Intelligent Delivery of E-Journal Content

Introduction:

This is both a celebration and an interim report. It's a celebration of seeing in PubSCIENCE some early results of Walt Warnick's vision of the Department of Energy's (DOE) product becoming "visible, used, and praised." It's also an interim report of the technologies, processes, and management that OSTI is exercising to bring PubSCIENCE to the knowledge workplace. This report details how OSTI, in collaboration with Soph-Ware Associates, is working to transform information to knowledge. The terms "Smart Document" and "Intelligent Delivery" are far more than sales hype; they are the primary drivers for OSTI's mission in a very competitive market. The requirements for valuable-useful-praiseworthy content deserve careful examination. Soph-Ware's product, called IDEA, offers an opportunity for PubSCIENCE to turn a portion of the DOE vision into a working reality.

What is PubSCIENCE?

The Department of Energy's PubSCIENCE initiative began in early 1996 with the initiation of a Small Business Innovation Research Grant to explore ways in which distributed sources of information could be accessed, managed and shared using the Internet as the delivery mechanism. As this initiative matured, it rapidly evolved as a means of delivering the Department's peer-reviewed scientific journal literature. The intent of PubSCIENCE is to facilitate the search, access and delivery of journal literature of interest to the DOE scientific community and the public.

As an organization with a rich history of Information Management within the Department's scientific community for the past fifty years, The Office of Scientific and Technical Information is seeking to establish PubSCIENCE as the premier source of Energy scientific and technical information. OSTI is vitally committed to the communication of the Department's research results to the Department's scientific community and the public. The journals that PubSCIENCE supports, in collaboration with its partner publishers, focus on Energy sciences, technology and medicine. As the focal point for access to the Department's research and development, OSTI is absolutely committed to lead the transition to the use of new communication technologies to enhance Energy scientific and technical information exchange.

Working with its publishing partners, PubSCIENCE seeks to add new dimensions to Journal electronic on-line publishing; the user will benefit from the ability to search across and access multiple journals, achieving efficiencies not previously realized. Working with the publishers' subscription policies, PubSCIENCE brokers subscriber access to participating on-line journals, from the individual user to Department-wide consortia subscriptions. Of course, much of PubSCIENCE's content is absolutely free and available to the public.

What is Soph-Ware's role in PubSCIENCE?

Soph-Ware is the provider of a product called Intelligent Document Exchange Architecture (IDEA). The company developed IDEA under a DOE Small Business Innovation Research (SBIR) grant from 1996-1998. Soph-Ware is enhancing and modifying that product to function as the delivery engine for PubSCIENCE. OSTI and Soph-Ware are collaborating this year to develop a procedure for accessing both DOE research publications and scientific electronic journal content. As the outcome of this engagement, OSTI is to assume total responsibility for the acquisition, production, hosting, and delivery of PubSCIENCE content.

How does content become valuable, useful, and praiseworthy?

A primary goal for PubSCIENCE is to add new and demonstratable value to that content. Another goal is to make that content more useful than it would otherwise be without the mediation and support of PubSCIENCE. And the most difficult goal of all is to elicit audible praise from the scientific and R&D communities. Valued, used, and praised—a tough mission for OSTI and some tough goals for PubSCIENCE.

What we have proposed as the best strategy for PubSCIENCE is for its delivery to be intelligent. Specifically, we have engineered the platform such that the user can see a Smart Document instead of just a Glass Article. Once the user perceives an e-journal document as smart, he or she is more likely to perceive the information in the article not merely as published content but as useful knowledge. And there's that K-word again. So whether this was what the DOE intended with PubSCIENCE or not, the Department has placed itself squarely in the knowledge management business. In data warehouse terms, PubSCIENCE is not just a data mart or an info mart. It is—or must soon become—a knowledge mart. (I'm hoping that no one will make the next logical transition, calling PubSCIENCE a K-mart!)

What is a Smart Document?

An e-journal article is not smart just because it's electronic, or just because it's searchable or repurposable or interactive or hypertexted or because it offers four SAVE-AS options for output. These are all familiar features of browsers, search engines, and COTS enterprise publishing tools. The intelligence of a really Smart Document—one that is most likely to become knowledge for the user—consists of three important varieties of knowledge:

1. Knowledge about its content

Thanks to structured markup technologies—SGML and XML for text and other ML's for audio, video, virtual reality, three-dimensional objects, music, mathematics—the content of a document bears knowledge about itself within itself. It knows about its own hierarchical structure: Set, Volume, Number, Article, Section, Paragraph (in the case of a journal). Because markup standards are accepted and promulgated, the article is able to expose that knowledge within any environment in which it finds itself.

2. Knowledge about its consumer

   Information becomes knowledge only by personal involvement with and ownership by the consumer. To me as an R&D manager, a publication begins to become value-added knowledge when I can bookmark it, edit it, transport it, and reuse it. Again, these are only out-of-the-box product features. But when a collection of e-journal articles also remembers my personal preferences, my session histories, my exceptions and exemptions to normal business rules, my precise authority to access the collection, the profile of my workstations, my normal search queries, and the current status of my subscription to each fee-based collection. . . then the collection begins to feel like knowledge.

3. Knowledge about its transaction

   A smart document knows the terms and conditions of its own transaction. This entails rights management, fee-based access, all security issues, links to publishers' internal accounting systems, print-deliver-bill-me (for a 275-page tutorial). These are business rules that govern our access and consumption of the e-journal content.

It may seem that we are carelessly blurring the distinction between data and metadata, content and information about the content, information and instructions for processing the information. In the spirit of object-oriented data bases, we are doing precisely that, and we're doing it deliberately. This integrated, three-fold intelligence of the Smart Document can move an e-journal out of its repository and into the consumer's mind.

How can PubSCIENCE become content-as-knowledge?

There is no guaranteed cookbook approach to creating knowledge in the mind of the consumer. But there is an essential agenda for an e-journal system that hopes to achieve it. Here is a brief profile of a PubSCIENCE that aspires to become a knowledge resource that is valued, used, and praised. More to the point, it describes a system that will survive even worse budget cuts, inevitable technology shifts, and adverse policy decisions.

1. PubSCIENCE is built upon an architecture. It is not simply a collection of COTS products. An architecture presupposes that someone has consciously designed for a particular function, just as in construction with wood and bricks. That design (1) specifies all of the components, (2) describes how everything is to be interconnected, and (3) predicts exactly how each component will interact with the other. The shopping cart approach—selection based on vendors' slicks—can yield a demonstratable system. But it will begin to hit various "brick walls" far sooner than a system based on an explicit information architecture.

2. PubSCIENCE's architecture is open. The Department cannot afford to become reliant on a vendor, either because of possible captivity to the vendor's terms and conditions or because of the vendor's possible demise.

3. PubSCIENCE is scalable. This week it is clear that PubSCIENCE can only succeed if it is "thin-client," requiring next to nothing special on the user's workstation. But the pendulum between thick and thin has never stopped swinging, so PubSCIENCE should not lock itself into any single blend of client and server resources. But "scalable" also means that PubSCIENCE must plan for its own best-case scenarios on the server side. Some e-journal systems already require supercomputers for pre-processing their data and very high-end, distributed computing power for delivery.

4. PubSCIENCE is a three-tier system. The client is too thin, and the publisher's data repository is either inaccessible or too busy. That familiar scenario has spawned the age of the mid-tier server. This is the computer or cluster of computers whose function is to field the user's requests, fetch the data from wherever on the planet the pieces may reside, preprocess the results, and then serve up pre-digested content to the client, along with some program code for further processing the data.

5. PubSCIENCE will be savvy about business rules. OSTI is currently engaged in serious intellectual property statesmanship, negotiating PubSCIENCE's electronic access to journal collections from several prestigious publishers. That diplomacy will mean nothing if the PubSCIENCE deployment engine does not faithfully execute those terms and conditions.

6. PubSCIENCE's content is structured. Whether you have good or bad feelings about SGML and XML, all serious published material will bear structured markup. That is true for the SGML material from the APS. It is true for 63 journals from five professional societies in the DLI test bed at the University of Illinois.

What does IDEA bring to the party?

Fortunately for PubSCIENCE, IDEA was born and incubated at OSTI, for which we all

thank the SBIR program. And while IDEA has electronic content applications far beyond e-journal delivery, it contributes heavily and specifically to PubSCIENCE in each of the areas described above. And it does that in the following ways:

1. IDEA is an architecture by design and by implementation. The over-arching requirement in the SBIR solicitation was that the architecture should account for (1) multiple and disparate platforms, (2) heterogeneous content, and (3) distributed stored objects. The delivered system incorporates and fully implements all of this. PubSCIENCE may not need this power over the next few weeks, but it is inevitable that it will in the future. Journals will incorporate non-text and non-graphical data. A single "article" can very well be a virtual entity, stored physically at various repositories worldwide. As for hardware and operating systems, no one can predict these with certainty beyond the current fiscal year.

2. IDEA is open. IDEA is indifferent as to which search engine returns the user's search results. It doesn't care what security scheme or product controls the user's access to the system. It allows for OSTI to elect for Oracle as the underlying database for the middleware, although it is currently supported by MS SQL Server. It can freely add conversion tools to the Transformation Manager.

   IDEA is also open and standards-based with regard to content. It can manipulate content in any format or notation. Most importantly, IDEA's smart locating mechanism allows for content fragments to be added, revised, moved about, and deleted at will without any harm to the repository.

3. IDEA is scalable. For the Air Force we presented an entire specialized implementation with all of the software, including NT Server, contained within a single laptop. In the demo area this evening, you will see PubSCIENCE delivering from a single Pentium III server. In the future, OSTI will surely need to upgrade and expand the PubSCIENCE server, possibly to incorporate multiple, distributed processors. That has already been implemented and tested.

4. IDEA is a mid-tier server plus a mid-tier application. Forrester research calls this combination—mid-tier server and middleware—an "application framework." IDEA incorporates its own mid-tier server, performing all of the system-level brokering and preprocessing functions described above. In addition, it is a middleware application specifically designed and implemented for electronic content delivery.

5. IDEA is driven by business rules. The system is capable of controlling and negotiating system access, whenever that becomes necessary. It is able to be proactive about rights management so that PubSCIENCE need not rely strictly on the publisher's password-protected Web pages.

6. IDEA speaks XML/SGML natively. IDEA can manipulate storage units of any content regardless of format. But with XML/SGML content, it can actually penetrate intelligently to the lowest internal structural level. So for an on-line report in which a particular table is updated every few minutes, IDEA will always serve up a totally current version of the report, including the up-to-the-minute revision of the table. And since XML/SGML data is entirely portable, it does not matter what authoring or editing tool prepared the content.

Why IDEA just for e-journals viewing?

IDEA is both a feature-laden mid-tier server and application middleware, more than is required for the e-journal content that the Department will host during the next few months. Were the Department never to impose any restrictions on access and were all e-journals to be single self-contained entities and were the PubSCIENCE server to operate forever on a single microcomputer, the capability of IDEA would be unnecessary. However, any one of the following scenarios will immediately require the functionality of IDEA as an open systems application platform:

1. An Energy Data Base research document comprises several distributed entities, located at various DOE laboratories and consisting of live, dynamically updated data.

2. The Department chooses to install a new search engine or an array of search engines, some that possibly aren't even developed yet.

3. Experience demonstrates that some "choke point" in PubSCIENCE becomes intolerable, requiring that separate computers process some or all of the various services: search, authentication, authorization, location, retrieval, pre-processing, transformation.

4. A publisher announces that its content is XML or SGML, that it will support more intelligent searches and mid-tier processing, and that it will negotiate dialogue between the user and authors, using software that it will freely provide to the mid-tier server.

5. The Department wishes to track and account for access to the EDB content. The Department commits to managing access and subscription to a particular e-journal or data base collection.

6. It is necessary to deploy on-the-spot customized editions of a particular publication, based on user profile, geography, affiliation, specialty, or any other criterion.

Summary

PubSCIENCE is entering the e-journals arena at a fortunate moment, because it can gain much from lessons learned in similar initiatives elsewhere. The challenge facing PubSCIENCE is to achieve and maintain a competitive advantage as a value-added e-journal host. Demonstrating intelligent deployment in a convincing manner is the best strategy for adding that value, thereby getting its product visible, used, and praised.