# From isolation to integration: re-shaping the serials data silos

*Based on a paper given at the 26th UKSG Conference, Edinburgh, April 2003*

**This article is about the so-called 'data-silo' phenomenon, in which content is isolated by the publisher or aggregator, and which generates an unreasonable cluster of destination points for the harried student or professor to visit, each with its own interface, set of functions and behaviors. A user has no confidence that what one can do on Site A (limit browsing by a community's terminology; bookmark at the article level) will also be true at Site X, and libraries have limited abilities to build services for local needs on top of such a disjointed data landscape.**

**DAVID SEAMAN**
Executive Director, Digital Library Federation

## Introduction

Academic libraries and scholarly publishers are part of a process whose end result is designed to be richer pedagogy, better-informed scholarship, and an increased democracy of access. Increasingly, scholarly publications move from author to reader electronically, which provides faster transmission of primary and secondary material, and increases the chance of timely discovery of relevant data.

Unlike electronic publications, printed journals and books on a shelf resist interrogation en masse – you cannot search through 200 printed books and journals at once – and reusing snippets of them to create a new publication or presentation is also complicated by the print medium. Electronic publications have the potential for speedy discovery and rich re-use; however, a potent combination of economic, production, and emotional forces have largely hobbled the ability of electronic books and journals to operate in rich aggregations that are shaped by the subscriber – the scholar and student – rather than the publisher or aggregator. The same publishing forces have also largely removed the end user's ability to interact richly with the online content, re-using, recombining, annotating, and enriching it to create new expressions of the work, personalized and tailored for a local audience. Instead, the data is accessible online but in formats (the web browser) that resist annotation, enrichment, or even on-screen reading, and the data is organized by publisher and aggregator, with no ability to search across them. If your 200 books and journals are housed in 30 different web-based aggregations, each of which looks and behaves differently, there is a significant barrier to working with that digital material en masse.

What I am going to say about data silos is here largely uncomplimentary. However, I do want to recognize the value of a well-crafted single product: when a coherent, articulated, limited body of material meets your needs it has a real value and can be a pleasure to work with. At present, the cost of that well-crafted product is too often that the content can only work in that manner – and cannot interact with similar content from other sources. We routinely are bemused by a sophisticated audience who nonetheless repeatedly exhibit a predilection towards simplistic, un-Boolean, searching habits ("we give them rich searching and they do one- and two-word unrestricted searches") and we bemoan the Googleization of our audience without any sense that they behave that way because of the hostile data terrain we present them with as they navigate across many sites.

## The Digital Library Federation

The DLF (http://www.diglib.org/) was created in 1995 to provide an organization to focus exclusively on digital library issues in the large academic libraries that are its founding members and partners. We are designed to be focused, nimble and agile. We are a collaborative space and an incubator: practical, strategic, and relatively small, by design.

Initiatives such as the Open Archives Initiative (OAI) (http://www.openarchives.org) and The Metadata Encoding & Transmission Standard (METS) (http://www.loc.gov/standards/mets/) have developed with deep involvement from the DLF as an organization and through the work of our allies and individual member libraries. I am not going to list all our initiatives here, but to highlight a couple that are relevant to this audience I can point to the ongoing E-resources Management initiative, actively creating a new XML format to hold the contents of the e-journal and other electronic licenses we all struggle to manage (http://www.diglib.org/standards/dlf-erm02.htm); we have a lot of time invested in production benchmarks and good practices (http://www.diglib.org/collections.htm); we tend to be involved early with new technologies such as Shibboleth – a new authentication method from the Internet 2 community, designed to work in a very different way to the current, often unsatisfactory, IP authentication (http://shibboleth.internet2.edu/) – and we are involved in both LOCKSS ('lots of copies keep stuff safe') (http://lockss.stanford.edu/) and the institutional repository movement.

What I will talk about in more depth is a 2002 DLF-sponsored survey that measures aspects of the behavior of scholars and students as they interact with scholarly information in their research and teaching: *Dimensions and use of the scholarly information environment*, freely available online (http://www.diglib.org/pubs/scholinfo/). The survey is large: 3,234 hour-long phone surveys of faculty members (930), graduate students (1,056), and undergraduate students (1,248) from over 400 institutions, from liberal arts colleges to large research universities. It is one of those bodies of information that has something in it to delight and irritate everybody. My comments are also informed by my impressions from a recent, whirlwind tour of the 34 university campuses and partner organizations that make up the DLF – a rite of passage for a new DLF Director. It is a luxury to visit many institutions quickly and in doing so to form a picture of what concerns, challenges, and delights us in our work with digital library content and users.

## The data silo

One of the common issues that came up in every visit was some aspect of the limitations of the 'data silo' phenomenon. We have a growing awareness that our faculty are irritated with the fact that the electronic resources we put in front of them simply don't play well together. This is not a new problem. It reminded me very much of a decade ago, talking to users and vendors about cd-based product, and its limitations (data isolated from disc to disc; different interfaces to learn; the problems of networking the data). By moving online we have solved one access problem but have made little inroads into the data isolation and integration issues; we can now get to isolated datasets from the comfort of our own homes, but we haven't solved the basic problem that things simply don't play well together across aggregators and across publishers.

This situation is not native to the way we work as academics in a university library. We don't shelve books by publisher. It doesn't make sense to do this because faculty and students do not work by publisher, and are largely unconcerned if a book is by X university press or Y scholarly publishing house except maybe when they are publishing a book themselves. Yet in too large a measure in the digital library, that is exactly how we present material – the ordering principle that governs a cluster of adjacent material is often a standalone website that presents a co-ordinated view of works by publisher X or aggregator Y. As a service industry, the academic library is faced with a real problem as we try to mediate between those piles of remotely-housed digital content and the needs of our users to access by subject, or author, or to re-package it within a learning management system or courseware package. Too often, we cannot re-shape that content, or enrich its metadata, in order to provide good local service.

## The imagined user

It is curious to sit back and think; well who did we think the user was when we started building digital libraries? I suspect we didn't think as deeply as we should about their daily work habits, on the whole. We thought a lot about access, a lot about design, a lot about pricing models, a lot about budgets, but you can still find examples on library web pages of interminable alphabetical lists of "stuff we subscribe to," arranged roughly by subject. Who is the user of such rich but disjointed lists? Either one who is so specialized that he or she only needs one or two ready-made aggregations, or somebody with so much time and expertise that they can go in and out of dozens of websites, learning their different ways of working and pulling the stuff together. The latter type of user is rare and – when spotted – is usually a librarian.

The current situation also tends to lead to a user who is passive – a user who visits content, but doesn't expect to aggregate it themselves, to bring it home, to re-shape it, to drop it into a desktop tool to analyze it. I was very struck at a recent meeting of the IMS, one of the organizations in the learning objects and courseware space, how almost every other paper had something to say about the need for teachers to have desktop tools to drop online content into in order to re-shape or enrich it or contextualize it before delivering it to their students. It is a need that is equaled I suspect in the research libraries, where faculty who work with online content, or build it, want to bring it home and have some control over it. It is part of the same urge that had us Xeroxing articles and sticking them in file cabinets, I suspect, in previous iterations of the scholarly endeavor.

## Surveying the scholarly information user community

So, what are we beginning to know? Well I am going to present simply this one survey by way of a glimpse of what users tell us. These data are all public, on the Digital Library Federation website, and it is a survey of over 3,000 phone interviews with undergraduates, graduates, and faculty, at small colleges through to big universities.

I don't know if it is a surprise, but one of the things we heard across the board is that print predominates. You may find something electronically, but finally you want to print it out. Printed content still often predominates as a source. There has been a fast uptake of electronic content, clearly: faster in research than teaching, and fastest with the undergraduates. This varies across disciplines – no surprise there – but the uptake of electronic content isn't what I would have expected, given the nature of the population that is being surveyed here – highly wired and wireless academic communities with very high computer ownership and computer access ratios. I would have expected these numbers concerning reliance on electronic content to have been somewhat higher. Either there isn't enough of the right material in digital form or it is too cumbersome to use. Or both.

> RESEARCH: 35% of the faculty members said they rely exclusively or almost exclusively on electronic sources, with the grad students reporting 49% [Law, 65%] Exception: arts and humanities – 25%
>
> TEACHING: About 25% of respondents rely all or most of the time on electronic sources for teaching – Exception: business (42%) and law (30%).
>
> UNDERGRADUATES: About half (49.2%) of undergraduates reported that they used electronic materials exclusively or almost exclusively [business students, 62.9%]

The survey has a lot to say about unmet needs. There is a level of frustration over the amount of material available, in finding material, in working with it, and users are concerned that they are not adequately trained to operate in this information space. They know it is difficult to get content out of this morass and they are not sure whether it is their fault or ours. There is a brief window of opportunity for us to fix the situation before they work out it is largely our fault as publishers and libraries.

> **Content**: About one-quarter of the faculty members (25.6%) and well over one-third (38.3%) of graduate students expressed a need for more online journals.
>
> **Time:** Nearly 40 percent (38.8%) of the total sample of respondents and 60 percent (60.2%) of the faculty reported "having enough time" as their major problem.
>
> **Retrieval:** Respondents expressed frustrations with finding information, determining its credibility, and analyzing it.
>
> **Training:** 38.4 percent of respondents saw having insufficient training on how to find information as an impediment.

Time (or rather the lack of it) was one of the things that came through this very strongly for me, that almost irrespective of discipline, one of the limiting factors that people said they had is they don't have enough time. This is important for us as we deliver services. It suggests in fact, that as we suspected all along, they don't have time to go in and out of silos. It is not only unattractive as a way of working, it is not feasible. So what do they do? Well, we know what they do. They find something, they stick with it, they go to Google, they go to Amazon, or they go other places instead of the services we offer them.

## The courseware space

I don't know what the situation is here in the UK, but in the States, most institutions now have a courseware system. It may be an off-the-shelf one such as WebCT or Blackboard, or it may be one that an institution has built locally. It is an online service in which you build a webpage for your class – a syllabus, a chat room, an e-mail function, and so on – often a service that is very popular with the faculty. However, it is extremely difficult sometimes to get the library's digital content into these class pages, which further lessens the value and reliance on the digital library content. We have faculty who are going to their libraries and scanning articles from print journals, to make a PDF file, because they want to link to that article from their courseware page. This is a perfectly reasonable thing to want to do. They have a syllabus and they want it to be linked. They have access to the electronic version of that journal in their digital library, but too often it is impossible to make a stable, persistent link at the article level between the digital library holdings and the courseware system. This is a very significant need. So, we are buying the paper copy, we are buying the electronic copy, and then we are paying again locally to digitize the article from paper, in order to satisfy a basic use need such as connecting it into the courseware system. This is not going to be a sustainable model, especially in the current financial situation. Like many of the problems with a failure of interaction between systems, it is rarely simply a technology problem. There are lots of technologies, standards, and emerging standards that help us in this space. The root of the problem tends to be other forces – habits, branding ("if I don't have a silo will they know it is mine?"), and ownership.

## Sustainable data behaviors

So what do we need? It is certainly not all doom and gloom, and we know from our experience with locally produced and/or loaded content that we can build much richer services on top of it and encourage much more sophisticated use than we typically see in the e-journals usage.

**Malleability:** We need the data that resides on publishers' sites to be much easier for us to re-shape for local customized delivery and analysis. This doesn't necessarily mean that the content needs to be loaded locally, but it needs to be built with the assumption that somebody else will be putting a front-end on it, will be mixing and matching it with content not in that repository.

**Multiplicity:** PDA, wireless, e-book, text-to-speech, and print on-demand are all here or coming, and content that can't go where the users are will under-achieve; it is not just a web world now. While the web is a wonderful delivery mechanism still and the predominant one, we are seeing a lot more interest in other delivery technologies. We need to look at our content with the assumption that it will need to work in a multiplicity of delivery formats. This doesn't necessarily mean that the publisher of that content has to publish it in every form known to man. It is enough if the customer has the ability to re-shape that content for what may be a localized need. So, we need content that is customer reshape-able, and with XML and style sheets we have the basic technologies and standards to do this.

**Management:** We need the ability for a library to build local services that allow users to interact richly across vendors. Publishers need to help libraries be data aggregation services for the libraries' customers, and the Open Archives Initiative and Open URL give us good tools to start to achieve this.

**Mix:** We currently invite our users to visit sites and watch them like TV channels: you go, you look, with all functions, aesthetics, and relational features provided by the publisher. It is not a particularly rich or personalized interaction. The alternative to 'data TV' is that of the data 'mix' – the ability to sample, re-use and re-package as a personal library, a classroom presentation, or to annotate, to build on, to cross-search items from disparate places by bringing them local.

## Concluding remarks

We need to work as a community to move us to a point where we create data that behaves freely, that 'plays well with others'. We need to evolve away from the current situation to one where the library and the scholar can mix and match across silos, and can create new views ordered by criteria other than publisher/aggregator. We need data that behaves freely within its licensed user community, to encourage richer use; if we want our users to be innovative we need to give them data and tools that help them innovate.

We need to bear in mind time and our users' lack of it. Google isn't predominant because our users think it has got everything we have got in our libraries. They know it doesn't, but they can get a result, often a useful (albeit an incomplete) one and they can get it on a time scale that makes sense.

Unchecked and unmodified, this data silo situation has – at worst – the potential to seriously undermine our ability to perpetuate the current library/publisher relationship in academia, and at best it continues to foster users who are limited in their ambition and development by content that fails to maximize its recombinant potential.

**David Seaman**
**Executive Director, Digital Library Federation**
**1755 Massachusetts Ave., NW, Suite 500**
**Washington, DC 20036 USA**
**Tel: 202-939-4762**
**Fax: 202-939-4765**
**E-mail: dseaman@clir.org**