R2 Consulting Services

Rick Lugg
Ruth Fischer

Print on Demand and
Virtual Approval Shelf

A Report for the Informed Strategits/Tri-College Consortium Project

December 17,2001

Part I of this report focuses on Print on Demand (POD) and short-run digital printing (SRDP) options as they relate to libraries. Part II concerns availability of and sources for extended descriptive metadata (jacket scans, annotations, excerpts, etc.) for new titles, and how that metadata might support "virtual approval plan" service.

Note: An LC Classification breakdown of new academic titles published in FY 00/01 was obtained from the Blackwell's Book Services Web site, and is not included in this document.

PART I: Print on Demand & Short-Run Digital Printing:

Significant changes in printing technology have occurred over the past five years. The economics of traditional offset printing, with its high set-up costs, has historically compelled publishers to print more copies of a book than they can sell, because with offset printing the individual unit cost declines when amortized over a larger base. But when all the copies don't sell, as is often the case with scholarly material, the press is left with an over-investment in inventory.

Digital printing of variable text, has been available for some time, largely due to the emergence of PostScript and PDF as machine-independent software. Digital printing allows publishers to produce small runs of books (from 25 to 300) as they are needed—making digital printing especially suitable to backlist titles that would otherwise be declared out of print, as well as newer titles that are expected to have lower demand. In some ways, this approach to printing, known as short-run digital printing (SRDP), is a natural fit for scholarly monographs.

But until recently, there have been barriers to its widespread adoption. First, the cost of offset printing has continued to decrease, making it a stronger competitor. Printing of color images is less satisfactory in digital printing. Book covers require a separate printer and process. And printing companies aren't yet seeing enough demand for digital printing to realize a reasonable return on the new equipment they must purchase.

But within the past two years, the quality of SRDP has improved, and a number of publishers, both trade and scholarly, have begun to incorporate it into their operations for print runs of 1 to 500 copies. SRDP does require that the content be in digital format, a factor we'll revisit shortly.

In R2's research and discussions with publishers about print on demand, it became clear that most of them think of SRDP and Print on Demand as one and the same. To quote

from the University of Chicago Press's Web site: "There is no absolute distinction between SRDP and print on demand (POD). POD is often used to describe the production of the quantity of books needed to fill a single order…[whereas] "SRDP often implies the production of small quantities (25 to 300) to satisfy both immediate demand and short-term inventory requirements."

SRDP also implies that the actual printing occurs at a printer's facility or a distribution center. The most far-reaching visions of POD posit a network of distributed printers or "kiosks" in copy shops or bookstores or libraries, where a single copy can be printed instantly, using new equipment such as the InstaBook Machine or the MIT Perfect Book machine mentioned above. Ultimately, once repositories of digital content are built, and digital print files can be easily moved across the Web, where the printing is done will be a matter of convenience.

Standards: PODi is a printing trade association that "evangelizes" about the advantages of print on demand, and promotes interoperability and standards. The most important standard in this industry is Personalized Print Markup Language (PPML), "a common print language for variable print, which is vendor- and format-neutral. It is an XML-based open standard that is configurable with the major POD file formats: PostScript, PDF, TIFF, and JPEG.

Sources for POD Content:

Lightning Source International (LSI): is a division of Ingram, and the largest print-on-demand provider in the US. They currently have a digital library of more than 100,000 "orderable" titles, from approximately 1,300 publishers. Since 1997, when they opened for business, LSI has printed more than 3 million POD units. In December, they announced a partnership with OCLC to create cataloging information for the Print on Demand title in the LSI digital library.

At present, most of the titles available through LSI are backlist, for which the publisher's stock of copies originally produced by offset printing has been depleted. However, this may change significantly in the future—especially because LSI is also a major distributor of eBooks in MS Reader and Adobe eBook formats. This means that LSI may already possess digital files for newer content, files that can be easily adapted for digital printing. For new titles with low demand, it's conceivable that a publisher never actually print copies for inventory—but rather rely on LSI to distribute the title in eBook or POD formats.

Replica Books: Like LSI, Replica Books is a subsidiary of a large distributor, in this case Baker & Taylor. We weren't able to establish an exact number of titles available through Replica. But some browsing of their title file suggests that Replica's title count is considerably smaller than Lightning's, and includes mostly older and public domain works. Replica's pitch to publishers early on had mostly to do with saving books from going OP, so this is not surprising.

But again, as with LSI, Baker & Taylor's Informata division is building a new product called "ed", which stands for e-content distribution. As "ed" is launched in the spring of 2002, Informata will begin building a library of digital content. It seems likely that they would tie this file to Replica's printing processes.

Major publishers: As noted in our earlier report on eBook availability, many large publishers are investing in new production systems that support format-neutral digital

repositories. Once these systems are in place, it will become routine for frontlist titles to be available for short-run digital printing. Newer titles, i.e., those produced in these new production systems, will already be in digital form, either PDF or some typesetting format. These are candidates for SRDP with minimal additional costs.

Conversion of backlist titles is more complex and expensive, and is likely to be done selectively. Typically, these titles have no usable electronic file, and must be taken apart and scanned to create a TIFF (or image) file. Because TIFF files are not searchable, a second file is created using OCR software, which, with some further work, will result in a searchable text file. This approach works reasonably well for text, less well for images. Cost of such conversion ranges from $.25-$.75 per page.

Many major publishers, of course, work with large printing partner such as R.R. Donnelly or Edwards Brothers to produce this content, but also work with print on demand providers such as Lightning Source.

Scholarly Publishing Print on Demand: In November, the University of Chicago Press, which operates the Chicago Distribution Center (CDC) on behalf of 20 mid-sized university presses, received a $1.5 million Mellon Grant to establish a short-run digital printing center and a digital book repository (called BiblioVault). Their ambitious plan calls for the conversion of 2,300 backlist titles (using scanning and OCR) and 2,850 recent titles (using existing electronic files prepared for online use) to create the Chicago Digital Distribution Center (CDDC). The CDDC will first enable titles for short-run digital printing (SRDP), and subsequently for distribution in various eBook formats.

To quote from their grant proposal, "SRDP…will allow publishers to produce small runs of books (25 to 300 units) as they are needed…." and will be based on "a repository (BiblioVault) for the digital files needed to produce these books."

In another announcement just this week, the CIC Consortium, which represents libraries and university presses at the University of Chicago plus the Big Ten, announced a formal initiative to " 1) identify collaborative models for university press operations and 2) develop and test a pilot cooperative e-publishing venture." As we've seen, e-publishing and print on demand are closely related, and this second large initiative by a group of university presses indicates the importance of these technologies for scholarly communication.

Analysis: POD and Libraries: To boil this all down, we can assume that POD is becoming increasingly available for selected backlist, and for books going OP. Over the next two years, more and more frontlist content, especially for scholarly material, will be available for SRDP—though the actual printing is likely to take place at printers or distribution centers. Within 2-3 years, distributed print on demand will be a reality. But how does any of this help the library, especially in terms of space planning? SRDP or POD books, after all, are still print books and take up just as much space as those produced by offset presses. However, here are some possible implications:

- Fear of OP diminishes: At least some decisions to acquire new titles are influenced by the short print runs that characterize scholarly publishing now. If a title is not ordered when first announced, it may not be possible to obtain it later. When POD is widely available, there is less risk in waiting to acquire a title – an therefore room for a different title on the library shelves.
- Just-in-Time Acquisition: In a publishing environment where eBook and POD delivery are prevalent, it's possible to adopt a different access strategy. For

instance, the library can purchase and load MARC records for any titles it wishes, but not actually purchase the material until a patron requests it. Given the speed with which eBooks and POD can be delivered, this may prove an attractive option, as it assures that any title purchased will be used at least once.

- More aggressive weeding: If it can be assumed that a title can be obtained again later if needed—via eBook or POD, perhaps libraries will more willingly discard material that is not circulating.
- Test market each title: A fully digital infrastructure among publishers gives libraries additional options. For a given title, a library might purchase eBook access first – then buy a print copy if demand or type of use warrants. Shelf space is not used until a title has proven itself.

Eventually, when distributed POD becomes common, it's conceivable that a library patron, having examined the library's eBook copy, may wish to purchase a print copy, a transaction which might be driven from the library to the local copy shop, with some revenue accruing to the library as a result.

PART II: The Virtual Approval Shelf:

In order for selectors at all three of the TriColleges campuses to participate in a shared approval plan, some method must be devised for selectors to share evaluation of each title. In traditional approval plans, the book itself is shipped to the library, and selectors decide to accept or reject, based on examination of the book. In a shared plan, selectors from all three campuses will want to evaluate the same books. Moving books to all selectors, or selectors to the books is inherently inefficient.

Some of our discussion during the meetings on October 18th revolved around creation of a "virtual" equivalent to this evaluation process. Essentially, we envisioned a scenario in which the various approval vendors would provide a list of which books were to be sent to the TriColleges, as well as a list of form selection titles. Those lists would be loaded into Innovative, and a shared view of the bibliographic records constructed.

But all agreed that more than a brief bibliographic record is needed to make an informed selection decision. We decided that I would investigate what types of extended metadata is available, including tables of contents, cover images, annotations or summaries, reviews, excerpts, and full text—and further, on what basis these sources might be licensed to libraries. It's especially important to know which sources can be reached through Innovative.

Innovative Interfaces: I spoke at length with Ted Fons, Product Manager for Acquisitions and Serials, and much of the necessary capability already exists, even if not integrated into a single product. First, and most importantly, Innovative is the only ILS vendor that has made provision for linking to external metadata from the Acquisitions module – other vendors have focused on linking from the OPAC only.

This means that approval or form selection records exported to Innovative from Book Bag or Collection Manager or GOBI can be collected into weekly selection lists, and that from the staff display of these brief bibliographic records, an ISBN link can be generated to whatever remote metadata sources the library chooses: Table of Contents records from Syndetics, for instance, or cover images and inventory information from B&T's Informata division, or a full MARC record from WorldCat. The only condition is that the target resource must have a predictable URL. Other indices besides ISBN (e.g., title) will

be incorporated into the next release. Depending on the extent of data sources wanted, it may be necessary to purchase Innovative's Web Bridge module.

We also discussed the need for some kind of transaction history to indicate which selectors from which of the TriColleges had reviewed a title, and a record of their decision—something as simple as adding a location code might work, but there are some wrinkles due to the fact that there are separate accounting units involved. This aspect will require further discussion.

Innovative is in the final stages of negotiations with Syndetic Solutions (see below) to license packages of Syndetic's metadata, and will in turn license it to libraries if wanted, but the system is open, and the library is free to use whichever metadata sources it wants.

Some work may be necessary to integrate the service further, but for Innovative libraries it does appear that most of the building blocks already exist. We then must address the question of what metadata sources are available, and especially for new titles, whether the information is timely enough to aid the selection decision.

Extended Title Metadata: Availability and Sources:

Baker & Taylor/Informata: Informata licenses descriptive metadata to system vendor such as Innovative Interfaces, epixtech, and Gaylord through its Content Server product.

Data currently available includes 625,000 jacket scans and 250,000 tables of contents. Full-text reviews from several important sources are under verbal agreements and will be incorporated within the next few months. Annotations, mainly from publisher marketing copy, will also be added.

It's unlikely that Content Server will make full text available, because of licensing considerations. However, it's possible that some excerpts that are now available in Title Source II, another Informata product, will be re-licensed for distribution through Content Server.

According to Bob Bogan, VP, Product Development at Informata, Innovative Interfaces has created links to this extended metadata not only from their OPAC module, but also from the Acquisitions module.

Although jacket scans, tables of contents, and reviews are important determinants in approval selection, they are needed at the point of selection for titles that are new. Only about 20% of the data in Content Server is for forthcoming titles. Full information is typically not available until after the book is published—meaning that very little information is available from Content Server at the time it is most needed to support an approval plan. It would be useful to test this in a very simple way, by keying ISBNs from a current AcBC invoice into Content Server, and quantifying the hit rate.

This will likely change over time, particularly as more publishers adopt ONIX, the new XML-based standard for exchange of book product information. Currently, only 20-25 presses are seriously pursuing ONIX implementation, so widespread availability of earlier advance information may be as much as two years away.

Syndetic Solutions: Syndetic Solutions's primary business is the creation and distribution of descriptive metadata to libraries, online booksellers, and their partners.

Data currently available includes 370,000 Tables of Contents (aggregated from Blackwell's and Ingram), 200,000 annotations (from Book News); first chapters and excerpts (though primarily for public library-oriented material), cover images (from Ingram), author biographies, and reviews from CHOICE. Syndetics also offers a unique service for Fiction titles, in which Genre, Sub-Genre, Location, Characters, and other elements are explicitly identified and tagged. Discussions underway with academic publishers (Routledge, Sage, Academic Press) for 2-3 page excerpts that might be displayed under Fair Use terms.

Syndetic's principals, Allan Graham and Jeff Calcagno, are both former Blackwell's employees, and understand the need to bring descriptive metadata into the selection process—rather than simply into the OPAC.

Because many of Syndetics's Table of Contents records come from Blackwell's, I checked quite thoroughly on how and when the TOCs are created, with both companies. When new titles arrive at the Blackwell's warehouse in New Jersey, the table of contents is scanned as part of the receiving process. The scanned image is sent electronically to Blackwell's approval offices in Oregon, where keying and formatting is done. (Some of this may be further outsourced, not positive.) But the turnaround time on this process is quite short—one week in most cases. This means that a Blackwell's/Syndetics table of contents record *should in most cases be available within a week after the title is profiled*. Again, this would need to be tested, but suggests that at least Tables of Contents might be available at point of selection.

Those TOC records are available both from Innovative (which licenses TOC data from Blackwell's) and from Syndetics. (They're probably also available directly from Blackwell's.) Syndetics has also partnered with Innovative to offer two "packages" of its other metadata—one comprehensive package, and another "pared down" version that excludes Publishers Weekly reviews, School Library Journal reviews, and first chapters. The timeliness of this information in relation to selection is not clear, but is also easily investigated.

Syndetics's future plans include provision of indexes and bibliographies, which will help with evaluation, and Z39.50 compliance of the Syndetics database.

OCLC Extended WorldCat: As of December 13, OCLC began receiving "evaluative" data, including cover images, book summaries, and author notes from publisher catalogs, as well as table of contents data from Ingram. In addition, of course, WorldCat provides a full MARC record and information on holdings for one's own institution and others—information which can influence a selection decision. The inclusion of Ingram's data is still quite new, so it's not yet possible to determine its timing in relation to new title selection.

OCLC Resolution Services is still in development but boasts an innovative design. Formerly known as Open Names, this is essentially a global registry for metadata, clustered around the ISTC (International Standard Text Code), a pre-ISBN identifier. Any metadata provider who has information about a particular work (e.g., "The Name of the Rose") or a product based on that work (e.g., a hardcover edition published by Harcourt Brace; or the original Italian edition) will be able to "register" that metadata with OCLC Resolution Services. When, for example, a WorldCat search retrieves an ISTC or ISBN for which descriptive metadata is wanted, the user can opt to view a menu of available metadata sources, and link to their sites.

This service is still being built, and discussions with metadata providers are just beginning. In addition, it would need to be able to accept a search or a link from an ILS system in order support selection. But it bears watching.

Bowker—booksinprint.com: This resource, of course, includes Bowker's *Forthcoming Books* data, as well as TOCs, cover images, and reviews. Bowker's data is sometimes hampered by the fact that they don't actually see the books, as B&T, Blackwell's, and Ingram do. Again, as far as timing, it would be important to work from a list of form selections or a current approval invoice to determine currency. Bowker does offer access to their Web site, which includes real-time inventory data for several vendors, via OptiWare. It's likely that the Innovative Web Bridge could also be used for this resource, though the representative I spoke with didn't fully understand what I was asking.

Analysis: The Virtual Approval Plan:

I'd like to go out on a limb here, and suggest that the TriColleges group is an ideal environment to attempt proof of concept for a virtual selection process—not only for individual libraries but for a consortium. There are a number of significant aspects to this opportunity:

1. TriColleges has a need to enable shared evaluation and selection of new titles.
2. TriColleges libraries share a library system, and a systems librarian.
3. That shared system is Innovative, which already has a number of the necessary features—especially the ability to link to outside resources *from the Acquisitions module*.
4. The TriColleges selection process requires coordinated communication at the individual title level—who has reviewed or selected it, who has reviewed and passed on it. This dimension will be critical in a consortial environment, but would best be tested in a small consortium like TriCo.
5. At minimum, it appears that Table of Contents, a full MARC record, and some information from publisher catalogs may be available at the time of approval review or form selection. This can be tested very quickly and very early in design of this service; i.e., we can gauge feasibility with minimal effort.
6. We can determine this feasibility at a relatively early stage, and at fairly low cost—enabling a clear "go" or "no go" decision.
7. In particular, Innovative, Syndetics, and Blackwell's expressed a great deal of interest in this project, and all offered to help in any way they could. They seem to realize the potential value of this approach.
8. A number of large publishers, such as Cambridge University Press, are experimenting with making TOC data available in advance of publication—which may improve the amount of evaluative data for selectors.
9. Looking not very far down the road, there will be eBook repositories of new titles to which this service might be pointed—perhaps allowing, say, 15-minute access to the full text of a new title to inform a selection decision. A selector could then decide between print and electronic format.
10. Development of a "virtual selection" facility now would demonstrate leadership and vision, and would position TriColleges to support coordinated selection of print and electronic monographs and coordinated selection among the three libraries.

Clearly, this would require investment of time and money, and would have to take its place among many competing priorities. I suggest it so strongly only because the situation, the moment, and the capabilities line up so well. It strikes me as an unusual opportunity, and an interesting, forward-looking project.