Applying Digital Content Management to Support Localisation

Alexander O'Connor[1], Séamus Lawless[1], Dong Zhou[1], Gareth J. F. Jones[2], Vincent Wade[1]
[1] Centre for Next Generation Localisation
Knowledge & Data Engineering Group
School of Computer Science & Statistics
Trinity College Dublin
Dublin 2, Ireland
[2] Centre for Next Generation Localisation
Dublin City University
Dublin 9, Ireland
www.cngl.ie
Alex.OConnor@cs.tcd.ie, Seamus.Lawless@cs.tcd.ie,
Dong.Zhou@cs.tcd.ie, Vincent.Wade@cs.tcd.ie
Gareth.Jones@compuing.dcu.ie

Abstract

The retrieval and presentation of digital content such as that on the World Wide Web (WWW) is a substantial area of research. While recent years have seen huge expansion in the size of web-based archives that can be searched efficiently by commercial search engines, the presentation of potentially relevant content is still limited to ranked document lists represented by simple text snippets or image keyframe surrogates. There is expanding interest in techniques to personalise the presentation of content to improve the richness and effectiveness of the user experience. One of the most significant challenges to achieving this is the increasingly multilingual nature of this data, and the need to provide suitably localised responses to users based on this content. The Digital Content Management (DCM) track of the Centre for Next Generation Localisation (CNGL) is seeking to develop technologies to support advanced personalised access and presentation of information by combining elements from the existing research areas of Adaptive Hypermedia and Information Retrieval. The combination of these technologies is intended to produce significant improvements in the way users access information. We review key features of these technologies and introduce early ideas for how these technologies can support localisation and localised content before concluding with some impressions of future directions in DCM.

Keywords: *Digital Content Management, Information Retrieval, Adaptive Hypermedia, Content Analysis, Open-Corpus Content, Multilingual Technologies*

1 Introduction

Digital Content Management (DCM) is concerned with the creation, transformation, storage, retrieval and presentation of information in digital form. At present, the most publicly visible resource available to DCM applications is the World Wide Web (WWW). The current approach to content management for web applications is very limited by the assumption that content is largely static and by providing access via search engines which broadly assume static file collections held individually on specific servers. However, this is rapidly becoming an outdated model of the way that most information exists on the web. Static file structures are giving way to web-based content-management systems, which compose responses dynamically using content stored in databases. This

content can be presented to the user in different ways depending on style, accessibility or security preferences. The web itself is becoming a collection of highly-diverse content management mechanisms. This is creating substantial challenges to the satisfaction of user information needs because this heterogeneity of data sources introduces complex obstacles to computational methods for managing content. In addition, with nearly 500 billion gigabytes of information being stored worldwide (Wray, 2009), the need to be able to find and index specific information has become a massive global challenge. In order to be able to better support users with complex information needs, it is also necessary to develop new ways of responding to users that go beyond the conventional ranked listing of documents.

Our belief is that an effective way to address these challenges can draw on two principal areas of research: Information Retrieval (IR) and Adaptive Hypermedia (AH). Research in IR underpins existing search engines such as *Google*, and enables efficient search for relevant documents among the billions of items currently available on the web. A particular challenge in the selection and presentation of this content is the increasingly multilingual nature of digital content. Effective DCM systems need not only to find and present content, but they need to do this in a multilingual environment with the output ultimately in a form that can be reliably and comfortably consumed by the user. Search and presentation already presents challenges; extending this to more personalised formats is considerably more demanding. AH research meanwhile focuses on the view that the power of digital content is its malleability. AH technology takes as its goal the creation of highly tailored, rich media presentations designed for the specific needs of the user. AH technology has its roots in eLearning systems, which teach complex concepts to students using rich media experiences. The main limitation of current AH techniques is that they have to-date focused on small, carefully controlled content sources, making them unsuitable for highly heterogeneous data sets such as the WWW. Further, as digital technologies proliferate, there is a compelling need to address the issue of documents authored or stored in different natural languages.

Localisation, based on the manual or machine translation of content, is thus a major concern and opportunity for DCM. Localisation is a mature domain with substantial industrial experience in many issues associated with managing corpora of content in different languages designed to serve different groups. It is impossible to attempt to address the challenge of effective global DCM without also addressing the language and localisation of content selection and presentation. DCM technologies thus aim to provide new functionalities for addressing emerging opportunities and challenges of localisation for dynamic multilingual content.

Work in DCM within the CNGL is seeking to integrate IR and AH technologies with language translation and input derived from existing experience with localisation. The goal of this integration is to develop novel and effective technologies for personalised responses to user information needs taking data from open, multilingual heterogeneous data sources. This paper introduces some of the key background technologies which are being investigated within the DCM track of the CNGL. The paper begins with background reviews of AH and IR, and of some of the specific techniques being used to make the wide variety of content available on the WWW more accessible. With these technologies outlined, the paper then describes some potential applications for different combinations of these technologies in the specific area of localisation. The paper concludes with some remarks on potential future directions in DCM research.

2 Adaptive Hypermedia

Conventional static web content management systems present the same responses to all users regardless of their preferences or other personal factors. However, these classic ‘one size fits all’ content delivery systems are simply not powerful enough, particularly as the WWW becomes

increasingly dynamic and multilingual. Web delivery systems and hypermedia systems are increasingly attempting to customise content so that it is relevant to the user or the context of use. This can be achieved by, for example, changing the presentation of the content for different screen formats, or by allowing users to alter its layout manually. These technologies separate some elements of content from presentation, for example by using stylesheets which can take account of personal or localised needs. However, the majority of these systems do not go far enough towards meeting individual needs.

We believe that AH technologies can make an important contribution to extending the limitations of current web-based content management systems. Such new systems can make it possible to deliver "personalised" views of a hypermedia document space without requiring programming from the content author. This is achieved by building a model of the goals, preferences and knowledge of the individual user, and using this model to dynamically compose responses tailored to the individual user. For example, a large component of the value of digital content in elearning is in the targeted delivery of that content to the right user in the right form. This is typically achieved using three specific component models: a user model, a content model and a domain model (Conlan 2004).

User models can be initialised by explicitly illiciting information from the user using a questionnaire or through the use of stereotypical user models. User models can also be evolved automatically through adaptation by simply observing the browsing behaviour of the user. In this way, the user model can continually adapt as the user works with the system and their preferences and knowledge develop.

At present, hypermedia systems are generally restricted to the use of "closed" content collections. It is assumed that the content to be used by these systems is authored so that the collection consists of pieces of content each of which covers some number of related concepts within a subject. These content pieces are typically annotated with highly structured metadata describing various features of the content. There are a variety of international standards for such document description. Some standards, such as Dublin Core (Dublin Core), support the creation of metadata to describe a document in a domain agnostic fashion. Some are more domain specific, such as Learning Object Metadata (LOM) in the eLearning domain. This metadata is generally added manually, meaning that the cost of producing content for use in AH systems is often very high. More recent systems focus on automatic, or semi-automatic generation of metadata as part of the content authoring process. Others focus on generating the metadata based on the context within which the content was originally developed. A third approach, explained below, focuses on inspection of content chunks to facilitate the generation of the metadata. However, manual annotation (metdata tagging) is still quite common due to the importance of accurate high quality metadata in AH systems. Because of the expense of authoring such content, one of the goals of research for AH systems is to maximise the exploitation and reuse of content in order to recoup the return on investment of content creation.

The third element of most AH systems, the domain model, contains a conceptual description of the subject area(s) of interest and a specification of the relationships between these concepts. By dynamically combining the domain model, user model and content model, AH systems can generate personal navigations of adaptively retrieved relevant content.

Having a detailed knowledge of the content when the system is designed is obviously rather different to the situation of the designer of an IR system where the designer often has very little knowledge of the features of the content which is to be indexed or searched. The challenges resulting from these contrasting approaches are addressed in more detail in section 4 of this paper. A good introductory review of current AH technology is contained in (De Bra, 1999) and (Brusilovsky, 1996).

AH systems use various models to generate a navigation through dynamically (adaptively) retrieved content. In addition to the user model, content model and domain model, more recent AH systems have also begun to use other models e.g. models describing the context within which the user is seeking information and a model of the device upon which the retrieved information is to be viewed (Conlan 2004).

2.1 Adaptive Hypermedia Functions

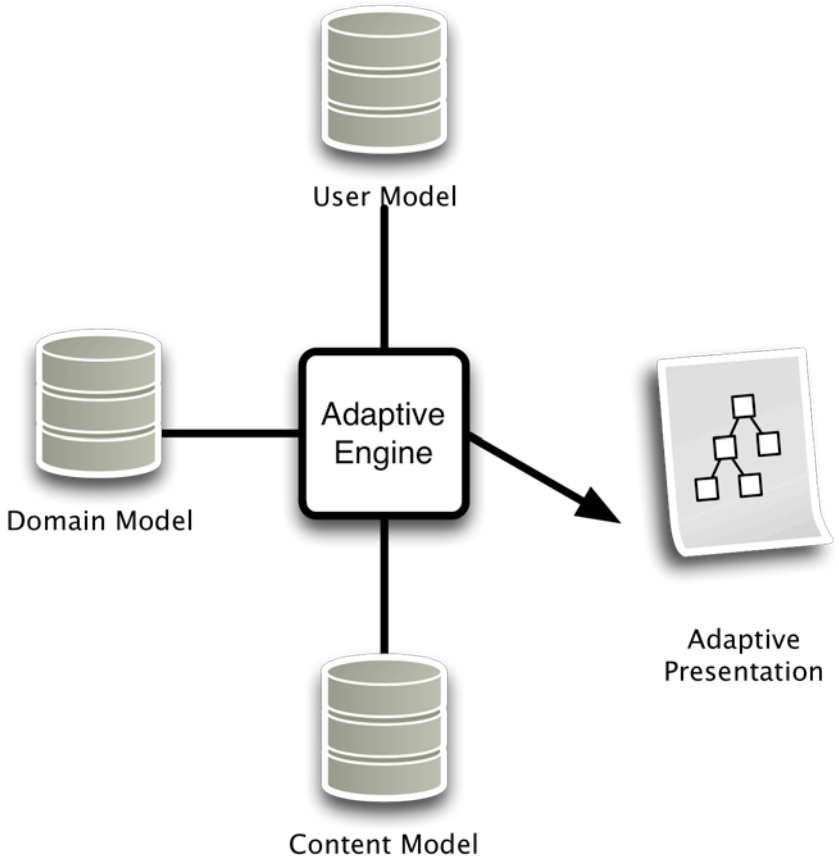

Figure 1: An Adaptive Hypermedia Framework, showing the influence of three models contributing to the creation of an adaptive presentation

An AH system can be thought of as supporting three functions:

• While the user is interacting with the system all, or selected, user actions are registered. Based on these actions, previous behaviour and other user-supplied or automatically-gathered information, the system builds a model of the user's knowledge about each domain model concept. The system seeks to model how much knowledge the user has about the concept and what information they have read about it, as well as other attributes about the user which can be used to tailor the adaptation process.

• The adaptive system reconciles the user model to classify all nodes (content pieces) into one of several group's depending on the user's current knowledge, interests and goals. The system manipulates links within nodes (and link destinations) to guide users towards appropriate, interesting and relevant information. This is called *adaptive navigation* in (Brusilovsky 1996).

• In order to deliver the content of a page at an appropriate level of difficulty or detail the system can conditionally show, hide, highlight or dim page fragments. This process is referred to in

(Brusilovsky 1996) as *adaptive presentation*. This is an important component in AH systems and is outlined in the next section.

2.2 Adaptive Presentation

Selecting content for presentation to the user is only part of the functionality of AH systems. In addition, they typically include elements of adaptive presentation. The selection of these pieces can have different consequences for the final presentation depending on the range of mechanisms employed to create the presentation. Adaptive presentation can enable:

• the provision of fundamental, additional or comparative explanations: For example the AH system can add important background knowledge for novice users. This can be achieved in two primary ways; either by including additional content, or by allowing for 'stretchtext'. In the first case, the system can attempt to predict which additional information might be needed by the user based on their user model, and create a presentation which includes additional content pieces directly in the text (De Bra, 1997). In the second case, the user can click on a particular term or element of the presentation, and the system will adaptively retrieve the appropriate content and present it separately.

• the provision of explanation variants: Depending on the user model, a variety of elements can be adapted: the level of difficulty, the links to related concepts, the length of the presentation, or the media type (text, images, audio, video). This can be done within a page or through guidance towards different pages in a process referred to as adaptive navigation support.

• the re-ordering of information: The user model can be used to vary the order in which information is presented to the user, similar to ranking in information retrieval. This can be used, for example, to create shorter or longer presentations, or to create presentations which are easier to browse by having summary information appear first.

AH is not just dependent on the existing hyperlinks within a document (or content piece). Adaptive link insertion allows for new, dynamically generated paths through the content space to be generated. This provides the appearance of new aggregrations of hyperlinked documents which are formed just-in-time for a particular user. This allows an AH sytem to, for example, annotate different link with a 'Traffic Light' metaphor, where the system can help the user make navigation choices that best suit their knowledge and preferences. This helps to keep the progress of a user through the content smooth and consistent. Several examples of adaptive navigation support can be seen in (Brusilovsky 2004).

3 Multilingual Information Retrieval

The standard aim of IR is to satisfy a user's information need. IR systems attempt to fulfil this objective by returning a list of potentially relevant documents to the user. In order to satisfy the information need the user needs to extract the information from the documents, typically by reading them. Depending on the information need and the structure and contents of the documents, this can be a very efficient or very inefficient process. An example of how this process can be very inefficient is if the document contains large amounts of information which is not of interest to the user, either because it is off topic or because the user is already familiar with most of it. If the user is seeking small new details buried deep within the content, this becomes a very labour intensive process. We believe that AH methods have scope to begin to address these weaknesses of current IR methods.

Algorithms for IR are typically based on statistical techniques which count the frequency or rarity of words in document collections. The most popular standard measure for this approach is referred to

as the tf.idf weighting scheme (Term Frequency - Inverse Document Frequency) (Salton and Buckley 1988). Tf.idf measures the importance of a particular term in a document when considered relative to that term's importance in the overall corpus. This is achieved by combining the term frequency within an individual document with its distribution in the collection as a whole. A simple document ranking can be produced by summing weights for terms (words) occurring in both the user's query and each document. Statistical techniques such as these also feature in probabilistic models, such as BM25 (Spärck Jones et al. 2000a) (Spärck Jones et al. 2000b), which shows increased resistance to noise in identifying particular documents in a collection.

In addition to ranking documents based on the match between a user's query and the document content, for hyperlinked document structures, such as the web, algorithms such as HITS (Kleinberg 1999) and PageRank (Brin and Page 1998) take advantage of the network structure to determine document significance dependent on their place within the network, but independent of their content. Web search engines thus provide an overall ranking of documents in response to a user search requests by combining content matching scores with a network-based measure of the document's likely importance to the user.

Multilingual Information Retrieval (MLIR) moves IR beyond the situation where the query and documents are expressed in a single language, to environments where documents may be a range of languages and queries can be performed in any of these languages. Within MLIR much attention has been focussed on the simplier problem of cross-language (or bilingual) information retrieval (CLIR) where search topics or requests in one language are used to retrieve documents in one other language.

Supporting MLIR requires two main features: adapting IR methods to the each document language and developing strategies for translating between query and documents languages to cross the language barrier. While early work in MLIR concentrated on text documents from published news sources, more recent work has extended this to explore various multimedia IR data sources including annotated photographic and medical images, spoken data sources, and multilingual web documents (CLEF).

In the case of CLIR there is the inescapable additional issue that while a retrieved document may be relevant to the information need, the user may not have sufficient knowledge of the document language to be able to identify it as such, and to extract the information they are seeking from it. Machine Translation (MT) or content gisting in context, based on bilingual machine-readable dictionaries, has been investigated as a means of accessing particular information or at least determining whether a document is relevant (Oard and Resnik 1999)(He et al. 2003)

MLIR is potentially a key technology in supporting the localisation of dynamic, or rapidly published content. MLIR can aid localisation by making content available across languages in response to user search queries and potentially by providing translation between queries and documents. It can also support the presentation of retrieved content in a culturally sensitive way.The next section summaries some of the key research in the area of CLIR.

3.1 Cross-Language Information Retrieval

In CLIR there is a linguistic mismatch between the queries that are submitted and the documents that are retrieved. To resolve this mismatch CLIR systems incorporate some facility for content translation to bridge this language barrier, an obvious requirement if query representations and document representations are to be meaningfully compared. The performance of a CLIR system is heavily reliant upon the success of this translation process, and therefore the tools and techniques used for automatic translation have formed much of the focus of the CLIR research community.

One of the main questions that arises when addressing a CLIR task is whether to translate the queries to the language of the documents, or the documents to the language of the queries. There are pros and cons with each approach. For example, translation of documents may be more reliable since there is extensive contextual information available. However, query translation does not increase the content storage overhead. Extensive experiments have been carried out comparing document translation and query translation, and the combination of both (McCarley 1999)(Oard 1998)(Oard and Hackett 1997). While document translation and combination methods can outperform query translation, these results have generally been set aside due to the prohibitive effort required to translate the complete document collection into the query languages to be supported and the storage of the index data associated with them. Thus, the overwhelming majority of CLIR systems today operate via query translation.

Typically, three types of resources are widely adopted for translation in CLIR: bilingual wordlists (or machine readable dictionaries) (Adriani 2000)(Gao 2002)(Liu 2005)(Zhou 2008), parallel texts (Chen and Nie 2000)(Nie 1999), and machine translation (MT) systems (Kwok and Dinstl 2007)(Wu 2007). Query translation has most often been effected using machine readable bilingual dictionaries. Unfortunately, bilingual dictionaries have an inherent tendency towards ambiguity. This problem stems from the choice of possible translations. A typical bilingual dictionary will provide a set of alternative translations for each term within any given query. Choosing the correct translation of each term is a difficult task, and one that can seriously impact the efficiency of any related retrieval functions. Disambiguation problems can be reduced by using phrase level translations (Ballesteros and Croft 1998), unfortunately it is difficult to develop high coverage phrase translation dictionaries. Parallel texts offer a value translation resource, unfortunately in most cases there is no parallel content available from which to establish a parallel translation resource for search queries. The final option of using MT can work effectively. However, MT systems are generally developed to work with well structured text. User queries are typically unstructured strings of search words and phrases, which can create problems for MT based translation. A further problem is that MT systems only exist for a limited number of language pairs, and it is extremely expensive to develop an MT system for a new language pair. A significant problem which can arise for all of these translation methods is the coverage of the translation dictionaries. If the user enters queries which use words or phrasal expressions outside the translation dictionaries, they will not be translated accurately. Errors in translation arising due to ambiguity, linguistic structure or dictionary coverage are the main source of degradation in retrieval effectiveness for CLIR. Methods are currently being investigated which allow general domain translation resources to be augmented with domain specific bilingual dictionaries automatically extracted from resources such as Wikipedia (Jones et al. 2008).

4 Merging CLIR and AH

While AH and IR share many objectives in satisfying user information needs, to date they have developed almost entirely in isolation. Consequently they largely use different technologies, and have different strengths and weaknesses. One of the key research themes within the DCM track of the CNGL is the hybridization of AH and IR technologies to address user information needs more effectively. For example, statistical methods used in IR typically return a set of ranked documents, however for some applications this may not be the most user-friendly method of accessing and presenting data, nor is the selection of a complete document necessarily a suitable answer for some information needs. On the other hand, the substantial need for metadata associated with AH techniques, as well as the need for content to be structured in a particular way, means that content needs to be authored manually for a specific system. Thus we are interested in exploring the introduction of personalisation features into IR, and to introduce IR techniques to AH, with the objective of satisfying a wider variety of users and their different information need types, over a variety of languages without requiring the addition of large amounts of manual metadata.

The first mechanism that we are investigating is the combination of personalisation and IR to improve multilingual query expansion. Using structures such as a domain model and user model, the system can make determinations about the subject domain of interest for a particular user and their queries. The presentation and selection of particular results from a personalised, expanded query can also be altered using AH techniques. For example, personalisation can be used to re-rank results, for example for previously unseen results, or for results in a particular format. Currently, the activity in this area being undertaken by the DCM track of CNGL relates to eliciting user models statistically, in order to drive improved personalised, intelligent response generation.

The second approach, which is detailed below, uses IR techniques to 'crawl' the WWW and selected digital content repositories to generate collections of content in particular subject domains. This open corpus content model can be used to make large quantities and varieties of web-sourced content more accessible for incorporation in adaptively composed presentations. Several substantial challenges remain, and the following sections outline some of the research which has been undertaken in the area of transforming unstructured data into AH content.

4.1 Open Corpus Content

Enormous volumes of content, which is varied in structure, language, presentation style etc., and is suitable for inclusion in AH presentations can be sourced via the WWW. This "open corpus" content, can be defined as any content that is freely available for non-commercial use by the general public or educational institutions. Such content can be sourced from web pages, scholarly research papers, digital content repositories, forums, blogs, etc.. While some AH systems, such as KBS Hyperbook (Henze 2000) and SIGUE (Carmona 2002), allow the manual incorporation of individual web-based content resources, the scale of open corpus content available is yet to be comprehensively exploited by AH.

In order to facilitate the large-scale utilisation of open corpus content in AH, methods of surmounting the heterogeneity of web-based content must be developed. This includes integrated means of content discovery, classification, harvesting, indexing and incorporation. The Open Corpus Content Service (OCCS) (Lawless et al. 2008) is an IR tool chain which has been developed in to address these challenges.

The process of content harvesting for AH is infeasible at runtime, since it requires a considerable amount of resources in order to discover and index the large volumes of data available via the WWW for a particular domain. A persistent document cache is therefore created, in order to ensure reliable content candidate selection during the creation of an AH composition. This cache ensures that: the content is permanently available, the content accurately reflects the index representation of the resource and that further content preparation can be conducted on the resources before incorporation into an AH composition.

Focused web crawling enables the discovery of content which meets pre-determined classifications from across the WWW. The OCCS applies focused crawling techniques to traverse the WWW and centrally collate open corpus content resources, categorised by subject domain, for use in AH compositions. The OCCS combines an open source web crawler called Heritrix [Heritrix] with an open source text classification library called Rainbow [Rainbow] to conduct focused crawls where discovered content is compared to a statistical model of a subject area to estimate relevancy.

Once content has been discovered and harvested it must also be indexed to make it more readily discoverable during the content candidacy process. There are numerous open-source content indexing solutions available, such as Lemur [Lemur] and Lucene [Lucene]. Some indexing tools such as Nutch [Nutch] and Swish-e [Swish-e] have also been integrated with a web crawler to form

openly available information retrieval tool chains. However, these tool chains typically utilise general purpose, rather than focused, crawling techniques and can be limited by the indexing methods employed. The OCCS combines its focused crawling functionality with an open source indexing tool called NutchWAX [NutchWAX] to implement web-scale subject-specific content discovery and indexing. NutchWAX enables the indexing and free text search of web archives, or collections of web-based content.

4.2 On-Demand Slice provision from Subject-specific Caches

The OCCS delivers an integrated means of content discovery, classification, harvesting and indexing which can be leveraged by AH systems. However, there remain several challenges associated with the incorporation of crawled content in an AH composition. The first and most obvious is that conventional assumptions regarding the granularity, format and presentation style of the content available to an AH system can no longer be made. AH systems have traditionally operated upon closed sets of content resources, where the system is aware of each individual resource, its format and characteristics and any relationships between resources in advance. When attempting to incorporate open corpus content, the AH system can no longer assume it possesses this detailed knowledge.

There tends to be an inverse relationship between the potential reusability of a resource and its granularity. The larger and more complex a resource, the more contextually specific it is and the more difficult reuse becomes. Fine-grained, conceptually atomic, resources are much more easily reused outside of the context for which they were created.

The OCCS delivers resources which have been harvested from the WWW and still contain contextually-specific information such as navigation bars, advertisements, banner images etc.. These resources can also be large, course-grained pages of text. To address the challenge of improving the reusability of the resources delivered by the OCCS, research is underway into the development of a framework which can serve requests for specific content from AH systems with precise resources which have been stripped of ancillary information (Levacher et al. 2009). This would allow the AH system to request resources which conform to specific granularity, format and presentation style requirements. This would solve some of the problems with seamlessly incorporating open corpus content into AH compositions.

This framework, which implements content analysis and on-demand resource generation, is outlined in Figure 2. Each component of the framework executes a specific task on the open corpus content. These tasks include: the structural disaggregation of the content to remove presentation-specific content and other extraneous content; a statistical analysis of the content to map the concepts detailed in different parts of the resource; and the on-demand resource provider, which fulfils requests from the AH system for specific pieces of content.

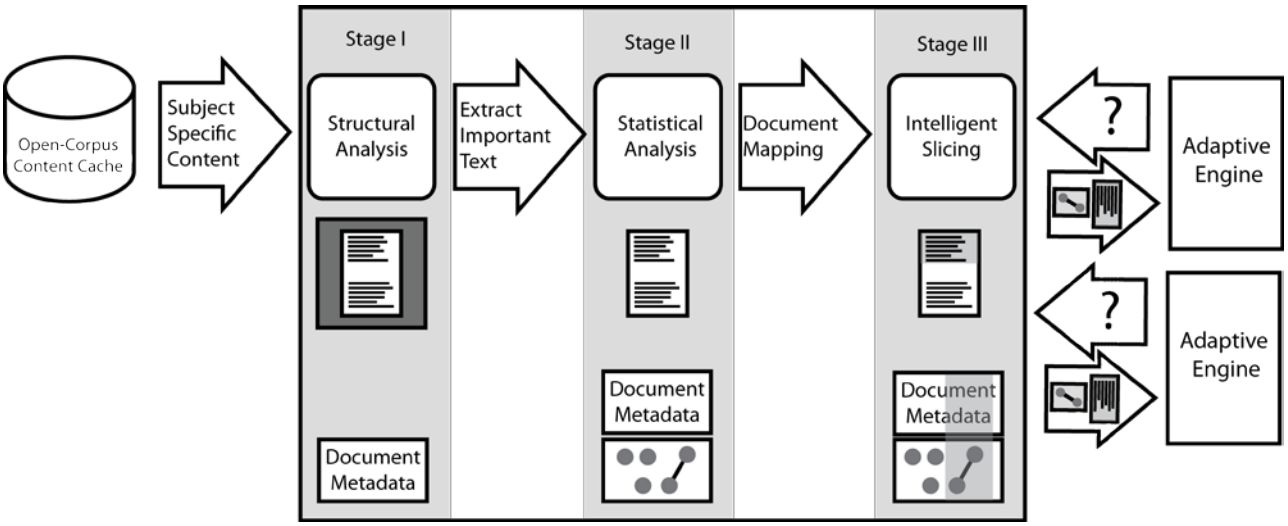

Figure 2: On-Demand Slice Generation Framework. The content to be analysed is drawn from an OCCS cache and is structurally and statistically analysed to extract the informative content and a conceptual map of that content.

The resource delivered to the AH system by this framework is referred to as a 'slice'. A slice is an abstract notion representing a conceptually atomic piece of information, originally part of an existing resource, which has been extracted to fulfil a specific information requirement. A slice can potentially be a composition of other slices from a number of resources. Appropriate metadata is also generated for the slice based upon the metadata of the parent resource(s). A slice is virtual in the sense that it only represents a subjective perspective of a particular resource and its description. The degree of complexity of a slice will match the requirements of the AH system which has requested it.

The content analysis stages of the framework are executed *a-priori* while the intelligent slicing task is executed at run-time. Each *a-priori* framework component generates a specific layer of metadata based upon the tasks and content analysis it has conducted. This metadata augmentation enriches the resource with any structural and semantic information which has been identified during the content analysis.

The first content analysis stage is structural segmentation, which is used to remove presentation-specific content and other extraneous content. Structural segmentation techniques include densitometric analysis (Kohlschütter 2009), DOM tree pattern analysis (Vieira 2006), isotonic regression (Chakrabarti 2007), vision-based techniques (Baluja 2006) (Cai 2006) and token-based approaches (Pasternack 2009). Once the structural segmentation has been performed, a statistical analysis of the resulting content is performed to create a conceptual map of the content. Statistical analysis approaches include the use of supervised learning techniques such as hidden markov models, dictionary and rule-based approaches and word-sense clustering. These stages allow the framework to reduce the resource down to its informative content, and identify what concepts are dealt with by the resource, and at what points in the content.

This framework represents a novel approach to content candidate provision in AH system, as it does not attempt to identify the required conceptual coverage, granularity or format of a resource in advance of presentation composition. Instead, the framework constructs a custom resource to fulfil the specific requirements of a request provided at runtime by the AH systems. The Framework architecture is currently under implementation within the DCM track of CNGL, with particular focus on the selection of structural analysis strategies.

5 Applications to Localisation

The localisation process can be viewed as a complex workflow of participants with different roles. Beginning with the content creator, the content itself must be translated into a target language and altered to conform to suitable cultural norms for a particular locale. The DCM track of CNGL aims to develop techniques which can be used to support a wide variety of localisation roles, beginning with the original author of the content, progressing through the translator who transposes the content to a new language, and ultimately to the reader who consumes the content.

There is considerable room for innovation in this area, both in the direct use of AH and IR techniques to improve content retrieval and presentation for the user, and in the use of these technologies to create greatly improved tool support for the participants in the process of generating localised content. This innovation can be viewed as affecting each of the different key roles in the localisation workflow, from the content creator to the localisation manager, Each role shares some concerns and has individual problems which can be addressed with improved digital content management.

5.1 In Support of the Content Creator

Translation reuse appears to be a common component of the overall goal of improving localisation effectiveness and efficiency. One approach to this goal is to make use of fuzzy string matching to retrieve previously-executed translations from a Translation Memory. In interviews with localisation service providers, one of the key challenges to this reuse is the tendency for content creators to seek to generate original content, which varies from writing guidelines.

There are two ways which DCM research can support improved content authoring. The first is to help in the generation of digital content, and the second is in creating tools which help authors to better comply with writing guidelines by making them more accessible.

Open-corpus AH techniques could be used to create subject-specific caches which contain information either from the open web, or from designated content sources. This can be applied to the creation of digital media through the reuse of ‘slices’, which can be re-composed in the formation of new documents, as described in section 4.2. Instead of a content author having to generate a completely new document, or substantially edit an existing document, it is potentially possible for an on-demand slicing service to create customised pieces of existing translatable content for use within a new document. There are several advantages to this approach over the manual authoring and merging of content, as the best content from the documents chosen over the whole corpus can be used, with the author able to concentrate on their combination rather than the laborious process of searching and separating slices manually.

The content which can be chosen to create the appropriate slices can be selected and prepared in advance, ensuring that the new document contains only material which is appropriate. Conventional document merging is a difficult and labour-intensive process, because there is a need to navigate and retrieve the correct content, then read large amounts of content to select appropriate parts for combination. This can make determining the provenance of pieces and the consistency of the document text more irregular. On-demand slicing can avert this difficulty by recording the transformations applied to a particular piece of a document during the content preparation workflow as well as its origin. The use of adaptive techniques to help select and compose the slices means that an AH system backed by an on-demand slicing service can provide content authors with a coherent skeleton of a document, rather than a collection of unrelated fragments.

By improving reuse at the document generation level, it is intended that the new documents which are created during this process will be more amenable to translation reuse, and will better comply with content creation guidelines because the constituent content is chosen from the corpus of documentation which best suits the localisation process.

The second method for supporting content creators is to make it easier for them to access and understand the policies which goven their authoring. As discussed above, AH systems can be used to select content which is most relevant to a particular user model. These techniques can be used to help content authors by tracking the documents which they are working on, and presenting only the relevant portions of particular policy documents. Familiar policy elements can be summarised, while new or particularly important translation and localisation guidelines and policies can be highlighted on a personalised basis.

This allows the author to have constant access to a tailored presentation of the translation and localisation guidelines and policies which they need to be aware of for a particular document, which is presented in a style which suits their preferences, and which remembers what they are familiar with. This is intended to help authors both by reminding them automatically of relevant policies, and also by reducing the overhead in switching between different documents, which might be governed by different policies. AH has been shown to be effective in the classroom learning situation (Conlan 2004) and the intention is to evaluate its effectiveness in this domain.

The effectiveness of these approaches will depend on several factors. The first, and most important, is that there is a need for the tools which are intended to support content authors to be effective in the generation of content which communicates effectively and can be translated at lower cost. There is a need to ensure that content preparation frameworks do in fact make the creation of content more effective, and that the adaptive documentation tools are in fact useful and deliver relevant instructions. The only way to measure this is with user trials, and the intention of our work in CNGL is to evaluate DCM research outputs in industrially-relevant scenarios with industry professionals. A combination of evaluation metrics will be employed that include industrial and academic concerns.

5.2 In support of the Translator

A human translator has several roles in the treatment of content. The first, and most obvious, is the creation of new translations for content. Depending on the translator's familiarity with the content, and their domain expertise, there can be substantial skill required in correctly translating the meaning of content in the correct sense.

The first application of DCM research in this area is the use of CLIR to help retrieve background material for translators from the open web. CLIR is effective in that it will return results of documents in a variety of languages, which can help the translator by providing them with similar texts in the original and target languages. This can help the translator to improve the quality of their translations by giving them access to background material on a particular topic in all the languages that they need.

The second major task of the translator which can be supported by DCM is in the creation of adaptive courses which help with guidelines, in the same way as to support content authors. These adaptive presentations are not limited to the rules and guides governing the translation and localisation process, but can also present relevant portions of terminological dictionaries, and can even make some simple semantic inferences from the domain models to help choose terminological hints specific to the translator, and the content which they are working on.

5.3 In support of the Reader

One of the key objectives of DCM research is to improve the presentation of content to the end user. In particular, adaptive presentation of information improves the ability of users to explore content, and encourages them to examine background material and related topics (Steichen 2009). Adaptive presentations based on a personal information need are of most benefit when there is a wide selection of candidate content appropriate to supporting different personalised requirements.

Traditionally, where 'personalisation' has simply meant offering users manual choices of language and locale, the content to be presented to the user has been prepared on a mass basis into specific culture/language combinations. As personalisation evolves to adaptivity, a more refined notion of how content can vary for individuals emerges.

Supporting fully personalised content in Localisation will likely require some changes to the assumptions about how content is transformed and managed during the localisation process. The composition of content pieces could allow for a more effective management of localisable material by allowing translation resources to be concentrated on the pieces of content which are most used in the compositions, while less important content can be relegated to machine translation, which can be less expensive.

A key advantage to the fact that personalisation systems can record a model of the user's behaviour and preferences is that this model can itself be reused. For example, as a user passes between different managed content sites, the user model can be allowed to propagate across these sites, cross-influencing each site which was independent up until that point. The notion of non-invasive adaptation means that personalisation in this case can have a subtle, but nonetheless powerful effect on the way content is presented, and the ease of the user's navigation (Koidl 2009).

Much of our work is at the early stages of planning and design. As the technologies mature, it is our belief that working with localisation professionals in the field will yield technological progress and mutually beneficial results through improved digital content management.

6 Conclusions

This paper has presented a brief introduction to the area of DCM research, through the specific subjects of AH and IR. Particular focus has been placed on multilingual, open-corpus techniques, as these are likely to be the best suited to the multi-cultural WWW of the future.The application of these technologies to localisation, and their evaluation within that domain is intended to create real improvements in the way content is prepared for presentation to the user.

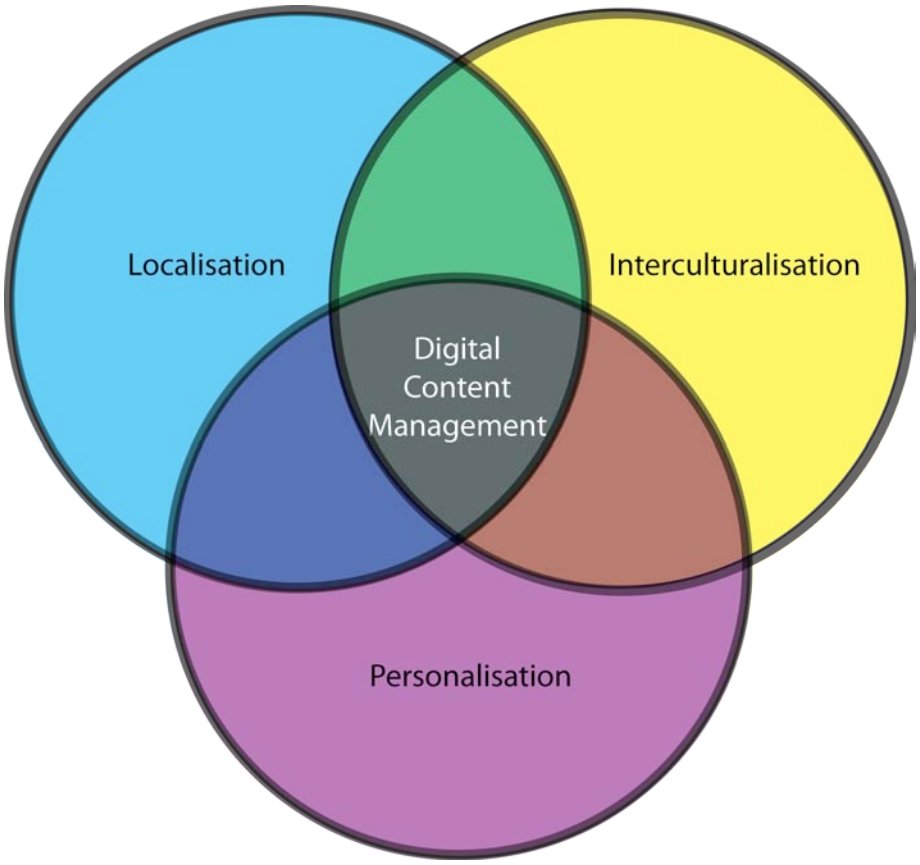

Figure 3: Locale, Culture and Personal Preferences combine to influence the future of DCM

Improvements in multilingual digital content management will result from a better understanding of the key influences which make for better answers to user's information needs. The first influence is the localisation of the document to the appropriate language. This is a fundamental requirement; it is difficult to envision a user being able to make use of a document which is not in the appropriate language. Translation of content can come in a variety of forms, and for a variety of costs, so the goal of reuse and targeting of effort seen in AH is of similar if not greater proportions in the localisation area.

The second influence of cultural appropriateness is deeply interlinked with the separation of presentation and content. There is a key need to be able to communicate content to the user with attractive and appropriate cultural norms, and once again, compatibility with the objective of DCM technologies, which seek to abstract the content of a document from surrounding extraneous features of a page.

The influence of personalisation is apparent, and has not until now been addressed in localisation. There are many advantages to being able to create specific content for individuals, and it allows targeting of effort across the content management workflow. The analytical aspects of DCM technologies also allows for statistics on the effectiveness and usefulness of particular content to be gathered, which can be fed back into the content preparation and analysis process for improved results.

Finally, as a rich media environment with clear performance metrics and an inherently multilingual approach, the localisation industry itself has the potential to benefit substantially from the addition of DCM techniques to improve the retrieval and presentation of the media that supports the process.

References

Adriani, M. “Using statistical term similarity for sense disambiguationin cross- language information retrieval”. *Inf. Retr.*, 2(1):71-82, 2000.

Ballesteros, L. & Croft, W.B. (1998) “Resolving ambiguity for cross-language retrieval”, In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in Information Retrieval*, pp 64-71, Melbourne, Australia

Baluja, S. (2006) “*Browsing on Small Screens: Recasting Web Page Segmentation into an Efficient Machine Learning Framework*”. In the Proceedings of the 15th International World Wide Web Conference, WWW2006, pp. 33-42, Edinburgh, Scotland.

Brin, S. and Page, L. “The anatomy of a large-scale hypertextual web search engine”. *Comput. Netw. ISDN Syst.*, 30(1-7):107-117, 1998.

Brusilovsky, P. (1996) ‘Methods and Techniques of Adaptive Hypermedia’,*User Modeling and User Adapted Interaction.* 6 (2-3) : 87-129

Brusilovsky, P. (2004), ‘Adaptive navigation support: From adaptive hypermedia to the adaptive web and beyond’, *Psychology Journal* 2, 7–23

Cai, D., Shipeng, Y., Wen, J.R. & Ma, W.Y. (2006) “Extracting Content Structure for Web Pages based on Visual Representation”. In the *Proceedings of the 5th Asia Pacific Web Conference*, APWeb 2003, pp. 406-417, Xi‘an, China. 23rd-25th April, 2003.

Carmona, C., Bueno, D., Guzmán, E., Conejo, R. (2002) “SIGUE: Making Web Courses Adaptive”. In *Proceedings of 2nd International Conference on Adaptive Hypermedia and Adaptive Web Based Systems*, AH2002, Malaga, Spain, 29-31 May, 2002. Lecture Notes on Computer Science, Vol. 2347. Berlin: Springer Verlag, pp. 376-379. 2002.

Chakrabarti, D., Kumar, R. & Punera, K. (2007) “Page-level Template Detection via Isotonic Smoothing”. In the *Proceedings of the 16th International World Wide Web Conference*, pp. 61-70, Banff, Alberta, Canada. May 8th-12th, 2007.

Chen, J. and Nie, J.-Y. “Parallel web text mining for cross-language IR”. In *Proceedings of RIAO-2000: Content-Based Multimedia Information Access*, pages 188-192, CollCge de France, Paris, France, 2000.

Conlan, O. and Wade, V. (2004) Evaluation of APeLS – An Adaptive eLearning Service based on the Multi-Model Metadata-Driven Approach. In *Proceedings of the 3rd ACM International Convference on Adaptive Hypermedia and Adaptive Web Systems* pp 192-195, Eindhoven

De Bra, P., Brusilovsky, P. and Houben, G.-T. (1999) ‘Adaptive Hypermedia, From Systems to Framework’ *ACM Computing Surveys*, 31(4),

De Bra, P. and Calci, L. (1997) Creating Adaptive Hyperdocuments for and on the Web. In *Proceedings of the AACE WebNext Conference*, Toronot, pp. 149-155

Gao, J., Zhou, M., Nie, J.-Y., He, H. & Chen, W. “Resolving query translation ambiguity using a decaying co-occurrence model and syntactic dependence relations”. In *Proceedings of the 25th*

Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 183-190, Tampere, Finland, 2002. ACM Press.

He, D., Oard, D. W., Wang, J., Luo, J., Demner- Fushman, D., Darwish, K., Resnik, P., Khudanpur, S., Nossal, M., Subotin, M. & Leuski, A. (2003) "Making MIRACLEs: Interactive Translingual Search for Cebuano and Hindi," *ACM Transactions on Asian Language Information Processing* 2(3):219-244

Henze, N. and Nejdl, W. (2000) "Extendible Adaptive Hypermedia Courseware: Integrating Different Courses and Web Material". In the Proceedings of the International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems, AH2000, pp. 109-120, Berlin: Springer-Verlag, Trento, Italy. August 28th-30th, 2000.

Jones, G.J.F., Fantino, F.,Newman E. & Zhang, Y (2008) "Domain-Specific Query Translation for Multilingual Information Access Using Machine Translation Augmented With Dictionaries Mined From Wikipedia", In *Proceedings of the 2nd International Workshop on Cross Lingual Information Access - Addressing the Information Need of Multilingual Societies (CLIA-2008)*, Hydrabad, India, pp34-41.

Kleinberg, J.M. "Authoritative sources in a hyperlinked environment". *J. ACM*, 46(5):604-632, 1999.

Kohlschütter, C. & Nejdl, W. (2009)"A Densitometric Approach to Web Page Segmentation". *In the Proceedings of the 17th ACM Conference on Information and Knowledge Managemen*t, CIKM 08, pp. 1173-1182, Napa Valley, California, USA. October 26th-30th, 2008.

Koidl K., Conlan O., Wade V. (2009) Non-Invasive Adaptation Service for Web-based Content Management Systems (DAH2009), International Workshop on Dynamic and Adaptive Hypertext: Generic Frameworks, Approaches and Techniques, Torino, Italy.

Kwok, K.L. & Dinstl., N. "NTCIR-6 monolingual Chinese and English-Chinese cross-language retrieval experiments using pircs". In *the Sixth NTCIR Workshop Meeting*, pages 190-197, NII, Tokyo, Japan, 2007.

Lawless, S., Hederman, L., Wade, V. (2008) "OCCS: Enabling the Dynamic Discovery, Harvesting and Delivery of Educational Content from Open Corpus Sources". In the Proceedings of the Eighth IEEE International Conference on Advanced Learning Technologies, I-CALT 2008, Santander, Spain. 1st-5th July, 2008.

Levacher, K., Hynes, E., Lawless, S., O'Connor, A., Wade, V. (2009) A Framework for Content Preparation to Support Open Corpus Adaptive Hypermedia In *Proceedings of International Workshop on Dynamic and Adaptive Hypertext: Generic Frameworks, Approaches and Techniques*, Torino, Italy.

Liu, Y., Jin, R. & Chai, J. Y. "A maximum coherence model for dictionary-based cross-language information retrieval". In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 536-543, Salvador, Brazil, 2005. ACM Press.

McCarley, J.S. "Should we translate the documents or the queries in cross-language information retrieval?" In *Proceedings of the 37th Annual Meeting of the Association for Computational*

Linguistics on Computational Linguistics, pages 208-214, College Park, Maryland, 1999. Association for Computational Linguistics.

Nie, J.-Y., Simard, M., Isabelle, P. & Durand, R. “Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the web”. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 74-81, Berkeley, California, United States, 1999. ACM Press.

Oard, D. W. and Resnik, P. (1999) "Support for Interactive Searching in Cross-Language Information Retrieval," *Information Processing and Management* 35(3): pp 363-379.

Pasternack, J. and Roth, D. (2009) “Extracting Article Text from the Web with Maximum Subsequence Segmentation”. In the Proceedings of the 18th International World Wide Web Conference, WWW2009, pp. 971-980, Madrid, Spain. April 20th-24th, 2009.

Oard, D.W. “A comparative study of query and document translation for cross-language information retrieval”. In *Proceedings of the Third Conference of the Association for Machine Translation in the Americas on Machine Translation and the Information Soup*, pages 472-483. Springer-Verlag, 1998.

Oard, D.W. & Hackett, P. “Document translation for cross-langauge text retrieval at the university of Maryland”. In *The Sixth Text Retrieval Conference (TREC-6)*, pages 687-696, NIST, 1997.

Salton, G. & Bucklley, C. (1988) “Term weighting approaches in automatic text retrieval” Information Prcocessing & Management, 24(5):513-523.

Spärck Jones K Walker S. & Robertson S. E. (2000) “A probabilistic model of information retrieval: development and comparative experiments Part 1”. *Information Processing & Management* 36(6):779-808.

Spärck Jones K Walker S. & Robertson S. E. (2000) “A probabilistic model of information retrieval: development and comparative experiments Part 2”. *Information Processing & Management* 36(6):809-840.

Steichen, B., Lawless, S., O'Connor, A., and Wade, V. (2009). Dynamic hypertext generation for reusing open corpus content. *In Proceedings of the 20th ACM Conference on Hypertext and Hypermedia* (Torino, Italy, June 29 - July 01, 2009). HT '09. ACM, New York, NY, 119-128

Vieira, K., da Silva, A., Pinto, N., de Moura, E., Cavalcanti, J. & Freire, J. (2006) “A Fast and Robust Method for Web Page Template Detection and Removal”. *In the Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, CIKM 06, Arlington, Virginia, USA. November 6th-11th, 2006.

Wray, R. (2009) ‘Internet data heads for 500bn gigabytes’, *The Guardian*, accessed July, 2009 http://www.guardian.co.uk/business/2009/may/18/digital-content-expansion

Wu, Y.C., Tsai, K.C. & Yang, J.C. “NCU in bilingual information retrieval experiments at NTCIR-6”. In *the Sixth NTCIR Workshop Meeting*, pages 133-139, NII, Tokyo, Japan, 2007.

Zhou, D., Truran, M., Brailsford, T., & Ashman, H. *“A Hybrid Technique for English-Chinese Cross Language Information Retrieval”*, in ACM Transactions on Asian Language Information Processing (TALIP) 7, 2, June 2008.

Footnotes

[Heritrix] Heritrix is the Internet Archive's open-source, extensible, web-scale, archival-quality web crawler project. Available online at: http://crawler.archive.org/

[Rainbow] Rainbow is a program that performs statistical text categorisation. Available online at: http://www.cs.cmu.edu/

[Lemur] The Lemur Toolkit is an open-source suite of tools designed to facilitate research in language modeling and information retrieval. Available online at: http://www.lemurproject.org

[Lucene] Apache Lucene is a full-featured text search engine library written entirely in Java. Available online at: http://lucene.apache.org/java/docs/index.html

[Nutch] Nutch is an open source web-search solution based upon Lucene. Available online at: http://lucene.apache.org/nutch

[Swish-e] Simple Web Indexing System for Humans – Enhanced (Swish-e), a flexible and free open source system for indexing collections of Web pages. Available online at http://www.swish-e.org

[NutchWAX] Nutch and Web Archive eXtensions is a tool for indexing and searching web archive collections. Available online at: http://archive-access.sourceforge.net/projects/nutch/

[CLEF] Cross-Language Evaluation Forum at: http://www.clef-campaign.org/