June 2006

Back Talk -- Janus Conference Key Challenges: Whacky and Non-totally Whacky Ideas

Anthony (Tony) W. Ferguson

University of Hong Kong, ferguson@hkucc.hku.hk

Follow this and additional works at: <https://docs.lib.purdue.edu/atg>

Part of the [Library and Information Science Commons](#)

Recommended Citation

Ferguson, Anthony (Tony) W. (2006) "Back Talk -- Janus Conference Key Challenges: Whacky and Non-totally Whacky Ideas," *Against the Grain*: Vol. 18: Iss. 3, Article 11.

DOI: <https://doi.org/10.7771/2380-176X.4959>

This document has been made available through Purdue e-Pubs, a service of the Purdue University Libraries. Please contact epubs@purdue.edu for additional information.

Predicting High-Circulating Titles for Public Libraries

by **Bob Molyneux** (Chief Statistician, SirsiDynix, 101 Washington Street S. E., Huntsville, AL 35801-4827; Phone: 256-704-7019; Fax: 256-704-7007) <bob.molyneux@sirsidynix.com>

At the last **Charleston Conference**, I discussed our work on predicting circulation patterns of library materials based on data from the **Normative Data Project (NDP)** (<http://www.libraryndp.info/>). We are attempting to predict which titles will have high circulation and which will circulate little. Using that knowledge, acquisitions librarians in public libraries would have more information about what to purchase and what not to purchase. This article gives a short overview of the talk and update where we are now. It turns out the problem is a bit more complex and the value of such a capability greater than we understood at the time.

The **Normative Data Project** is a huge pool of potential discovery. It collects data on various aspects of U.S. public libraries. For example, it includes data from the **U.S. National Center for Education Statistics**, geographical information system (GIS) data from **GeoLib** (<http://www.geolib.org/>), and **U.S. Census** data on all American public libraries. In addition, there are detailed data from the integrated library systems (ILS) of a smaller group of "contributing" libraries. For these libraries, we have additional information on holdings and circulations. We have such information as **Dewey** and **LC** classes, class of borrower (adult, juvenile, etc., but **no** personally identifiable information!), day and time of day of the circulation, title, and format such as book, DVD, and so on).

The first understanding of the problem as discussed in Charleston is: With the information collected in the **NDP**, can we predict titles that will have high circulations? Such a capability would be useful in spending end-of-year money to maximize circulations or in building new public library collections as is done with "opening day collections." Currently these collections are built based on prescriptive lists. But if we can predict which titles will circulate the most, we can be relatively certain that materials bought to be a part of the circulating collection will, in fact, circulate.

Version 1

Right now, using the **NDP**, we can do a good job of predicting a set of titles that will circulate. Demographics of the library's community affect how the library is used. We can match the demographics of a target library with those from the set of contributing libraries based on default market areas around each of the 17,000 public libraries (that is, systems and branches) in the United States. This comparison will identify a set of libraries similar to the target library and reveal what has circulated at those libraries and produce a bibliography of titles and formats that will be sought and tailored to that library's users. We can do this comparison by any of the many dimensions of the data, whether format, type of borrower, **Dewey** class, and so on.

Although this capability is an advance over other methods, there are currently two key limitations to this approach.

One is that a public library will have materials that do not circulate a great deal either because they are reference or because the titles are important for other reasons. For instance, most public libraries will own *Hamlet*, in spite of the title's circulation patterns.

In these cases, we would need to analyze holdings information on the contributing libraries like the target library to discover which titles are in the non-circulating collection and which titles are in the low-circulating collection that still would be important to own. From the holdings at similar libraries, we could build a virtual "core collection" of such items that is based on what *is owned* at like libraries and not from bibliographies of what libraries *should own*. It is hard to say what such core collections will look like. It is reasonable to speculate that most public libraries will own *Hamlet* but only the larger ones will own **Shakespeare's** lesser play, *Cymbeline*. It would seem that the size of the library's population served is an important variable for these both reference and low-circulating titles.

The other limitation concerns the time it takes to compile the **NDP** data because that delay affects the titles that are available. If the **NDP** is three months behind the current titles, any new title from **Stephen King** or **JK Rowling** since the **NDP's** latest compilation will not appear in a list of high-circulating titles although, clearly, such items will circulate. So, such a collection would always miss the latest titles as, of course, would occur with published lists.

Version 2

This latter limitation forced a rethinking of the whole matter, and it is clear that the problem as presented in Charleston — predicting titles that will circulate — is incorrect as stated. The correct way to state the problem is that we should predict the attributes of the titles that will circulate. Authors would be such an attribute in some cases — as the two cited above make clear.

Another probable attribute is the user population's demographic characteristics, which we can estimate in the **NDP**. For instance, one **NDP** user discovered in a set of libraries he examined (where the population had a relatively low percentage of high school graduates) that titles about completing a high-school equivalency were high-circulating items — but these kinds of materials did not turn up in lists of high-circulation items where the population had a high percentage of college graduates. It is scarcely headline making news that a population where many have not completed high school would want more books on high-school equivalency from their library than another library's population where all have completed high school. It is another thing, though, to demonstrate such a relationship with data. That demonstration is evidence that demographics will provide a useful set of attributes in a predictive model.

We might imagine other such attributes, but there would be those we would miss if we set out to list them and, in any case, data mining provides us a method to discover attributes that are both obvious and those that are not obvious. Essentially anything we know about the demographics, past circulation patterns, and the types of books that have circulated may be useful in telling us which attributes of titles will be useful to predict high-circulating titles.

The **NDP** runs on **Microsoft SQL Server**. The new version, **SQL Server 2005**, will have a "predictive analytics" capability based on **SPSS's Clementine**¹ and allow us to test whether this approach will be useful.

If this "predictive analytic" capability yields useful information, then the next decision would be whether to order the full-blown version of **Clementine** or a similar piece of software from another vendor. These applications are expensive and such a step will not be taken lightly.

continued on page 92

Back Talk from page 94

our embracing of the library friendly publishing alternative. I realize that this paragraph can get me in personal trouble so let me assure you all that I do not jump up and down with glee when our acquisitions department sends its annual million dollar check to the Netherlands. But, I am equally non-enthusiastic about the need to subscribe to both library friendly and expensive journals for the same subjects. Consequently, I cannot salute a new call for the subsidy of the publishing enterprise. We already gave to the **SPARC** effort. For me, my job is to give

people reading material and reading material costs money. Moreover, I won't feel better, if in the future, library book budgets are reduced so that the funds can be transferred to the faculty who have to pay the page charges associated with the open access movement.

So, I think the ideas associated with digitizing past and present materials, to preserve our printed legacy, and to negotiate with electronic publishers in a transparent and ethical manner are all good goals/challenges. But the goals to return to the past and resurrect core collection lists in today's electronic environment and to get everyone to contribute to a parallel open access publishing empire are both whacky ideas.

Back Talk — Janus Conference Key Challenges: Whacky and Non-totally Whacky Ideas

Column Editor: **Anthony (Tony) W. Ferguson** (Library Director, University of Hong Kong; Phone: 852 2859 2200; Fax: 852 2858 9420) <ferguson@hkucc.hku.hk>

At a recent **Academic Libraries** discussion group session of the May 2006 meeting of the **OCLC Members Council** meeting we discussed the six major collection development challenges/solutions identified at the October 9-11 **Janus Conference on Research Library Collections** at **Cornell University**. I missed attending the conference myself. Based upon the information about the conference circulated at the **OCLC** meeting, I find some of these challenges/solutions are good, what needs to be done. Others are a bit whacky and deserve to be ignored/buried in the historical refuse pile of librarianship.

Among the better goals or challenges identified is the need for libraries to: (1) "reconvert the scholarly record." That is, digitize everything published thus far. I am going to assume that the goal is to digitize everything that appeals to scholars everywhere. Having said that, I wonder if this isn't a whacky idea. For example, should we digitize all of the pathfinders produced by all librarians everywhere? If so I will try to dig up something I once wrote about how to find information about blue 1957 Chevrolet cars — a well meaning attempt to teach the use of **Boolean** operators. If these pass the test, how about all Sabbath day printed programs of all churches in all countries for the benefit of homiletic historians? Well, yes these examples are admittedly a bit extreme. But since the **Google 5** have already agreed to do 10 or 20 million monographs, and since almost everyone has seemed to have caught the institutional repository fever and are bound to digitize all of their intellectual output, the goal of reconverting the scholarly record should not be seen as a completely whacky

idea. Moreover, it may be reasonable to think that the rest of us should be able to top off these other efforts and digitize the remaining 40 million or so items.

The next challenge, on my list of non-totally-whacky ideas, is to (2) "ensure objects published in the future are available in digital form." Since my library is celebrating the acquisition of its first millionth volume, I find I have to support this aspiration at least in concept. I do need, however, to confess that one of the **Janus** suggestions for implementation includes the recommendation that "research libraries [need to] agree to shift to e-journals, reference books, textbooks, government documents and other areas like electronic books" by 2008. My guess this suggestion was probably designed to be a bit provocative just to get everyone's attention, but I doubt that sufficient changes in the scholarly food chain will happen over the course of the next 18 months to make the complete switchover to digital only possible — but on the other hand if **OCLC** and **RLG** can come together, who knows what else might happen when hell freezes over again in early 2008?

Another non-whacky idea is the (3) goal to "develop and maintain methods to maintain traditional and digital holdings for the long term." The implementation suggestion for printed materials is that "research libraries will create a working group charged with developing models for coordinated print archiving at the national level." Their job will not be easy but they could have thought up some other whacky but effective ideas: Every man, woman and child in the world should agree to burn at least 10 books in order to reduce the size of the problem. So their sug-

gestion to refer the problem to a committee seems fairly

sound. For e-materials, however, the implementation ideas for this otherwise acceptable preservation goal do start to wander off in the direction of wishful thinking, if not wackiness: Libraries need to view the "preservation of e-journals as a kind of insurance" and "libraries must invest in a qualified archiving solution," etc. While this may sound a bit idealistic if not whacky, the last time we got into the permanent preservation solution we got equally emotional about slow fires and started hiding information on silver halide film. The current ideas/wishes on how to preserve e-content are at least as good as the earlier multi-million dollar preservation/reclass project.

The final non-whacky idea is to (4) support the **ICOLC** like principle to "negotiate collectively with publishers on the best possible access to e-content." This includes pushing for the right to use e-resources in course-management software, for e-reserves, for **ILL**, for institutional repositories, that agreements should be transparent, and to insure that perpetual access and preservation is accounted for in license agreements.

In the totally whacky column I have to place two well-meaning but bad ideas: Core collections and taking \$50,000 from everyone's library materials budget to invest in alternative "publishing structures." The core collection idea suggests that we need to "define items that compose a core collection in each discipline and ensure that all research libraries have access to the same core." This probably grew out of the fact that too many librarians over 50 attended the conference. First of all we were trained to know about Books for College Libraries and we get nostalgic about its demise. Secondly, we were the Peace Core generation who believe that all men are born equal and it is our job to make that aspiration a reality. Yet, this idea ranks very high on my list of things for librarians to do while Rome burns — that is if we have any traditional ideas that have not already been reduced to charcoal. This idea is simply no longer needed since the target beneficiaries of our core collections already have their core collections: whatever they can click to on the first two pages of **Google** — if they have extra time on their hands; or, what they can find in Wikipedia if they are in a hurry.

The \$50,000 invest in alternative "publishing structures" idea is an amplification of a now sacred tenant of our rejection of the evil commercial publishing empire religion and

continued on page 93

ADVERTISERS' INDEX

37, 47	ABC-CLIO	49	ASME INTERNATIONAL	73	FRONTIERS IN ECOLOGY AND THE ENVIRONMENT
41	ACM	5	ATG	15	IEEE
7	ACS	83	BASCH SUBSCRIPTIONS, INC.	79	McFARLAND
69	AM. ECONOMIC ASSOCIATION	3	BLACKWELL'S BOOK	96	MIDWEST LIBRARY SERVICE
53	AM. INSTITUTE OF PHYSICS	2	BLACKWELL PUBLISHING	59	MPS TECHNOLOGIES
23	AM. PHYSIOLOGICAL SOCIETY	95	BOOK HOUSE	75	PALINET
33	AM. PHYSIOLOGICAL SOCIETY	14	BROCKHAUS/GERMAN BOOKS	25	PROJECT MUSE
43	AM. PHYSIOLOGICAL SOCIETY	17	BRODART/MCNAUGHTON	65	RITTENHOUSE
67	AM. PHYSIOLOGICAL SOCIETY	89	THE CHARLESTON ADVISOR	19	SAGE
77	AM. PHYSIOLOGICAL SOCIETY	10	THE CHARLESTON REPORT	39	SERIALS SOLUTIONS
87	AM. PHYSIOLOGICAL SOCIETY	27	DUKE UNIVERSITY PRESS	61	SPRINGER
29	ANNUAL REVIEWS	57	EASTERN BOOK COMPANY	35	THOMSON
11	APPI	13	EBSCO PUBLISHING	31	WILEY
45	APPI	9	EMERALD	21	YBP
51	ASBMB	63	ENDOCRINE SOCIETY		

Ad Manager: Edna Laughrey Internet: elaughrey@aol.com
phone: 734-429-1029; fax: 734-429-1711