A Publisher's Perspective

More Successes Than Failures

Tim Ingoldsby

The University of Illinois Digital Library Initiative (DLI) has offered scientific publishers a unique opportunity to participate in a digital library research project oriented toward answering research questions of great interest to publishers. This discussion will represent the views of the publishing partners who contributed to the DLI project. Participation was very rewarding for the publishers and the individuals who attended the annual project briefings and learned from the renowned Illinois research team.

In preparing this article, a review was performed of quarterly reports submitted to the National Science Foundation since the project began in 1994. Also consulted were other publishers who contributed materials to the project to obtain their impressions of the preliminary outcomes of the project. I am grateful for their contributions.

AMERICAN INSTITUTE OF PHYSIC'S HISTORY IN THE PROJECT

When the University of Illinois was awarded one of the six DLI grants in 1994, the American Institute of Physics (AIP) was not listed as a partner.[1] However, soon after the announcements, a session was held at the American Association for the Advancement of Science's (AAAS) annual meeting featuring the principal investigators for each of the awardees. During that AAAS meeting, I had a chance encounter with Bruce Schatz in the exhibit area, where I was demonstrating AIP's recently released pioneering online physics journal, *Applied Physics Letters Online.*[2] The journal was a weekly letters compilation and had been available in full text SGML since October 1994. Schatz was having trouble finding publishers for the DLI project that could contribute science or engineering journals

in SGML. He invited AIP to join the project, and we enthusiastically accepted. Since *APL Online* was a research project anyway, we saw this as an opportunity to get two results from a single development project. We would be able to learn from the successes and failures of each system and approach.

Since we already had a large body of some 1,000 valid SGML articles, *APL Online* became the first journal to be incorporated into the DLI system. This was extremely valuable for AIP, as we became the first publisher to benefit from the careful analysis of our SGML and the successes and failures encountered by the project team as they crafted a powerful SGML-based search and retrieval system. We were so pleased with the early results that we offered two additional journals in SGML[3] to the project in mid-1996. The power of our SGML implementation was further validated when these two new journals were incorporated into the project with almost zero additional effort.

By this time, the American Institute of Physics had made a business decision to establish our own online journal service and move all sixteen of the AIP physics journals online by January 1997. As a result of this extensive project, our ability to perform additional research in house was drastically reduced, and the DLI became our primary research and development center. Much of the look and feel of the user interface for our own service came from lessons learned in the DLI project. Another result of our preoccupation with our new online service was our decision to provide some financial support to the DLI project so they would continue to perform some routine SGML production services that we had earlier agreed to take over so the project staff could concentrate on bringing the materials submitted by other publishers online.

Another aspect of the financial agreement with the DLI was the development of a distributed repository at AIP. By mid-1997, our Online Journal Publishing Service (OJPS) was stable and fairly robust, and we were anxious to model a distributed database environment, which we believe to be one of the likely eventual environments for the online research environment. We agreed to host the first distributed repository for the project. As this article was being written, the DLI staff was replicating the database structure and was in the final stage of testing before AIP journals began to be searched on our Woodbury, New York, server by users of the DLI system on the Urbana-Champaign campus.

REVIEW OF THE RESEARCH: TESTBED

From a practical standpoint, the development of the testbed and DLI's demonstration that it is possible to use SGML from many publishers in a single coherent system is the project's most significant accomplishment. DLI research programmers had to develop a system that could receive

SGML files conforming to at least six different DTDs and merge them into a uniformly searchable system. The volume of material must have been staggering for researchers that were used to small test databases. The American Institute of Physics was delivering fifty to sixty articles weekly, and the American Physical Society an average of sixty to seventy more. The programmers developed auto-processing systems that resulted in their ability to add new issues to the collection in a matter of a few hours. Many weeks I have found a new issue of *APL Online* available on the DLI even before it is mounted to our own OJPS even though both services receive the material at essentially the same time and the DLI has to process full articles, whereas our OJPS only deals with bibliographic information and abstracts. The OpenText search engine selected by the project staff has been thoroughly tested, and practically every capability it provides has found a use in the DLI implementation.

The testbed group also made great strides in interface design. The American Institute of Physics recognized early in the project that the collaboration between the testbed group and the User Studies/Social Sciences team would produce a very functional interface design. The use of a custom client, developed with Visual Basic (VB) rapid prototyping tools, allowed frequent updates to the functionality in reaction to user preferences. However, the reliance on VB also had a drawback—i.e., the client needed to be installed on every workstation every time it was updated. The custom program approach also was contradictory to users' overwhelming preference for a standard Web browser. Perhaps the project staff stayed with the custom client too long, but the Web DeLIver client that was developed in late 1997 benefitted from the rapid prototyping phase.

The final validation of the excellent work of the testbed group came with the announcement that DARPA was providing a grant to continue the testbed for an additional three years. We expect to see additional accomplishments during this phase.

REVIEW OF THE RESEARCH: USER STUDIES

One of the hidden joys of this project has been the opportunity to interact with the User Studies group led by Ann Bishop at the University of Illinois at Urbana-Champaign. They have been a joy to work with, and AIP has drawn on their expertise for our own user studies. Publishers have much to learn from the social science aspects of this research. We often feel that we know what our customers *need,* but Bishop's research shows that we do not necessarily know what they *want.* We have applied what we have learned from the User Studies group's use of various tools and feedback-gathering strategies to our own interactions with our customers. Their interview techniques have been very instructive. Another valuable insight has come from their instrumentation of the client to gather feedback when

it is fresh in the mind of the user. Bishop's group has contributed much of value to the process of understanding how researchers approach their tasks. Now that the testbed is to be extended to the wider Committee on Institutional Cooperation (CIC) community, user studies will have even more validity based on the expansion of the test population.

REVIEW OF THE RESEARCH: SEMANTIC RESEARCH

This group, led by Bruce Schatz and Hsinchun Chen, has produced the most forward reaching "Star Wars" research results. Their work is definitely more "research" than "development" but, even so, they have produced results that have meaning. They are leading the way to the future, even though, for most publishers, their efforts are beyond our event horizon at the present—with one exception. The IODyne research tool developed by Eric Johnson is a concrete example of the type of search aid that will assist the scientific research community in the near future. We at AIP are very excited about applying the concepts used by Johnson in doing keyword searches to enable more effective use of AIP's Physics and Astronomy Classification Scheme (PACS) codes in searching physics research.

Schatz and Chen have excelled in educating publishers about the obstacles to semantic retrieval. They have also found interesting uses for supercomputers—e.g., building concept spaces for large research collections. They have also enhanced the simplistic approaches of single word searches by developing noun phrase algorithms that improve the quality of machine searching results. They are also doing the fundamental research with concepts such as self-organizing maps and algorithms for automatic categorization. Here is an area where research performed in one of the six digital library projects has applicability and benefit for all of the other projects as well.

REVIEW OF THE RESEARCH: SYSTEM EVALUATION

This final component of the research for the DLI project has been the least visible in terms of accomplishments and continuing efforts. It is possible that much of the reason for this has been the way that the world has changed during the past four years. It is instructive to read the original proposal in which NCSA's Mosaic was expected to be the vehicle for much of the information delivery. It is sad to note that most of the "newbies" to the Web have never even heard of Mosaic, the compelling application that drove the rapid expansion of the World Wide Web. Even so, I have been personally disappointed by the lack of contributions to this project by the NCSA and Computer Science group. The quarterly reports of the project document the diminishing impact and value of this aspect of the project.

HOW ONE PUBLISHER HAS PROFITED FROM INVOLVEMENT IN THE DLI PROJECT

The American Institute of Physics has already begun to profit from our participation in the DLI through the process of technology transfer. Our efforts to develop the AIP OJPS have profited from the research activities of the DLI project. Our search interface, in particular, owes much of its heritage to the Visual Basic client prototyping. SGML to HTML conversion routines developed at AIP have been improved by discussions with DLI staff. The DLI problems with the display of SGML full-text articles, particularly the problems with rendering special characters and mathematical formulas, convinced us to remain with PDF as our primary full text "deliverable" for now.[4] Our service has also attracted other society customers. We now deliver thirty-six journals through our Online Journal Publishing Service. The awareness of our partnership with the DLI project has been cited by many societies as a factor in their decision to work with AIP's online service. Even so, it will be 1999 or later before the OJPS can deliver features that have been present in a research mode within the DLI project since 1997.

MAJOR SUCCESSES OF THE DLI PROJECT

In keeping with the theme of this conference, there was an attempt to identify the key successes and disappointments of the project. Clearly, the development of a "proof of concept" cross-publisher large-scale federated repository is the DLI's greatest achievement. At least until sometime in 1999, the DLI testbed will remain the largest sci-tech SGML article collection in the entire world. This testbed is beginning to deliver the promise of SGML: searchability across many titles, programmatic linking between articles published in journals of different publishers, and powerful fielded searching.

The DLI project has also made great strides in achieving a functional interface. This interface now incorporates results gleaned from interviews and other diagnostic processes that define how scientists and engineers use journal articles in the research process. Schatz and Chen have also produced impressive achievements in advancing the quest for semantic federation. These strides will lead to systems that make such tools commonplace in the next decade.

MAJOR DISAPPOINTMENTS OF THE DLI PROJECT

While it would be too strong to call them "failures," there have been two significant disappointments regarding what had been proposed for the project. Clearly, the project has failed to deliver adequate display of SGML-based scientific literature. If SoftQuad's Panorama had worked as

anticipated, AIP and other scientific publishers would be rushing to convert from PDF full-text delivery to SGML. This remains the most difficult problem for publishers and may require government support to solve.[5]

The other disappointment is the final size of the testbed and the user population that was proposed to interact with it. Instead of the 100,000 documents and 20,000 users promised in the original proposal, the final numbers will be closer to 50,000 documents and 1,000 users. The publishers participating in this project are perhaps most disappointed by this shortcoming. Our judgment is that the testbed group stayed with the Visual Basic client for perhaps six months too long, thereby not leaving enough time to expand the user community to its fullest possible extent. However, the follow-on DARPA grant should permit the achievement of both goals.

LOOKING TO THE FUTURE

For its part, the American Institute of Physics is looking forward to continued collaboration with the testbed and user studies groups through the DARPA grant and the establishment of an industrial partners program for publishers. We would like to see more journals added to the testbed and will offer all sixteen AIP journals (which have been in full-text SGML since the beginning of 1998) on the distributed repository at AIP. We want to see the testbed group turn its efforts to solving the SGML math display problem. If MathML emerges as a viable solution, tools will need to be developed to convert ISO12083 math markup into MathML markup. The special character aspect of the display problem is well on its way to being solved by the STIX font project[6] that recently submitted to the UNICODE standards body a proposal to add every character required for mathematics, physical sciences, and life sciences to the standard code set which is, or shortly will be, supported by every browser developer. We have also suggested to the testbed group the development of a joint DLI-2 proposal to address these issues and others that are being faced by the scientific and engineering research community.

ACKNOWLEDGMENTS

AIP wishes to take this opportunity to thank Bruce Schatz, Bill Mischo, Tim Cole, Ann Bishop, and their associates for offering us the opportunity to participate in a research project of this calibre. The cooperation between university and publishing has been beyond expectation.

NOTES

[1] The American Institute of Physics was a participant in the unfunded proposal, *Science Quest*, submitted by the University of Maryland.

[2] *Applied Physics Letters Online* was the first physics journal made available through OCLC's Electronic Journals Online program.

[3] *Journal of Applied Physics* and *Review of Scientific Instruments.*

[4] Results from our electronic journals online experience with OCLC also confirmed the difficulty of using HTML full text supplemented by GIFs of special characters and formulas.

[5] However, the recent development of the Extensible Markup Language (XML) standard, its widespread endorsement by browser and rendering software vendors, and (most importantly) the development of the Math Markup Language (MathML) standard written in XML offer perhaps the best hope for a solution to this serious problem.

[6] A project of major scientific publishers including AIP, APS, ACS, IEEE, AMS, and Elsevier Science.