1 Content batch creation

As the aim of the experiment is to investigate whether resource reuse improvement, through the use of slicing, is achievable on any open corpus content, the need to initially harvest a truly random set of open corpus pages was necessary. Hence, prior to conducting the experiment, a group of five independent English teachers, from two different schools, were asked to arbitrarily select nine pages of their choice from the web (combined total of 45 pages) and perform a set of tasks. They were first asked to select a set of pages of their choice from Wikipedia. They were then asked to create a second, larger set of pages, harvested from any source of their choice (as along as it consisted of news articles see 3.2). Finally, a third set of pages chosen by teachers was required to be on a specific set of topics T. English teachers were then asked to select fragments of the pages harvested (according to specified granularities and tense requirements) and manually annotate tenses encountered within these extracts. Finally, they were asked to extract fragments from a portion of the pages harvested in tasks 2 and 3 and manually focus these fragments respectively upon tenses annotated and content referring to individual topics T (content not about these topics was discarded). These last two sets of pages created represent content batches FAM and FTM respectively (section 4.2). Manually produced fragments contained an average of 189 words and 14 annotations.

All of the extracts created were subsequently converted into grammar e-assessment pages. The entire collection of pages collected in each set was then harvested from the web, in their original form, to produce batch CBN. CAS characteristics of extracts manually created, were identified and fed into the open corpus slicer as parameters. The entire content set in its original form was then sliced with these parameters to produce the page set CBO (including set CBW) with similar CAS characteristics as its manual equivalent. The Wikipedia pages downloaded were then matched to the set contained in the database (see section 3.2) and sliced (using again the same CAS parameters as their manually created equivalent) using the closed corpus slicer to produce batch CBC. Since both CBC and CBW, were sliced using respectively closed corpus and open corpus methods from identical source of pages, any improvement in reuse differences detected between the two sets can only be the result of slicing performances between closed corpus and open corpus slicing. Finally, in a similar way, original pages used to create FAM and FTM were fed into the open corpus slicer. The slicer was provided with annotation and topic focused parameters to produce FAA and FTA respectively. As in the previous case, these two parallel content sets are derived from the same original pages. This allows a direct comparison between the performance of manual and automated focused content.

4.4.2 Evaluation scenario

Once the eight content batches had been produced, a user trial was conducted which consisted of a task-based evaluation using real-life linguistic assessment needs. Each user was initially asked to select their interface language of choice and specify their level of ability in the English language. They were subsequently invited to perform successively the same activity eight times in succession, using eight different pages randomly selected from each content batch. Each piece of content was associated with a unique colour and users were unaware of either how the content was created or what content batch was being presented to them at each task. Each original page (regardless of the content batch it belonged to) was only shown once to each user. The trial performed consisted of a traditional language learning activity that is encountered in most textbooks (section 4.1). The activity was divided into four individual tasks:

User Task 1: The first task required users to read the text presented to them and fill in any blanks encountered (there were 10 gaps on average per page) with the appropriate verb and tense specified for each case (Figure 2). If users felt slices were unreadable or inappropriate for this exercise, a "Trash Slice" button could be pressed to select another page. The ability to provide users with accurate grammar assessment questions is highly sensitive to the precision of the annotations delivered by the slicer, therefore this task aimed to evaluate both the annotation quality and overall reuse appropriateness of the slices presented.

User Task 2: The initial task did not necessarily require users to read the entire content presented to them (only the sentences containing gaps). For this reason, users were then asked to summarise what they felt the entire piece of content was about (as in traditional textbooks) in their own native language.

User Task 3: Finally, although previous tasks are clearly dependent upon the quality of the slices presented to users, the ability of a slicer to correctly scope (or focus) an original document on a particular topic or set of annotations certainly was not measured. For this reason, users were asked an additional set of questions, which assess the appropriateness of the scope of each slice. The concepts selected for presentation were chosen to be accessible to as wide an audience as possible. This enabled users to accurately judge the scope of the content presented. Once

completed, the answers received for slices that were manually focused (FTM & FAM) were compared to those that had been automatically generated (FTA & FAA) by the open corpus slicer. Although the authors are aware that an activity which was directly dependent upon the scope of slices would have presented a more convincing argument, these results nevertheless provide an initial indication of a slicer's ability to correctly focus native content. A further experiment which is currently in progress, assesses slices that are produced through re-composition and addresses this issue in a different application domain (section 6). Additional questions related to task 1 and 2 were also included in order to reinforce quantitative measurements. All questions presented to users offered a 10 point Likert scale (ranging between "strongly disagree" and "strongly agree"). A scale with no mid-point was was deliberately used to enforce user preference however slight. The order of sentiment in the scales was also randomized between questions in an attempt to ensure that users were genuinely attentive when answering each question..

User Task 4: Finally, users were asked to order a set of colours, corresponding to each content presented, based on their perceived quality. In order to balance any effect of order bias, each user was presented content batches according to a Latin square design distribution. The entire experiment was available online to the public and users were asked to perform the entire set of tasks in a single session without any interruption. Task completion time in addition to user interactions were tracked throughout each activity. Finally, as part of a larger experiment comparing the automated production of language e-assessments with respect to its manual equivalent, non-native English speakers were also invited to perform the experiment. Hence, the interface of the slice consuming application Slicegap (including the questionnaires) was also translated into Spanish and French.

4.5 Results

This section presents a summary of the findings observed throughout this experiment in relation to each hypothesis. The full detailed list of results on the 36 variables measured can be accessed online[11]. A total of 41 users across 7 different countries performed the experiment. Most of these users performed the experiment using the English interface (en=59%, non-en=23%) and rated themselves as native or advanced English speakers (Native=46%, Advance=28%, Intermediate=12%, Beginner=3%, Other=1). Finally, the number of times the "Trash Slice" button (section 4.4) was clicked throughout the experiment was statistically insignificant. For this reason, it is not further discussed within the rest of this paper.

4.5.1 Reuse Improvement (H1)

H1.1: As pointed out in section 2, a part of maximizing the reuse potential of a previously published resource requires the ability to decontextualize this content from its original setting. Hence, users were asked directly, for each content, whether "*in addition to the main content, a lot of material displayed on the page was irrelevant to the task (such as advertisement, menu bar, user comments..)"*. Figure 3 presents the results obtained over a Likert scale ranging from 1 to 10 (section 4.4.2). Content batch CBN, containing pages in their native forms, achieved the worse score with a mean equal to 5.67 while closed corpus (CBC) and open corpus (CBO) content batches achieved much better scores (CBC=1.76, CBO=2.53) with paired t-tests confirming results are indeed significant (p=0.001). Additionally, when comparing closed corpus (CBC) and open corpus (CBW) approaches on an identical set of pages, a mean difference of 0.879 between CBC and CBW was measured (p=0.015). This result suggests that, although the open corpus approach to slicing did achieve a very good performance with respect to its closed corpus equivalent (91% similarity), users did notice a minor performance decrease in it's ability to separate the main content of the original resource. When asked whether *"the level of irrelevant material (e.g: menus, advertisement, user comments etc…) present in the content prevented me from pursuing this task effectively"*, CBN again achieved the worse score (3.43) versus closed (CBC=1.48) and open (CBO=1.64) content batches (with paired t-tests confirming these results are statistically significant (p<=0.01)).

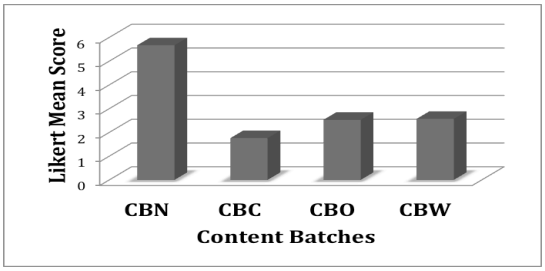

Figure 3 De-contextualisation Performance

This observation, suggesting poor de-contextualisation of original resources impacts the reuse of such content, was confirmed when measuring the average time, per answer provided by users, necessary to complete the exercises. As can be seen in Figure 4, users required an average of 88% more time to complete the exercises using non-decontextualized resources in their native form (CBN) than it did for fully de-contextualised resources (CBC). This finding appears to confirm the assumption stated by Lawless et al [7] earlier (see section 2). Furthermore, differences in results measured between CBC and CBW for both the question and average time measured were insignificant (p=0.77 & p=0.58). This suggests that although a performance decrease in de-contextualising original resources for the slicer using an open corpus approach was noticed by users, this minor performance decrease didn't have any impact upon the reuse of the content produced.

H1.2: As mentioned in section 3.1, the ability to right-fit a resource in its native form into a slice containing only parts of the original document about a specific topic or set of requested annotations, is important with respect to the capacity of a slicer to deliver information objects matching specific content requirements and use cases of various AHS. Content batches FTM and FTA represent such a case of original resources being focused (or scoped) on an arbitrarily chosen topic. When users were asked whether they *"felt parts of the content presented were about topic <T>"* (<T> being the topic in question), the manually focused content batch (FTM) obtained the best score with a mean equal to 9.09 versus the open corpus automated equivalent (FTA) which obtained 8.55. A p value of 0.82 suggests this mean difference between manually and automatically generated contents is insignificant. This result appears to suggest the SOC slicer was capable of automatically identifying which parts of the documents did refer to specific topics.

[11] https://www.scss.tcd.ie/~levachk/TCDWebsite/data.html

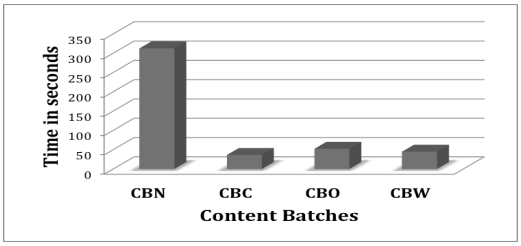

Figure 4 De-contextualisation Impact

Additionally, when asked whether *"a large amount of content delivered was not related to topic <T>"* the manually focused content again achieved the best score in comparison to the open corpus slicing approach with a mean equal to 3.07 and 3.86 respectively for FTM and FTA. The paired t-test (p=0.159) suggests once more that the difference measured between the manual and automated approach is insignificant. As a sanity check, the answers for the same question asked for native content were also measured. Since pages within content batch CBN are presented as entire documents, they are necessarily unfocused; hence a very poor focused performance should be measured. As expected, a mean value equal to 6.43 for CBN suggests users felt this native content wasn't focused correctly on topic T. When asked the reverse (ie:*"all the content presented was related to topic <T>"*), the results lead to the same conclusion with FTA and FTM obtaining respectively 7.36 and 8 (p=0.411) while CBN received a score of only 1.23. These results appear to indicate that not only could the open corpus slicer SOC correctly identify parts of the document on a particular topic, it could also minimize the amount of un-related content delivered, and offer slices with boundaries tightly confined to that specific topic. Moreover, paired t-tests (p=0.159 & p=0.411) do suggest content focused automatically using the SOC slicer achieved similar scores to its manual equivalent. Finally with respect to open corpus content being correctly focused upon annotations requested, a similar pattern can be observed. When asked whether *"apart from any clutter information (advertisements, menus etc…) displayed within the page, I felt a large amount of content delivered to me was un-necessary for the grammar task being performed"*, content manually focused upon annotations (FAM) achieved the best score (3.53) with the automated approach presenting a minor difference (mean=3.73) considered insignificant (p=0.783). A sanity check with CBN again obtained the worse score of 5.87. These results indicate the automated open corpus slicer could correctly focus native content upon annotations requested with a performance similar to it's manual equivalent.

H1.3: In order to measure the accuracy of annotations delivered to slice consumers, a sample set of identical pages annotated by all teachers (section 4.4) was compared to those produced automatically by the open corpus slicer. Individual manual annotations, obtaining the largest agreement among teachers, were used as a golden standard. Precision and recall was calculated for the annotations produced by the open corpus slicer as well as for each manual annotator. An F score equal to 0.88 (precision=0.86, recall =0.92) was measured for the open corpus slicer. This result suggests the annotator identified most items to be annotated while producing a minority of errors during the process with respect to human annotators. This high correlation between automated and manual annotations hence appears to confirm the quality of annotations produced by the automated open corpus approach to slicing. Additionally, when performing the same measurement for each human annotator individually, with respect to the golden standard, an average F score of 0.77 (Precision=0.89, Recall=0.68) was obtained. This result shows that although human annotators did a better job at correctly annotating the fragments, some disagreements did occur; suggesting mistakes produced by the automated approach could be subject to interpretations. Additionally, the recall score indicates that many human annotators missed several annotations in comparison to the automated approach.

4.5.2 Quality Decrease (H2)

H2.1: As described in section 4.3, the process of structurally fragmenting original resources into individual pieces and subsequently merging selected fragments together to form slices, can lead to structural incoherencies within the final content delivered. For this reason, users were asked directly if *"parts of the original web page presented were merged inappropriately together (eg: end of menu merged with start of a paragraph)"* (Q1). Table 1 depicts the results obtained per content batch. As can be observed, very good performances were measured across all content batches, with t-tests estimating any differences encountered as insignificant. These results suggest that the process of slicing resources didn't produce any significant increase in structural incoherencies in comparison to the original content in its native form. Moreover, a mean difference of 0.688 observed between CBC and CBW was considered insignificant. This would indicate that, although measurements for the open corpus slicing approach were higher than for closed corpus slicing, these results do not provide enough evidence to suggest open corpus slicing leads to any significant decrease in structural coherency of slices produced. When asked the reverse (Q2), opposite values were also measured, with similar t-tests, which confirms these observations.

Table 1 Structural Coherence

		Mean	Comparison	Mean Dif	p
Q1	CBN	3.38	CBN v CBC	0.25	0.734
	CBC	3.13	CBN v CBO	0.28	0.662
	CBO	3.65	CBC v CBW	0.688	0.357
Q2	CBN	8.19	CBN v CBC	0.313	0.608
	CBC	8.5	CBN v CBO	0.531	0.479
	CBO	7.65	CBC v CBW	0.75	0.118

H2.2: While previous measurements aimed at estimating any decrease in quality of the content delivered from the point of view of visual layout, the readability of any content produced for reuse is of course critical. This research defines readability as being composed of two major components, namely i) the ability of content to be easily understood, and ii) the extent to which the flow of reading is broken. The second component refers directly to the notion described by Debasis et al [8], in other words, the extent to which pronouns, for example, referring to earlier subjects missing from a fragment, affect the ease of reading and comprehension of a text. Figure 5, depicts the results obtained for the first component measured. The first pair of questions aimed at asking users directly their opinion with respect to how easily could content presented to them be understood. The second pair relates to measuring any impact upon the difficulty in summarizing content. As can be seen, results obtained across content batches are very similar. While native (CBN) and closed corpus content slicing approaches interchangeably achieved better results across questions, open corpus content sliced using the open corpus slicing approach (CBO) constantly performed lower than the latter two. Paired t-test between CBC and CBW content however estimated any mean differences measured between content produced using closed corpus or open corpus slicers as insignificant ($p > 0.06$ for all questions). As a sanity check, and in

order to make sure all sentences within content presented were read (regardless of whether they contained gaps to be filled), users were asked to summarize the meaning of the content presented to them. Among all the summaries provided by users across content batches, only less than 5% were rated as "weak". Despite being on topic, these answers were too short (less than 5 words) to determine whether the overall meaning of the content was correctly understood or not.

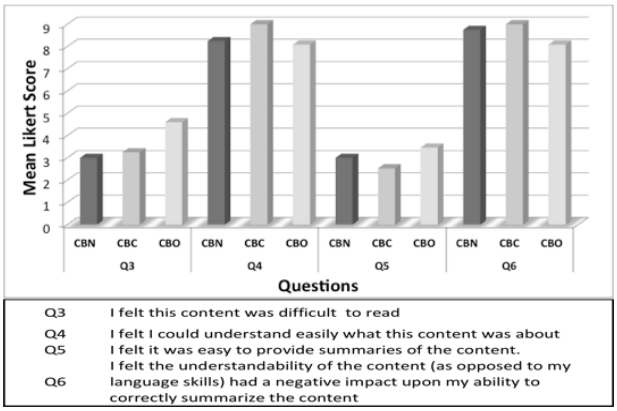

Figure 5 Slice understandability

Nevertheless, the overall quality of summaries submitted clearly supports previous results obtained. Since close to a third of users, who participated in the trial where non-native speakers, a Pearson correlation analysis between answers submitted to each question and English level of users, was also performed. Correlations measured for each variable combination ranged between -0.2 and 0.2, which indicates no direct correlation between answers provided and the level of users. Additionally, when users were asked whether they felt *"their language skills had any negative impact upon [their] ability to correctly summarize content"*, means measured across content batches were all inferior to 2.30. With respect to the second component of readability, there exists, to the best of our knowledge, no automated mechanism to evaluate whether pronouns or topics mentioned within a fragment referred to earlier subjects mentioned in omitted parts of the original content. Hence users were asked directly whether this situation occurred or not for each content presented to them (Q7). Results depicted in Figure 6, indicate CBN obtained the best results, which is expected since the entire original resource was presented to users. Closed corpus content (CBC) presented no statistically significant differences with its native alternative (CBN) ($p=0.685$), while results for open corpus approach (CBO) on the other hand, presented a small statistically significant decrease with respect to CBN (mean dif =1.03, $p=0.037$) for confidence intervals of 95%. This result suggests open corpus slicing can indeed slightly deteriorate the flow of reading of open corpus content, which makes common sense. However when asking users if this situation had a negative impact upon the readability of content presented to them (Q8 & Q9), no statistical difference could be noticed between contents ($p>0.300$). Hence although, a slight decrease in the flow of reading for content sliced using an open corpus approach could be noticed, the resulting impact upon the overall readability of content delivered was marginal.

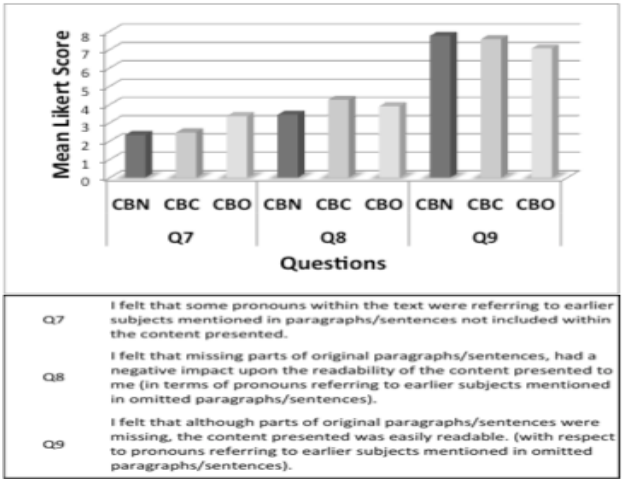

Figure 6 Slice reading flow

4.5.3 Open Corpus Slicing Performance (H3)

Throughout the results presented previously in section 4.5.1 and section 4.5.2, slices produced by the open corpus slicer consistently depicted lower performances in comparison to their rule based closed corpus or manual equivalent. However, in most cases, differences measured between slices produced by the two closed and open corpus slicers were statistically insignificant. Two exceptions to this rule occurred (both when comparing CBC and CBW content batches): the first occurred with respect to the ability of the SOC slicer to separate the main parts of pages from any clutter (H1.1), while the second occurred when analysing any broken flow in readability (H2.1). Albeit, both cases were anticipated, (since this slicer operates without any prior knowledge in the targeted resources to process) differences measured between slices produced from open and closed content were very small. Moreover, when measuring the effect both cases had upon user experience, no significant differences could be noticed.

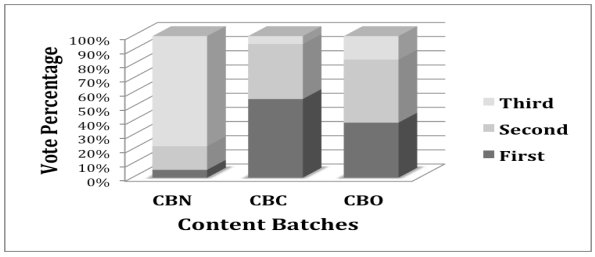

Figure 7 Content batch ordering

4.5.4 Results Summary

In summary, the results of the evaluation strongly indicate that the ability of slicers to de-contextualize resources from their original settings and focus this content upon specific topics or annotations, improves the reuse of open corpus resources (H1). Although, the flow of reading in slices produced, deteriorated in comparison to the original resources (H2.2), no significant impact upon the understandability of the content could be measured. These experiments also showed how, although open corpus achieved lower performances than closed corpus, any difference tended to be very small and its measured impact insignificant. Finally, after being presented with each content, users were asked if *"Overall, [they] felt this content was adequate to perform such an*

exercise". Content semi-automatically (CBC) and automatically (CBO) sliced achieved the best scores, with means respectively equal to 8.89 and 8.57 in comparison to native content (CBN), which only achieved 6.85 ($p<0.034$) (Figure 8). Moreover, when asked to order contents in order of preference (the lower the value the better), 56% of users placed CBC content in first position (Figure 7), closely followed by CBO, with the overall majority placing content in its native form (CBN) in the last position. This result does confirm a rule-based approach to slicing upon closed corpus offers the best reuse improvement of native content. However, if the source of content to slice is unknown in advance, the open corpus slicing approach will perform with similar performance with very little impact from the point of view of user experience. These overall results reinforce the trend observed across variables, which suggest the reuse of existing resources was improved through the approach of slicing, with open corpus slicing reuse achieving very close performance to its closed corpus equivalent.

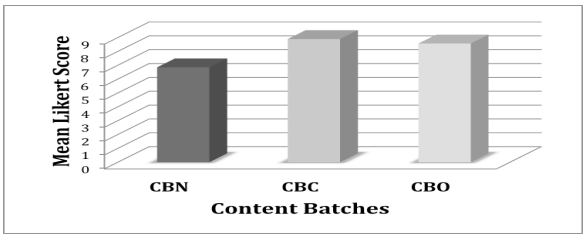

Figure 8 Content batch preferences

5. DISCUSSION

The measurements obtained by the user trial have shown encouraging results for the large-scale full reuse of open corpus material within various OAH systems through an open corpus slicing approach. Despite the identification of minor performance differences between open and closed slicing techniques, the impact upon the user experience appeared to be insignificant. It may be argued that the reuse scenario (including the choice of what constitutes discerning meta-data) selected for the purpose of this experiment only represents one specific use case of open corpus reuse. Indeed, further investigations will be required to determine the range of reuse scenarios possible, which slicing approaches could support. However, this experiment strongly suggests that the ability to go beyond the one-size-fits-all delivery of open corpus content and automate the right fitting of such resources for reuse within various content consumers is possible, with a minimal decrease in quality of this original content. Hence, for at least a selected number of use cases, large scale automated "full" reuse of such resources in AHS is possible. Subsequent steps (section 6) will be required to demonstrate whether these individual slices indeed be re-composed within third party AHS in various reuse scenarios.

6. CONCLUSION

This paper has presented a new approach to open corpus reuse through the automated customization and right fitting of heterogeneous resources available on the web called slicing. An initial implementation of this approach was successfully applied within the application domain of a language e-assessment scenario by reusing random selected resources available on the web. Evaluation results provide evidence that slicing improves the reuse of open corpus resources while minimizing any decrease in the quality of original content delivered. The results also appear to suggest open corpus reuse through slicing achieved very close performance to its closed corpus equivalent with minimal impact upon user perception. This provides evidence that large scale automated reuse of open resource is achievable. Further investigations are currently integrating Slicepedia with third party adaptive eLearning systems targeted at young teenagers in the area of natural sciences.

7. ACKNOWLEDGEMENTS

This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (www.cngl.ie).

8. REFERENCES

1. Aroyo, L., Bra, P. De Embedding Information Retrieval in Adaptive Hypermedia: IR meets AHA!". *New Review Of Hypermedia And Multimedia 10*, 1 (2004), 53 - 76.
2. Brusilovsky, P., Chavan, G., and Farzan, R. Social adaptive navigation support for open corpus electronic textbooks. *AH'04: Proc. of the 3rd int. conf. on Adaptive Hypermedia and Adaptive Web Based Systems*, (2004).
3. Dieberger, A., Jose, S., CoWeb - Experiences with Collaborative Web spaces. In D. Lueg, Christoph and Fisher, ed., *From Usenet to CoWebs: Interacting with Social Information Spaces*. Springer, 2002, 155 - 166.
4. Fernandez, M., Lopez, V., Sabou, M., et al. Semantic Search Meets the Web. *ICSC'08: Proc. of the int. conf. on Semantic Computing*, (2008), 253 -260.
5. Henze, N. and Nejdl, W. Adaptation in Open Corpus Hypermedia. *IJAIED'01: Int. Journal of Artificial Intellligence in Education*, (2001), 325 - 350.
6. Kohlschütter, C. and Nejdl, W. A Densitometric Approach to Web Page Segmentation. *CIKM'08: Proc. of the 17th int. conf. on Information and knowledge management*, (2008), 1173-1182.
7. Lawless, S. Leveraging Content from Open Corpus Sources for Technology Enhanced Learning., 2009.
8. Leveling, J. and Jones, G.J.F. Utilizing sub-topical structure of documents for Information Retrieval Categories and Subject Descriptors. *Utilizing sub-topical structure of documents for Information Retrieval*, (2011)
9. P, B. and Pesin, L. Adaptive Navigation Support in Educational Hypermedia: An Evaluation of the ISIS-Tutor. *Journal of Computing and Information Technology*, (1998).
10. Pennacchiotti, M. and Pantel, P. Entity extraction via ensemble semantics. *int. conf. on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics (2009), 238.
11. Speretta, M. and Gauch, S. Personalized Search Based on Search Histories. *int. conf. on Web Intelligence*, (2005).
12. Steichen, B., Connor, A.O., Wade, V., and Oconnor, A. Personalisation in the Wild – Providing Personalisation across Semantic , Social and Open-Web Resources. *int. conf. on Hypertext and Hypermedia*, (2011), 73-82.
13. Zhou, D., Goulding, J., and Truran, M. LLAMA: automatic hypertext generation utilizing language models. *Hypertext and Hypermedia*, (2007).