

Enabling access to information and knowledge: A report on new tools – *Enable and Unify* – for conducting federated search of distributed content, extraction and management of metadata, and development of user-centered views of content

Steven Shaw
Chief Learning Officer
Eedo Knowledgeware Corp.
steven.shaw@eedo.com

Marie-Claude Lavoie
Educational Technology, Concordia University
marieclaude.lavoie@gmail.com

Abstract

Current thinking concerning individual and organizational performance questions the ability of conventional, formal training strategies to address gaps, given the volume and rate of change in modern business environments. Increasingly, there is an emphasis on the role of improved information access and informal learning, both in the fields of learning and KM. However, despite the advances in technology to capture and store information and knowledge, our ability to search for useful information remains limited (and may even be degrading as information mounts). In this paper we discuss the problem and evaluate some existing solutions: centralized content management strategies and platforms, and federated search. We present a new solution in the form of advanced federated search tools and a metadata management application that are part of a new release of an industry-leading learning content management system, ForceTen, from Eedo Knowledgeware Corp.

139 words

Keywords: information search and retrieval; distributed content; metadata; federated search; taxonomy management

Limitations of Training in Support of Organizational Performance

It is becoming increasingly evident that traditional, formal approaches to training in organizations are inadequate as a response to the performance issues facing us with the advent of modern, competitive global markets and the growing pre-eminence of knowledge work and the role of the knowledge worker.

In the past, two problems have been recognized as endemic to training. First, training is rarely directly associated with the performance gaps attached to specific individuals. In most cases, we do not know exactly what are the individual gaps. And even where this is known, through some form of needs analysis, the set of tools available to respond has generally been inadequate. Most typically, one-size-fits all courses – which serve no audience very well – are developed and deployed across the board. Given that the learning is not contextualized within the different specific roles found across the organization, transfer of learning, even where the content is related to performance problems, very often approaches zero.

Second, even if individual performance is measured and somehow managed, the further link to organizational performance and results is usually missing. With rapid changes to technology, business strategy, tactics and goals, and the related adjustments that occur with respect to products, services, structure, workflow, and business rules, it is difficult to imagine how, with traditional tools and methodologies, whole training curricula can possibly be re-oriented and rewritten in a cost-effective, timely manner.

Informal Learning and JIT Access to Information and Just-enough Learning

There is increasing recognition that the volume and rate of change in organizations is now simply too overwhelming to be addressed only through the development of formal training programs. The cost of developing programs is too high, the impact generally too low, and the time to delivery is too long. Informal learning and access to information "just-in-time" (JIT) and in a "just enough" format are now being viewed as key elements in any response to the dynamic nature of businesses and markets (Weintraub *et al*, 2002; Dickoever, 2002).

The fields that deal with human performance in the workplace have responded with a growing set of methodologies and tools: rapid authoring tools for e-learning development; learning content management systems (LCMS) platforms to support effective, timely content management; performance support tools (to reduce the dependency on time-consuming, expensive training), and; knowledge management strategies designed to capture and disseminate new knowledge as it evolves. On a conceptual level, the emphasis on frameworks that promise quick access to new information as a response to the challenge posed by rapid change can be seen very clearly in both the spheres of training and performance support and of knowledge management. Some influential thinkers include Sam Adkins, who has written extensively concerning workflow-driven JIT learning (Adkins, 2003), and Yogesh Maholtra (2000), who has argued it is necessary to advance information strategy to "internet time" in order to support business performance – "real-time learning for the real-time enterprise".

"Access" is Problematic Despite Portals, Vortals, Content Management or OM Systems

Despite the growing recognition that "access" is key and the rapid development of information technology, including search engines, access to information remains problematic. In the knowledge management domain, it is fair to say that technologies designed to capture or store knowledge and information (repositories, data mining techniques, the internet, etc.) have advanced further than the tools that are currently available to make that knowledge available. For example, according to Pederson (2001), as of October 2000 there were some 2.5 billion static pages on the World Wide Web along with over 550 billion "invisible" or dynamically generated pages. He estimates that the most powerful search engine, Google.com, can only retrieve 0.1% of these. This marks a significant reduction from a study by Lawrence and Giles's (1999), only two years earlier, which concluded that all of the search engines combine could potentially retrieve 16% documents published on the web. The Semantic Web initiative of the W3C has sought to address this by developing schemes to automate the semantic indexing of content, but it is arguable that little of practical value has followed from the Semantic Web initiative to date.

Within organizations we find we are faced with the same challenges that confront us in our attempts to search for useful information on the web, as well as some additional ones. The amount and variety of information has grown. Much of it is not indexed or even universally accessible. Some content is dynamic and so therefore invisible to conventional indexing and retrieval methods and tools, as well as most advanced technologies. Finally, a growing proportion of content is not textual, but rather is comprised of images, sound, or video. There are no simple schemes or techniques for capturing and categorizing multimedia documents.

From a straightforward perspective of aggregation, the situation just described presents a considerable

challenge. One solution for organizations is to create a central repository or database, and store and index content within this repository. This process, however, requires significant effort and resources. Each document needs to be sorted and tagged and, typically, not all of this can process can be automated. In XML-based systems there is a big investment in creating and maintaining the required style sheets and DTDs, and in training content producers in their use.

More telling, even, is the reality that any larger organization is inevitably constructed of silos, and there will likely be more than one source or repository of information. Portals do provide a single point of access to distributed on-line information from such repositories, and are increasingly an important element of IT architecture for business. But they are far from an ideal solution. Perhaps the most significant shortcoming is that they impose a single view of the content – they are defined with respect to a community of users who share common tasks and interests. Hence, they are not likely to serve all audiences effectively. “This is especially true for internal corporate portals, where different functional and organizational groups and lines of business may have substantially different needs for information access and organization. Examples include sales and marketing, best practices, competitive intelligence, research and development, and general corporate resources.” (Mack, Ravin, & Byrd, 2001)

In response to this predicament, more recently vertical portals, or *vortals*, have emerged as IT solutions to information access. These are specialized portals that provide in-depth capabilities that are highly focused on a vertical segment of an organization or field. This approach, however, poses its own set of problems. Effectively, *vortals* reify the gaps or white spaces (Rummler & Brache, 1995) within an organization - the areas between silos or vertical components of the organization where information and issues can escape any attention. The end result is that much useful information and knowledge, though captured, is never applied because it is never accessed.

Moving away from the literature and experience of IT solutions in commerce, we find similar claims in the KM realm. As Dzbor *et al.* (2000) recognize, there are fundamental problems that prevent organizational memory systems from working efficiently. Organizational memory (OM) – the way an organization applies past knowledge to present activities – can significantly influence the competitiveness of an enterprise. But, employees must have straightforward and fast access to OM. “Employees and knowledge gained during their engagement are often lost in the dynamic business environment. Those who stay with the company are often unaware of critical resources that remain hidden in the vast repositories” (Dzbor, Paralic, & Paralic, 2000).

Overall, portals and centralized content management systems suffer from basic limitations:

- it is difficult to aggregate and index information; much useful information remains isolated on individual hard drives, network drives while other information is not captured because it is unstructured or dynamic
- it is near impossible to get individuals across an organization to apply a fixed indexing or metadata scheme with any great degree of reliability and accuracy; professionals are required to do the job
- the schemes are either too simple to be useful to a wide audience with different information needs, or too complex to be useable or applicable and too effortful to apply in the first place.

The problem of aggregating information is very significant in general in the field of content management. Some organizations have attempted to implement centralized repositories with content management platforms. The difficulties with this approach have been summarized above. Even where some success is had with indexing schemes and retrieval – the result of a long, expensive, iterative process of analysis and testing involving information specialists and a variety of stakeholders within the organization – the difficulty remains that much information will continue to fall outside the confines of the repository. There is also the circumstance that only a small proportion of organizations adopt enterprise content management strategies, along with supporting technology such as Documentum, to begin with. There are probably considerably less than a thousand true enterprise content management scenarios in the world today (some subset of the Fortune 1000, given that such implementations cost in the range of 7-8 figures in US dollars, and the worldwide market for content management is considerably less than a billion dollars). Among

those organizations, only a small proportion of content is managed in, and accessible through, the repository. Estimates in large organizations may be as low as 1%, this figure influenced in part by large volumes of legacy content that cannot easily or quickly be moved into the repository and indexed. Still, even with an extensive program to automate indexing and move content into the centralized repository, it is to be expected that a large volume of information and knowledge will continue to be created and stored in other repositories, in the form of unstructured documents, and on individuals' local hard drives and media.

Some partial solutions to this problem exist and have been tested commercially. For example, in the world of academic libraries, a number of commercial (e.g., Metafind and Metalib) and open source tools have been developed to support "federated" searches across repositories. These have existed for over decade. In general, these are based on the Z39.50 standard as a protocol for cross-database communication, combined with a variety of additional standards that must also be supported in any useful system: SQL, Dublin core, XML, MARC, HTTP and others.

While these systems work, the basic problem is that they really amount to a single login to access multiple databases via a portal, together with basic functionality for multiple searches (such as de-duplication of results). Each database will have its own unique metadata, fields, controlled vocabularies, and syntax. A federated search using these tools is really a lowest common denominator search that will only operate using common elements across the repositories that are combined through the search. The search will not exploit any unique features of the repositories' metadata schemes, thus reducing the power of the search.

Content management needs a new, distributed paradigm.

In general, it can be argued that most content management and LCMS vendors are operating with a faulty paradigm; namely the paradigm of information control based on a central repository and monolithic indexing scheme, as critiqued above. A recent content management industry report from Gartner Consulting (2001) touted the emergence of distributed content management and peer-to-peer content networks as the next killer application in the content management space. But the reality, in the intervening period to 2005, has proven otherwise.

At the same time, desktop technology is rapidly evolving to enable advanced forms of search for local content. For example, Google offers technology for image search and for email (Gmail) and the new Mac OS X Tiger offers both customizable information display widgets (the "dashboard") and advanced search capability via the "Spotlight", a search utility that exploits metadata and can be focused on specific categories of content (email, image files, documents etc).

The growth of personal information management and search technology is also reflected in the emergence of a wide variety of technologies for searching specific kinds of digital content such as images, video, melodies based not on metadata, but on pattern recognition. In many ways, the state-of-the-art in enterprise content management and LCMS needs to advance – needs to accommodate the distributed nature of information, and needs to incorporate more advanced categorization and search capabilities.

Advanced Technologies for Information Access

The future of enterprise content management, in our view, does lie in the direction of distributed content management. This means enabling federated searches across multiple repositories, incorporating local media within searches, and managing metadata in such a way as to provide the ability to view and manage content according to the needs and requirements of different segments within an organization. The reality is that organizations will always contain silos, and much useful information will never make its way into any content repository. Content management strategies and technologies in our space (learning content management) need to reflect this reality, if individuals are to have access to the information they require to support performance.

In response to these challenges, emerging technologies seek to overcome the limitations of existing approaches as regards categorization of content and access to distributed content. The following sections

report on *Enable* and *Unify*, two new tools integrated into ForceTen, an industry-leading LCMS developed by Eedo Corp.

What follows is a brief description of how Eedo Corp has developed the capability for federated search and metadata management within its current 3.0 release of the learning content management system, ForceTen. We focus primarily on the taxonomy management tools that allow for taxonomy development, and for reconciling the different metadata schemes inherent across different repositories and with different document formats stored locally. The conference presentation will feature a demonstration.

Enable

Eedo's LCMS, ForceTen release 3.0 supports federated search across a variety of repositories, subscription-based information services, and also local media, through network connections and the browser, with a tool named *Enable*. Connectivity is based on a variety of protocols and a set of robust XML-based APIs.

In addition to establishing the necessary basic connectivity and interoperability across sources and repositories, *Enable* will also harvest metadata native to the different repositories and embedded in documents. A set of basic functions allows users to leverage common elements, or construct advanced searches using elements of different components. Some basic reconciliation of or management of metadata elements are possible within *Enable*.

Unify

A second application, called *Unify*, can be layered on top of *Enable* to provide a more powerful approach to federated search. *Unify* replicates much of the functionality of stand-alone taxonomy management tools, such as Wordmap's taxonomy server. Using *Unify*, a ForceTen user or administrator can modify and, if desired, reconcile the taxonomic schemes of any ForceTen repositories. When used in conjunction with *Enable*, the same functionality is extended to other information sources connected through federated search, including (with appropriate permissions), local files, email and subscription information services.

The basic functions include:

- *Managing ForceTen connections*: create or delete a connection, edit a connection, load information from a connections
- *Managing taxonomies*: navigate a taxonomy tree, transfer content, create new nodes, delete nodes, view objects associated with a node, modify permissions for access to nodes
- *Searching for objects (at different levels of aggregation)*: preview items, view properties and modify some properties. Search results are viewed in an object viewer, or under the taxonomy tree, according to user preference. Search can be limited by selected courses, databases, connections, object types, subjects, and metadata (e.g., language)
- *Managing courses*: export courses, delete courses, view course details
- *Managing metadata*: view metadata, create custom metadata, reclassify metadata, delete metadata, associate metadata, unassociate metadata.

In combination with ForceTen's support for portlets, these tools provide a powerful mechanism for integrating information resources across an organization, and for establishing metadata schemes or views that reflect the needs of a particular audience better than common elements (lowest common denominator search) or perhaps the native scheme presented within a particular repository.

At the same time *Unify* also provides the basic capability to manage an organization's taxonomy over time. While taxonomies are correctly viewed as living entities, or works in progress, that should evolve to meet the emerging needs of users, the reality is that most taxonomies are relatively static – they remain fixed at the point of implementation. The reasons for this are twofold. There is, first, often a lack of planning, so that no processes and mechanisms for on-going taxonomy governance are ever established. Secondly, there are the problems associated with taxonomy revision from a technical standpoint. Typically, revising a taxonomic scheme within a content management system is a task that requires significant effort. Objects

and documents must be reclassified using scripts, and to some extent, manually. The process is time-consuming and subject to error. *Unify* provides the functionality required to manage taxonomies over time, in a relatively efficient fashion, and with reduced risk of error. Having this capability integrated with the ForceTen platform also encourages organizations to address the organizational issues associated with taxonomy governance, and set in place the required criteria, roles and workflow.

Other developments

Currently, these applications are based on robust technology: APIs and simple schemes for implementing permissions, such as SMPT. In the future we look to implementing new iterations using emerging paradigms such as Service Orientated Architecture and information buses, and advanced standards and technologies for creating user-centered views of content – in particular Topic Maps. This work is already advanced. For example, in 2004, Eedo participated in a major research project funded by Industry Canada, in conjunction with Kontenstu Corporation and Concordia University, to develop and assess Topic Map technology. As part of this work we demonstrated the feasibility of navigating disparate repositories by conjoining topic maps reflecting different metadata schemes (Shaw *et al* 2004a; Shaw *et al* 2004b).

Our commitment to distributed content development is also reflected in earlier innovations, such as the implementation of a distributed learning content model. Under this model, parent-child relationships are established among implementations of ForceTen. An organizational element with a ForceTen repository can declare other environments (other repositories) to be child environments of their own environment. In this scenario, an organization, *A*, that is a child environment of another organization, *B*, can access content contained within *A*'s repository and can sequence and incorporate this content within their own. However, they cannot modify the content accessed from the parent. The scheme is intended to support situations where, for example, a manufactured product (such as an aircraft) may involve subsystems manufactured by different firms. Within this scheme the customer, for example, may be empowered to aggregate and sequence the appropriate training materials for a given configuration of the manufactured product. However, the content is maintained by the various subcontractors who created the related subsystems. This scheme offers a cost-effective way for customers to create the training they need, or for the prime manufacturer to create customized training for customers more cost-effectively (Shaw, 2004c).

References

Adkins, S. (2003). *Workflow-based e-learning: Next generation enterprise learning technology*. Retrieved December 18th 2004 from ASTD's Learning Circuits: <http://www.learningcircuits.org/2003/aug2003/adkins.htm>

Bradley, P. (1995). *Towards a common user interface*. *Aslib Proceedings* 47 (7), 179-184.

Dzbor, M., Paralic, J., & Paralic, M. (2000). *Knowledge Management in a Distributed Organisation*. In L. M. Camarinha-Matos, H. Afsarmanesh, and H-H. Erbe, editors, *Advances in Networked Enterprises. Virtual organizations, Balanced Automation, and System Integration*, pages 339--348. Kluwer Academic Publishers, September 2000.

Dickover, Noel T. (2002). *The Job Is the Learning Environment: Performance-Centered Learning To Support Knowledge Worker Performance*. *Journal of Interactive Instruction Development*, Vol. 14, No.3 pp. 3-9.

Garshol, L.M. (2004). *Metadata? Thesauri? Taxonomies? Topic Maps! Making sense of it all*. *Journal of Information Science*, 30 (4), 378-391.

Gartner Consulting (2001). *The emergence of distributed content management and peer-to-peer content networks*. January 2001.

Lawrence, S. & Giles, C.L. (1999). *Accessibility of information on the web*. Nature, 400, July, 107-109.

Malhotra, Y. (2000). *Knowledge Management for E-Business Performance: Advancing Information Strategy to 'Internet Time'*. Information Strategy: The Executive's Journal, 16(4), 5-16.

Mack, R., Ravin, Y. & Byrd, R. (2001). *Knowledge Portals and the Emerging Digital Knowledge Workplace*. IBM Systems Journal, Vol. 40, No. 4.

Pederson, A. (2001). Semi-automated web resource discovery and analysis: An approach based on interactive machine learning principles. Dissertation. Agder University College. Norway.

Rummler, G., & Brache, A. (1995). *Improving Performance: How to Manage the White Space in the Organization Chart* (2nd ed.). Jossey-Bass: San Francisco, California.

Shaw S., Venkatesh, V., Lowerison, G., Dicks, D., & Zhang, D. (2004a). *Search-and-Retrieval Tools in E-learning Applications: An Empirical Comparison of Search Engines and Topic Maps in an Educational Context*. Proceedings of ELearn 2004: World Conference on Elearning in Corporate, Government, Healthcare and Higher Education. Washington, DC, Nov 1-5, 2004.

Shaw S., Venkatesh, V., Lowerison, G., Dicks, D., & Zhang, D. (2004b). *Topic maps – User-friendly search engines for broadband applications*. Ninth International Conference on VSMM. Montreal, October 15-17. Electronic proceedings only.

Shaw, S. (2004c). *Distributed Content Models -- A new innovation in Learning Content Management Systems and Strategies*. Proceedings of ELearn 2004: World Conference on Elearning in Corporate, Government, Healthcare and Higher Education. Washington, DC, Nov 1-5, 2004.

Weintraub, Robert S., Martineau, Jennifer W. (2002). *The just in time imperative*. Training and Development, Vol 56, No. 6, pp. 50-58.