1. Interest in usage data originally driven by
   a. Availability of ILS circulation data
   b. New book-browser application prototype (StackLife) for Harvard online catalog, and opportunity to offer more context for search results
2. Conceived of as a measure of ? of catalog’s resources; generally valid term difficult to find
   a. Harvard’s “community engagement” or “community usage” with resources
   b. “Scholarly importance” of catalog’s resources
   c. “Relevance” of catalog’s resources
   d. “Recommendation index”
   e. “Most popular” items in catalog
3. Working with intuitive notions of data’s meaning, intended to be provocative in soft-launch, proof-of-concept, explorative way
4. Endless issues emerge concerning meaning of the data
   a. What does a checkout mean when a book may be checked out and never read?
   b. What does a checkout suggest when non-circulating materials of similar contents never appear in the available data or in-house patron browsing not captured?
   c. In what sense is a book checkout similar to an e-download of a journal article?
5. StackScore
   a. Single metric needed to allow StackLife to organize search results and heat-map dynamic collections (“stacks”)
   b. Name evolution: “ShelfLife” => “ShelfRank” => “StackLife” => “StackScore”
5. Usage data at Harvard is primarily circulation and reserves data and hence mainly applies to monographs in online catalog
   a. 88% of collection is books
   b. Supports use-case of StackLife, though might be less useful for other use-cases (anything leveraging journal output, use of visual materials collection, online catalog browsing, etc.)

6. Have started to merge e-journal usage stats (COUNTER compliant) with monograph usage
   a. Graphic example of apples v. oranges problem
   b. Top StackScores in most general subject queries return e-journals
      i. Format facet in book browser addresses this issue
# Materials Formats in Harvard Online Catalog

<table>
<thead>
<tr>
<th>Records</th>
<th>Format</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>12,105,655</td>
<td>Book</td>
<td>88.2%</td>
</tr>
<tr>
<td>624,369</td>
<td>Serial</td>
<td>4.5%</td>
</tr>
<tr>
<td>188,042</td>
<td>Map</td>
<td>1.4%</td>
</tr>
<tr>
<td>154,033</td>
<td>Notated Music</td>
<td>1.1%</td>
</tr>
<tr>
<td>154,028</td>
<td>Sound Recording</td>
<td>1.1%</td>
</tr>
<tr>
<td>150,725</td>
<td>Manuscript</td>
<td>1.1%</td>
</tr>
<tr>
<td>124,699</td>
<td>Other</td>
<td>0.9%</td>
</tr>
<tr>
<td>110,361</td>
<td>Video/Film</td>
<td>0.8%</td>
</tr>
<tr>
<td>96,398</td>
<td>Book Part</td>
<td>0.7%</td>
</tr>
<tr>
<td>21,938</td>
<td>Collection</td>
<td>0.2%</td>
</tr>
</tbody>
</table>
Usage Data Types and Metrics at Harvard

1. Checkouts and recalls of items in ILS (from ILS)
   a. Item barcode
   b. ILS ID
   c. Timestamp
   d. Patron’s Harvard status (undergraduate student, graduate student, faculty)
   e. Patron’s Harvard school affiliation (Arts & Sciences, School of Design, etc.)
   f. Library at which transaction recorded
2. Reserve placements (from ILS)
   a. Timestamp
   b. ILS ID
   c. Course
3. Course-text assignments (from university bookstore faculty order data)
   a. Timestamp
   b. ILS ID
   c. Course
4. Number of Harvard libraries holding given item (holdings data)
5. E-downloads (from COUNTER data; mainly e-journals)
   a. COUNTER format
   b. Journal-level counts
Anonymization

1. Raw transaction-level data
   a. University ID’s deleted
   b. Timestamps no more granular than transaction day (hours, minutes, seconds deleted)
   c. Incoming record order randomized to break apart possible clustering

2. StackScore aggregation and computational transformation
   a. All data aggregated since 2002 (installation of latest ILS)
   b. Raw score totals
      i. Weighted
      ii. Scaled
Usage Data Metric Weightings

1. Survey of Harvard librarians about relative importance of proposed usage-data metrics
2. Asked to weigh relative importance for measuring “scholarly impact” of each proposed metric on scale of 0 to 100, from “not at all important” to “extremely important”
3. Areas surveyed
   a. Collections metadata (number of Harvard libraries holding work, number of copies acquired)
   b. Patron usage metadata (checkouts, recalls, renewals, etc.)
   c. Harvard course-usage metadata (number of times an item placed on reserve or on course reading list)
   d. Online catalog metadata (number of times a record is clicked, exported, associated e-book link clicked, etc.)
4. Respondents asked to weigh relative importance of each metric compared to other metrics in same category
5. Roughly 130 responses received
What metadata is most important for determining a work's scholarly impact?

ShelfRank measures and analyzes a work's use by the Harvard community by collecting various types of metadata. But, in order to determine a work's scholarly impact and make useful recommendations, ShelfRank needs to know the relative importance of each metadata type. Please use the sliders to tell us how important each of the following metadata types are in conveying a sense of a work's scholarly impact in relation to its peers. (We'll ask you to tell us about individual component metrics for each metadata type in the next five questions.)

<table>
<thead>
<tr>
<th>Metadata Type</th>
<th>Not at all Important</th>
<th>Very Unimportant</th>
<th>Neither Important nor Unimportant</th>
<th>Very Important</th>
<th>Extremely important</th>
</tr>
</thead>
<tbody>
<tr>
<td>Harvard collections metadata (e.g., number of Harvard libraries holding the work, total number of copies held by Harvard libraries, etc.)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HOLLIS catalog metadata (e.g., number of times a work's record is viewed, exported, emailed, printed, etc.)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patron usage metadata (e.g., number of times a work is checked-out, recalled, requested from the Harvard Depository, etc.)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Harvard course usage metadata (e.g., number of times a work is required reading for a course, placed on course reserve, etc.)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### 1. What metadata is most important

<table>
<thead>
<tr>
<th>#</th>
<th>Answer</th>
<th>Average Value</th>
<th>Standard Deviation</th>
<th>Responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Harvard collections metadata</td>
<td>52.71</td>
<td>25.39</td>
<td>134</td>
</tr>
<tr>
<td>3</td>
<td>Patron usage metadata</td>
<td>76.88</td>
<td>18.53</td>
<td>136</td>
</tr>
<tr>
<td>4</td>
<td>Harvard course usage metadata</td>
<td>80.82</td>
<td>17.87</td>
<td>136</td>
</tr>
</tbody>
</table>

### 2. Harvard collections metadata

<table>
<thead>
<tr>
<th>#</th>
<th>Answer</th>
<th>Average Value</th>
<th>Standard Deviation</th>
<th>Responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Number of Harvard libraries holding a work</td>
<td>51.58</td>
<td>24.84</td>
<td>130</td>
</tr>
<tr>
<td>1</td>
<td>The fact that a work is included in the Harvard collections</td>
<td>52.56</td>
<td>26.28</td>
<td>133</td>
</tr>
<tr>
<td>3</td>
<td>Number of copies of a work held by Harvard libraries</td>
<td>56.35</td>
<td>23.44</td>
<td>130</td>
</tr>
</tbody>
</table>
Computing StackScore at Harvard

1. Aim is to compute a single score for a bibliographic item in ILS
2. Aggregated raw result for each item
   a. Raw transaction data anonymized (patron ID’s deleted, record sequence randomized, timestamp’s hours-minutes-seconds deleted)
   b. Transaction events counted up for each usage data type globally since 2002
3. Weighting and scaling of each item
   a. Each transaction-event count is multiplied by weighting factor derived from librarian survey
      i. Course reserves and assigned texts: 0.64
      ii. Checkouts, e-downloads, recalls: 0.58
      iii. Holding libraries: 0.28
   b. Each usage data type for which there is data is summed together into a raw total score
   c. Number of distinct raw total scores computed across all items in ILS
   d. Distinct raw total scores sorted and divided into 100 evenly distributed groups for final scaled score
### StackScore Computation at Harvard

<table>
<thead>
<tr>
<th>raw transaction count x weighting</th>
<th>weighted counts</th>
<th>weighted total</th>
<th>scaled StackScore</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 faculty checkouts x 0.58</td>
<td>1.16</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 reserve placements x 0.64</td>
<td>1.28</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>2 holding libraries x 0.28</td>
<td>0.56</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
LD4L Use Case 5: Leverage Usage Data

Use Case 5.1: Research guided by community usage

Example story: As a researcher, I want to find what is being used (read, annotated, bought by libraries, etc.) by the scholarly communities not only at my institution but at others, and to find sources used elsewhere but not by my community.

This use case requires understanding of the relevant community of the user. This would require them to be authenticated and community inferred by some means/data from their identity, or for community to be specified as part of the discovery process, or for community to be inferred as part of the discovery process.

Use Case 5.2: Be guided in collection building by usage

Example story: As a librarian, I would like help building my collection by seeing what is being used by students and faculty.

Example story: As a subject librarian, I would like to see what resources in my subject area are heavily used at peer institutions but are not in my institution’s collection.

This use case is essentially a business analytics tool that would help libraries make best use of collection building activities and funds. This would be useful at both institutional or cross-institutional levels.
StackLife

1. Proof-of-concept project for ILS catalog book browsing
   a. Simulation of stacks browsing through visualization of book stacks as main context rendering technique
   b. Leveraging usage data as expressed in StackScore to drive presentation of search results and heat-mapping of book stacks
2. Adopted by Harvard University Library as complementary discovery tool into online catalog
<table>
<thead>
<tr>
<th>Title</th>
<th>Author</th>
<th>Year</th>
<th>StackScore</th>
</tr>
</thead>
<tbody>
<tr>
<td>The annotated Origin</td>
<td>Darwin, Charles, 1809-1882.</td>
<td>2009</td>
<td>53</td>
</tr>
<tr>
<td>On the origin of species</td>
<td>Darwin, Charles, 1809-1882.</td>
<td>1964</td>
<td>52</td>
</tr>
<tr>
<td>Darwin's Origin of species</td>
<td>Browne, E. J. (E. Janet), 1950-</td>
<td>2006</td>
<td>33</td>
</tr>
<tr>
<td>The origin of species by means of natural selection</td>
<td>Darwin, Charles, 1809-1882.</td>
<td>1936</td>
<td>15</td>
</tr>
<tr>
<td>Darwinism, war, and history</td>
<td>Crook, D. P. (David Paul).</td>
<td>1994</td>
<td>11</td>
</tr>
<tr>
<td>Darwin's Origin of species</td>
<td>Browne, E. J. (E. Janet), 1950-</td>
<td>2006</td>
<td>09</td>
</tr>
<tr>
<td>Genetics and the origin of species</td>
<td>Dobzhansky, Theodosius Grigorievich,</td>
<td>1951</td>
<td>08</td>
</tr>
<tr>
<td>Darwin's Origin of species</td>
<td>Browne, E. J. (E. Janet), 1950-</td>
<td>2006</td>
<td>07</td>
</tr>
<tr>
<td>Systematics and the origin of species from the viewpoint of a zoologist</td>
<td>Mayr, Ernst, 1904-2005.</td>
<td>1999</td>
<td>06</td>
</tr>
<tr>
<td>The origin of species and the voyage of the beagle</td>
<td>Darwin, Charles, 1809-1882.</td>
<td>2003</td>
<td>05</td>
</tr>
<tr>
<td>The origin of species by means of natural selection, or, The</td>
<td>Darwin, Charles, 1809-1882.</td>
<td>1982</td>
<td>05</td>
</tr>
<tr>
<td>preservation of favoured races in the struggle for life</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Metaphysics and the origin of species</td>
<td>Ghiselin, Michael T., 1939-</td>
<td>1997</td>
<td>05</td>
</tr>
</tbody>
</table>
Availability


Advanced Bibliographic Data

StackScore 53

Faculty checkouts: 6
Undergrad checkouts: 105
Graduate checkouts: 1
Holding libraries: 6
About the DPLA Bookshelf

The bookshelf is an easy way to search DPLA’s books, serials, and journals. The darker the shade of blue, the more relevant the results. Click on a spine for details and related images. Book thickness indicates the page count, and the horizontal length reflects the book’s actual height.
Haystacks

1. Proof-of-concept project for subject-related book browsing and collection-level analytics
   a. Usage data drives visualization of UI components
      i. Individual items in search results dimensioned according to their StackScore (dot size maps to magnitude of StackScore)
      ii. Subject bars dimensioned according to number of items acquired in that subject area
   b. Subject data derived from Library of Congress Classification Outline: analyzed into hierarchical data map allowing drilling down from any given class into to its immediate sub-class
HAYSTACKS A NEW WAY TO LOOK AT HARVARD'S LIBRARY

The 250 most popular items out of 12,976,734.

**AIDS**

**AUTHORS**
Not Available

**PUBLISHING DATE**
1987

**CALL NUMBER SUBJECT**
Medicine – Internal medicine – Specialties of internal medicine – Immunologic diseases – Immunodeficiency

**LIBRARY OF CONGRESS SUBJECT KEYS**

**LIBRARY OF CONGRESS CALL #**
RC607.A28 A344415

ADD THIS ITEM TO YOUR STACK
The 250 most popular items out of 11,547 about evolution.

**Evolution And The Theory Of Games**

**AUTHORS**

**PUBLISHING DATE**
1982

**CALL NUMBER SUBJECT**
Science -- Biology (General) -- Evolution -- Special aspects of the subject as a whole

**LIBRARY OF CONGRESS SUBJECT KEYWORDS**
Evolution, Mathematical models, Game theory.

**LIBRARY OF CONGRESS CALL #**
QH371.M325 1982
The 250 most popular items out of 942 about evolution, found in the Andover-Harv. Theol collection.
Usage Data Modeling

1. UsageData.owl ontology being developed to handle the following, among others
   a. Transaction type
   b. Transaction date-time
   c. Transaction patron
      i. Patron’s affiliated school
      ii. Patron’s status
   d. Transaction’s associated library
   e. Transaction’s associated internal ILS ID and barcode

2. Many usage-data specific classes and properties needed to be created from scratch

3. Heavy re-use of prov-o for collections and event handling
Vocabularies Used in UsageData.owl
A Good, Dumb Way to Learn From Libraries

Too bad we can't put to work the delicious usage data gathered by libraries.

Research libraries may not know as much as click-obsessed Amazon does about how people interact with their books. What they do know, however, reflects the behavior of a community of scholars, and it's unpolluted by commercial imperatives.

But privacy concerns have forestalled making library usage data available to application developers outside the library staff, and often even within. And the data are the definition of incompatible: Libraries collect them in different formats at different levels of granularity and at different time scales, making them hard to work with.

But suppose we could get at it. Library search engines could be tuned to what's shown itself to be relevant to their communities. Researchers could explore usage patterns over time and across disciplines, schools, geographies, and economies. Libraries could be guided in their acquisitions by what they've learned from the behavior of communities around the corner and around the globe.

We can dream, but solving the policy and technical problems intelligently would take many years and probably more will than we can muster. If only there was a big, dumb way to start putting community-usage data to work quickly.

So, here's an idea: Any library that would like to make its usage data public is encouraged to create a "stackscore" for each item in its collection. A stackscore is a number from 1 to 100 that represents how relevant an item is to the library's patrons as measured by how they've used it.
StackScore in LD4L and Beyond

1. Weaknesses
   a. Self-reinforcing loop of keeping long tail dark
   b. Different StackScore algorithms (different data and weightings) across institutions
   c. Based on incomplete usage-data profile; for example, not captured at Harvard
      i. Materials usage becoming increasingly digital => paper-based circulation statistics becoming increasingly less relevant (more true for journals than monographs)?
      ii. Non-circulating materials
      iii. No citations
   d. Compromised by apples v. oranges problem within own institution’s data
      i. Is an e-download of a journal article equivalent to a book checkout?
      ii. Is an acquisition of an extra copy of an item by a holding library similar as an expression of “engagement” to an undergraduate checkout?

2. Strengths
   a. Mitigates privacy concerns by aggregating anonymized events; no transaction-level data made available
   b. Single similarly scaled metric across institutions offers possibility of comparing usage across those institutions and relying too heavily on a single institution’s perspective on usage data

3. Too dumb, or just intelligent enough, to be useful?
1. http://stacklife.harvard.edu