Open Corpus Learning Content; Harvesting Knowledge to provide Equitable Access to Education for All

Inequitable Access to Education and Technology in a Knowledge Economy

Séamus Lawless

Knowledge and Data Engineering Group, Trinity College Dublin, Ireland

e-mail: slawless@cs.tcd.ie

Abstract – For generations education and educational institutions were closed books, available only to those privileged enough to attend these institutions. However the internet has unlocked the door to knowledge for countless individuals and communities around the world. With recent advances in internet technologies and social interactions online, eLearning is fast becoming an integral part of everyday life. However structuring learning experiences and tailoring them to the individual remains a huge challenge. eLearning systems that are capable of adapting a learning experience with regard to localization, cultural diversity and personalisation remain tied to educational institutions as they are heavily reliant upon specifically authored educational content. However, the possibility now exists to make education and eLearning a reality for all the global community. By leveraging the vast amounts of open corpus learning content available in digital repositories, corporate training repositories, whitepapers, the Worldwide Web, etc., educational experiences could be generated, tailored and delivered to users around the world. The ability to harness this resource would help to redress the imbalance that exists in access to education and knowledge availability worldwide.

Introduction

Knowledge has always been at the heart of Humankind, from the entrustment of traditions within a community, to the teaching of life skills to the young. Knowledge and access to knowledge has had a profound affect on shaping the social structure of our world. The countries, communities and individuals who have control over access to knowledge have long been the privileged, wealthy and powerful. This inequality in the distribution of knowledge, and the ability of communities to educate themselves, has produced vast social and economic divides in the global community. The ability to attain knowledge is essential to all human endeavours be they personal, social or economic.

The dawn of the Internet has inadvertently created a whole new knowledge economy. We have an opportunity for the first time, to create open access to education and information. No longer is attendance at a higher-level educational institution the sole method to acquire knowledge on a particular subject or area. No longer are the levels of personal development an individual can achieve dictated by their standing in society or social status. Knowledge is fundamentally different from physical goods and services, information resources can be copied, can be shared and have proliferated across the globe. There is no restriction on how many times a piece of information can be shared and reused, no constraint on how many people can use information to educate themselves on a topic.

The continued rapid development of internet technologies and the recent advances in online social interactions have meant that electronic learning experiences, or eLearning, has fast become an integral part of everyday life. From searching the internet for specific resolutions to problems or information on particular topics of interest to formal structured learning courses, the education possibilities and the levels of information available online are immense.

However the availability of formal structured learning experiences is still limited to the arena of the educational institution. eLearning systems, and in particular adaptive eLearning systems that can tailor the educational experience to the individual are heavily reliant on bespoke content. This content has to be authored within the educational institution and made available to the eLearning system. This not only continues to place the strain of content creation on the pedagogues, it also restricts access to this knowledge for the global community outside the realm of formal educational institutions. The imbalance of access to formal educational development and knowledge acquisition thus remains in this instance.

However, the opportunity to redress this imbalance and make formal education a reality for the entire online community does exist. The internet is a vast archive of information and knowledge resources on almost every subject imaginable. Knowledge resides in web pages, scholarly papers, digital content repositories, commercial training repositories, forums, blogs, etc. the list of resources is almost endless. The latent information contained within these online resources can be identified and leveraged for use in formal elearning offerings. The ability to identify, harvest and deliver open corpus learning content within formal elearning offerings produces a scenario whereby formal educational experiences can be delivered to the global community. These experiences can be adapted to account for localization, cultural diversity, personal preferences such as learning style and goals, and also to provide learning experiences for those with disabilities.

The vast array of knowledge on the internet is a legacy of all the individuals worldwide who have contributed to this new global community. Through this technological revolution, the spirit of cooperation, collaboration and community that pervades humankind has flourished and the dissemination and propagation of knowledge has prospered. The ability to utilise this knowledge, and use open corpus content to create educational experiences would help to redress the imbalance that still exists in access to education and knowledge availability worldwide.

eLearning and it's place in the World

eLearning is usually employed as a term to describe any computer-enhanced learning experience. Simply put, elearning is an online educational experience that is delivered in a synchronous tutor-driven structure or in an asynchronous learner-driven structure. In its infancy, eLearning was simply a means of providing access to information online, however next generation eLearning environments are responding to demands for greater levels of interaction and tailoring of learning experiences. Educational environments are attempting to satisfy these demands by supporting such functionality as personalisation, adaptivity and on-demand learning object generation [Brusilovsky, 01].

With the emergence of new web 2.0 technologies eLearning is shifting from being a passive activity to an interactive experience where the learner becomes more dynamically involved and engaged with each learning experience. Technologies emerging from the adaptive hypermedia (note. 1), semantic web (note. 2) and distributed computing (note. 4) communities are being widely employed in online learning. It is now common for learning experiences to be supplemented with quizzes and tests to assess each learner's progress. Forums are a common way of creating social communication and collaboration between individuals as they progress through a learning experience. Video, Audio and Interactive multimedia content is now an invaluable resource for use in the delivery of online learning experiences. The popularity of sites supporting folksonomies such as Flickr [Flickr] and YouTube [YouTube] are testament to this fact.

eLearning has rapidly become an integral part of the education process with almost all public higher education institutions delivering some course content online. This provides an opportunity to unlock access to education for many people around the world. It is no longer necessary to be one of the privileged few who can attend a higher education institution. Access to education via online eLearning is now a reality. More and more systems are becoming freely available that provide learning opportunities to the global community. This provides individuals and communities with the prospect of acquiring knowledge that can help them to improve their political, social, personal and professional development. Education can be provided to people as never before possible, which can aid the self-sufficiency of communities in the developing world.

Personalisation – Tailoring individual learning experiences

Next generation eLearning environments are delivering personalized "just-for-you" learning experiences [Dagger, 2005]. The benefit of such learning is that it can be dynamically tailored to the learner's goals, cultural background, language, prior experience, learning style and learning preferences. The experiences can also be tailored for the benefit of learners with disabilities, for instance the provision of audio for those with impaired vision. This empowers the individual as the learning experience and the learning activities are more suited to that particular student.

Contextually tailoring learning experiences for the individual is one of the most compelling developments to emerge from the field of eLearning, and is one of the grand challenges for the next generation of learning systems. It allows learning offerings to address individual learner's motivations and promote each learner's engagement with the educational process, which in turn promotes greater effectiveness, efficiency and student empowerment.

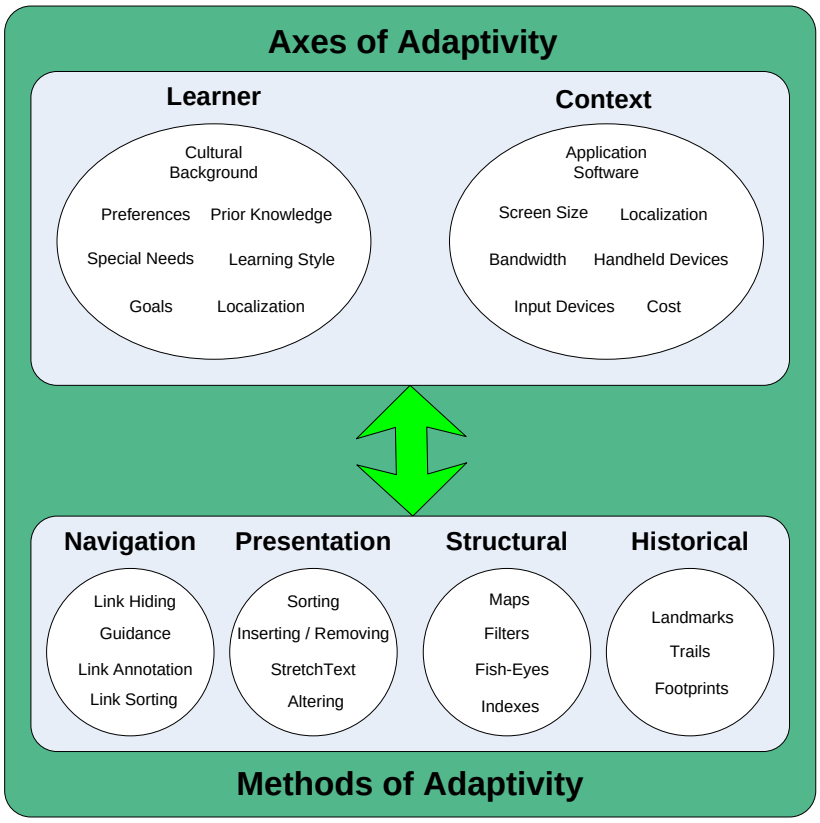

Fig.1. Axes and Methods of Adaptivity

A learner's attributes that may contribute to the personalisation of a learning offering and the aspects of context in that system are collectively known as the axes of adaptivity. The methods that may be used to personalize the learning offering based upon these axes are known as the methods of adaptivity (Figure 1). A learner's cultural background, localization information, learning style and preferences and any special needs that the learner may have can all be incorporated into the personalized learning experience [Conlan, 04].

One of the main problems with the current generation of personalized eLearning systems, is that they are reliant on bespoke proprietary content. This means that the ability of the system to use content authored externally is severely compromised. It also means that such eLearning systems are restricted to use within educational institutions where content can be authored to produce suitable learning offerings.

There is a move in the next generation of elearning systems to separate the system functionality and intelligence from content in the system. This creates a scenario whereby personalized eLearning systems will no longer be reliant on proprietary content. Leveraging open corpus content for use in such systems would mean that educational experiences could be constructed dynamically and delivered to users around the world, regardless of their location or social status.

Open Corpus Content

Open corpus content can be defined as any content that is freely available for use by the general public or educational institutions. Content can be sourced from web pages, scholarly research papers, digital content repositories, commercial training repositories, forums, blogs, etc. Many countries are now investing in national digital content repositories to encourage the reuse of learning resources. Merlot

[Merlot] in the United States of America and the National Digital Learning Repository [NDLR] in Ireland are just two examples of such repositories.

Open corpus content can exist in numerous formats. It can be wrapped and tagged as a structured learning object, digital resource/asset or an entire structured course. It can take the form of video, audio or graphical streamed data. It can be stored as documents, pdfs or slideshow presentations. Any piece of text, regardless of its granularity, from single paragraphs to entire textbooks, novels or plays can constitute learning content. The worldwide web is a vast warehouse of information that would be ideal for incorporation into learning experiences for people the world over. If this vast amount of learning content could be leveraged for use in personalised elearning systems, dynamic personalised elearning offerings could be generated for the global community.

Leveraging Open Corpus Content to deliver Education to the Global Community

To facilitate the use of open corpus content in eLearning environments, mechanisms for sourcing, harvesting and delivering this content need to be identified and made accessible through flexible services. There are several challenges that need to be addressed to successfully implement such services. In order to cope with the numerous sources, formats and attributes of open corpus content, any service that hopes to leverage such content will need to perform a series of operations to ensure accuracy and consistency in content identification and harvesting. A web crawler will traverse digital repositories and the Worldwide Web creating a metadata cache of sourced learning content.

Inconsistencies in both the semantics and structure of the metadata content descriptions applied to the content will need to be addressed. The variety of metadata standards applied to content on the Worldwide Web currently limits the interoperability of content and systems. Semantic inconsistencies in describing content make the identification of suitable content a difficult task. These interoperability issues effect the identification of suitable open corpus content for incorporation into structured learning experiences. A mapping from the current metadata standard applied to the content to a canonical metadata model is required to resolve this problem. A fixed vocabulary is then used to ensure consistency in the metadata descriptions of content.

In the case of some content discovered on the web there may be no associated metadata descriptions. Automatically generating descriptions of open corpus content that has no, or insufficient, metadata information will be a significant challenge in leveraging open corpus content for educational purposes. Thoroughness and consistency in this semantic information generation is essential to ensure accurate identification and retrieval of content. The number of tools emerging in the research community that facilitate the automatic or semi-automatic generation of semantic content descriptions is an indication of its importance in the progression and development of the Worldwide Web. Future technologies will rely on computers being able to automatically analyse and comprehend the information contained within web content.

Annotea [Annotea], developed by the W3C project, uses an RDF based annotation schema to describe annotations as metadata. Metasaur [Metasaur], developed by the University of Sydney, supports the creation of metadata for learning objects, using a standard vocabulary ontology that is generated automatically. This ontology can then have terms added to improve accuracy. This ensures consistency in the terms used to tag content. Semtag [Dill, 2003] is a system that has been developed using Seeker, which is a platform for large-scale text analytics. Semtag was created by the IBM Almaden Research Centre. The system crawls through web pages and performs an automated semantic tagging of each page using the TAP ontology. Semtag uses a Taxonomy Based Disambiguation (TBD) algorithm to ensure the correct classification of content in its tagging. IBM's LanguageWare [LanguageWare] is a text analytics tool than can be employed in the classification of content that has no associated descriptions. The tool analyses content for vocabulary found in an ontology to gain an understanding of the contents function, topic, structure etc.

When suitable content has been retrieved it needs to be delivered to the elearning system in a format that can be used during the generation of personalised eLearning offerings. To enable this, the content needs to be formally re-structured into a learning object (LO). Several research projects are focused on addressing this issue, for example, the iClass project [iClass] suggested the development of a component called the Learning Object Generator [Brady, 2005]. The open corpus content service

could provide appropriately grained and suitably tagged content to such a learning object generation environment which could then deliver the learning object back to the elearning system.

Issues such as digital rights management, intellectual property, security and ensuring the conceptual and aesthetic flow of learning offerings will also be major challenges in the implementation of a service that can deliver open corpus content to the global community. However, it is felt that these issues constitute major research projects individually and as a result are considered out of the scope of this current research and as such will not be addressed by this work.

An Open Corpus Content Service

It is proposed to provide an open corpus content sourcing function as an autonomous service that can replace the current method of content sourcing in the architecture of elearning environments. This will minimize the impact on the current architecture of such systems.

Content discovery is the first issue the system must address. A web crawler is implemented to navigate selected digital repositories and the Worldwide Web extracting content on a specific topic. This is referred to as a focused web crawl. The crawler traverses the internet following hyperlinks discovered on each website it encounters. As each page is identified it is downloaded and passed to a text classifier. The text classification tool [Rainbow] analyses the content to ascertain its subject matter using a list of key words and positive and negative training sets created in conjunction with the Google API [Google]. If the content is compatible with the focus of the crawl it is included in a subject specific cache of content.

Metadata descriptions are attached to each piece of learning content, however when dealing with content extracted from the Worldwide Web no assumptions can be made as the quality or quantity of these descriptions. In situations where the metadata descriptions attached to discovered content are insufficient or do not exist, automatic/semi-automatic analysis of the content and generation of metadata descriptions will be required to take place.

Existing metadata descriptions of content may be structured using conflicting or incompatible standards. Mappings to a canonical metadata model will be implemented using a fixed ontology of terms to standardise the metadata descriptions used to describe content stored in the cache. These mappings are required to ensure both accuracy and consistency in content description and the results of searches performed on the content index. Once this is complete an entry can be made in the candidate content cache for that content object.

The subject specific content cache is then parsed and catalogued by an indexing tool [NutchWax] to create a searchable content index. This index will be used to identify suitable learning content for inclusion in a specific learning offering. Semantic and Syntactic content requirements can be extracted from the eLearning systems used to generate the personalized eLearning designs. Information regarding the nature of the content required, both technical and semantic, can be extracted without manual intervention by the user. These requirements can then be manipulated by the open corpus content service, and structured into search queries. These search queries can then be used to search through the content index to identify suitable candidate content.

Once suitable content has been identified, it can be extracted from the content cache, and passed to either an LO Generation service or to the eLearning system itself, depending on the requirements for how the delivered content needs to be structured for incorporation into the learning offering in question. If passed to an external LO Generation Service it will be re-structured and sequenced into a learning object. This learning object is then passed back to the eLearning system. If content is passed directly back to the sourcing eLearning system, that system will be responsible for ensuring that conceptual and aesthetic flow of the learning offering is maintained.

This open corpus content service could be deployed as a service for a personalized eLearning system. Users from across the global community could use this eLearning system to generate structured learning experiences tailored to their own preferences and adapted to suit the cultural diversity and locality of individual. This ability would produce a massive step forward in providing equitable access to education and knowledge for people across all social and economic divides.

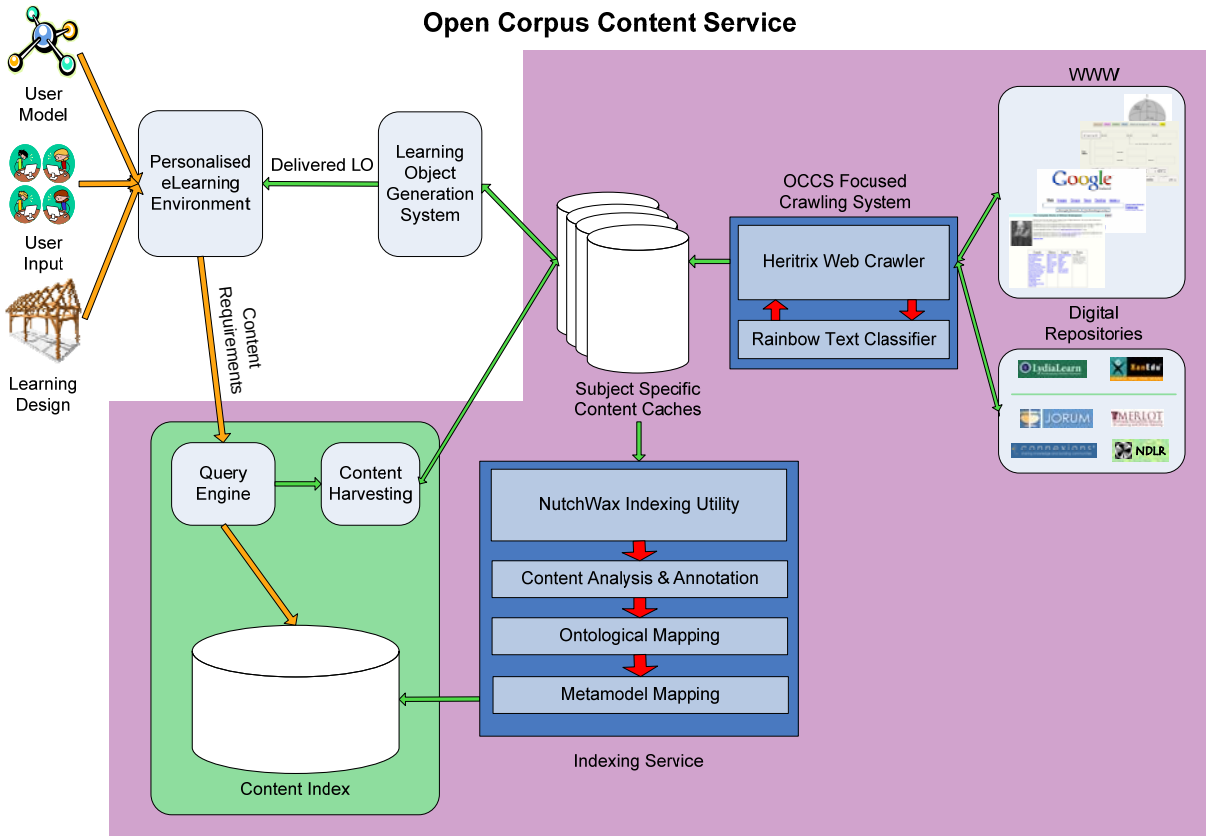

Fig.2. An Open Corpus Content Service

Conclusions

Equitable and open access to knowledge and education is paramount in ensuring the development of the global community across the boundaries created by national, social and economic divides. The internet has created, for the first time in history, the ability to share and replicate information with minimal cost across all communities worldwide. This gives life to a scenario whereby education is no longer reserved for those in positions of privilege, for those who can attend higher level education institutions in developed nations. For the first time, education can be delivered to all the global community, regardless of their physical location, social status or financial resources.

The internet is a vast warehouse of knowledge created by individuals and available to all. A huge cache of the worlds collective knowledge has been generated that, if leveraged, could provide the basis for educational experiences on almost every subject imaginable.

Emerging technologies have provided us with the opportunity to identify, analyse, harvest and manipulate this huge vat of knowledge. We can traverse the web using web crawlers and create a cache of content. Content can then be accurately described using semantic tagging tools and metadata mapping. We can categorise this content using text classifiers and indexers. This index can then be searched to identify learning content that matches a list of specific requirements.

Technologies in the field of eLearning and Adaptive Hypermedia have provided us with the ability to tailor learning offerings for the individual. Learning experiences can be adapted, using various adaptation methods, to account for localisation information, cultural diversity, individuals learning style, the learner's goals and preferences and much more.

When these technologies are combined with the vast array of open corpus content available it creates an exciting and powerful educational tool. Web services can allow individuals to create structured learning experiences that achieve specific educational goals. Communities of learning could evolve to support knowledge development in specific areas. We are provided with the ability to use this knowledge to create open and equitable access to education for all the global community.

Notes

Adaptive Hypermedia

Traditional hypermedia systems allow the learner to freely navigate between nodes by following links in an extensive network of information and knowledge. This unstructured style allows exploratory and inquiry-based learning with a high degree of student control. However there are problems with such systems. The potential for learner disorientation is great in such a large, unstructured knowledge base. This is often referred to as "lost in hyperspace". The systems also fail to accommodate users with varying interests, goals and needs. Adaptive hypermedia systems aim to provide solutions to these problems. Adaptive systems can incorporate domain knowledge and knowledge of the user. An adaptive system might provide a different body of information or present the information in a different fashion based on characteristics such as a user's learning style, language, geographical location or previous experience. Essentially adaptive hypermedia systems are all hypermedia systems which use some attributes of the user to adapt various visible and informational aspects of the system for that individual.

Semantic Web

The Semantic Web is a vision of a computer comprehensible Worldwide Web. Computers cannot accomplish many complex tasks on the web without human intervention as web pages are designed to be read by people, not machines. The semantic web is a vision of web pages that are understandable by computers. Currently, the Worldwide Web is based primarily on pages of information written in HTML. However, HTML as it is generally deployed, has limited ability to classify and describe the blocks of information on a page, apart from the specific roles they play in the page's organization and desired visual layout. In the vision of the Semantic Web, descriptive semantics are associated with all content stored on the web. The machine-readable descriptions enable content managers to add meaning to the content. If the semantic web becomes reality, it will be comprised of ‘intelligent machine agents’ that can accomplish complex tasks without human assistance.

Distributed Computing

Distributed computing focuses on the design of distributed, open and scalable systems. This area of research has evolved from the use of computers in the formation of networks. Distributed computing implements decentralised and parallel computing, using two or more computers communicating over a network to accomplish a common objective or task.

Acknowledgements

This research is funded by the Embark Initiative of the Irish Research Council for Science Engineering and Technology; funded by the National Development Plan.

References

Annotea, A W3C Lead Project, Available at http://www.annotea.org

Brady, A., Conlan, O., Wade, V. 2005. Towards the Dynamic Personalized Selection and Creation of Learning Objects. In *Proceedings of World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education, Vancouver, CA, 2005 E-Learn 2005*, 1903-1909

Brusilovsky, P. 2001. Adaptive Hypermedia. *User Modelling and User-Adapted Interaction*, Springer, 87-110.

Conlan, O., Wade, V. 2004. Evaluation of APeLS - An Adaptive eLearning Service based on the Multi-model, Metadata-driven Approach. In Nejdl, W., De Bra, P. (Eds.), *Adaptive Hypermedia and Adaptive Web-Based Systems: Third International Conference, AH 2004, Eindhoven, The Netherlands, August 23-26, 2004, Proceedings (Lecture Notes in Computer Science)*, 291-295

Dagger, D., Conlan, O., Wade, V., 2005. eLearning Without Borders - A Support Framework for Reusing Educational Strategies, *In Proceedings of Education Without Borders International Student Conference 2005, Abu Dhabi, UAE.*

Dill, S., Eiron, N., Gibson, D., Gruhl, D., Guha, R., Jhingran, A., Kanungo, T., Rajagopalan, S., Tomkins, A., Tomlin, J., Zien, J. 2003. SemTag and Seeker: Bootstrapping the semantic web via automated semantic annotation. In *12th International Conference on World Wide Web, Budapest, Hungary*, 178-186.

Flickr, a photo sharing product from Yahoo! that allows semantic tagging of images. Available at http://www.flickr.com

Google Search API, An API for applications to query Google using SOAP and WSDL standards. Available at http://code.google.com/apis/soapsearch/

iClass, Intelligent Distributed Cognitive-based Open Learning System for Schools (iClass), European Commission FP6 IST Project. http://www.iclass.info

LanguageWare, IBM LanguageWare Linguistic Platform for semantic analysis. Available at http://www-306.ibm.com/software/globalization/topics/languageware/index.jsp

Merlot, Multimedia Educational Resource for Learning and Online Teaching. Available at http://www.merlot.org

Metasaur, University of Sydney's Metasaur Project, Available at http://www.it.usyd.edu.au/~alum/demos/metasaur_hci/

NDLR, National Digital Learning Repository, Available at http://www.learningcontent.edu.ie

NutchWax, A bundling of the open source search engine Nutch and extensions that can be used to index web archives. Available at http://archive-access.sourceforge.net/projects/nutch/

Rainbow, A statistical text classification tool, developed at Carnegie Mellon University by Andrew McCallum. Available at http://www.cs.cmu.edu/~mccallum/bow/rainbow/

YouTube, a video sharing website that allows tagging and sharing of streamed media content. Available at http://www.youtube.com