Usage Data

Paul Deschner
Harvard University
Feb. 24, 2015

Usage Data at Harvard

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"

Usage Data at Harvard (continued)

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

Records	Format	Percentage
12,105,655	Book	88.2%
624,369	Serial	4.5%
188,042	Map	1.4%
154,033	Notated Music	1.1%
154,028	Sound Recording	1.1%
150,725	Manuscript	1.1%
124,699	Other	0.9%
110,361	Video/Film	0.8%
96,398	Book Part	0.7%
21,938	Collection	0.2%

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.)

Not at all Important	Very Unimportant	Neither Important nor Unimportant	Very Important	Extremely Important

0 10 20 30 40 50 60 70 80 90 100

Harvard collections metadata (such as the number of Harvard libraries holding the work; the total number of copies held by Harvard libraries; etc.)

HOLLIS catalog metadata (such as the number of times a work's record is viewed, exported, emailed, printed, etc.)

Patron usage metadata (such as the number of times a work is checked-out, recalled, requested from the Harvard depository, etc.)

Harvard course usage metadata (such as the number of times a work is required reading for a course, placed on course reserve, etc.)

1. What metadata is most important

#	Answer	Average Value	Standard Deviation	Responses
1	Harvard collections metadata	52.71	25.39	134
3	Patron usage metadata	76.88	18.53	136
4	Harvard course usage metadata	80.82	17.87	136

2. Harvard collections metadata

#	Answer	Average Value	Standard Deviation	Responses
2	Number of Harvard libraries holding a work	51.58	24.84	130
1	The fact that a work is included in the Harvard collections	52.56	26.28	133
3	Number of copies of a work held by Harvard libraries	56.35	23.44	130

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

raw transaction count x weighting	weighted counts	weighted total	scaled StackScore
2 faculty checkouts x 0.58 	1.16		
2 reserve placements x 0.64 	1.28	3	1
2 holding libraries x 0.28 	0.56		

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

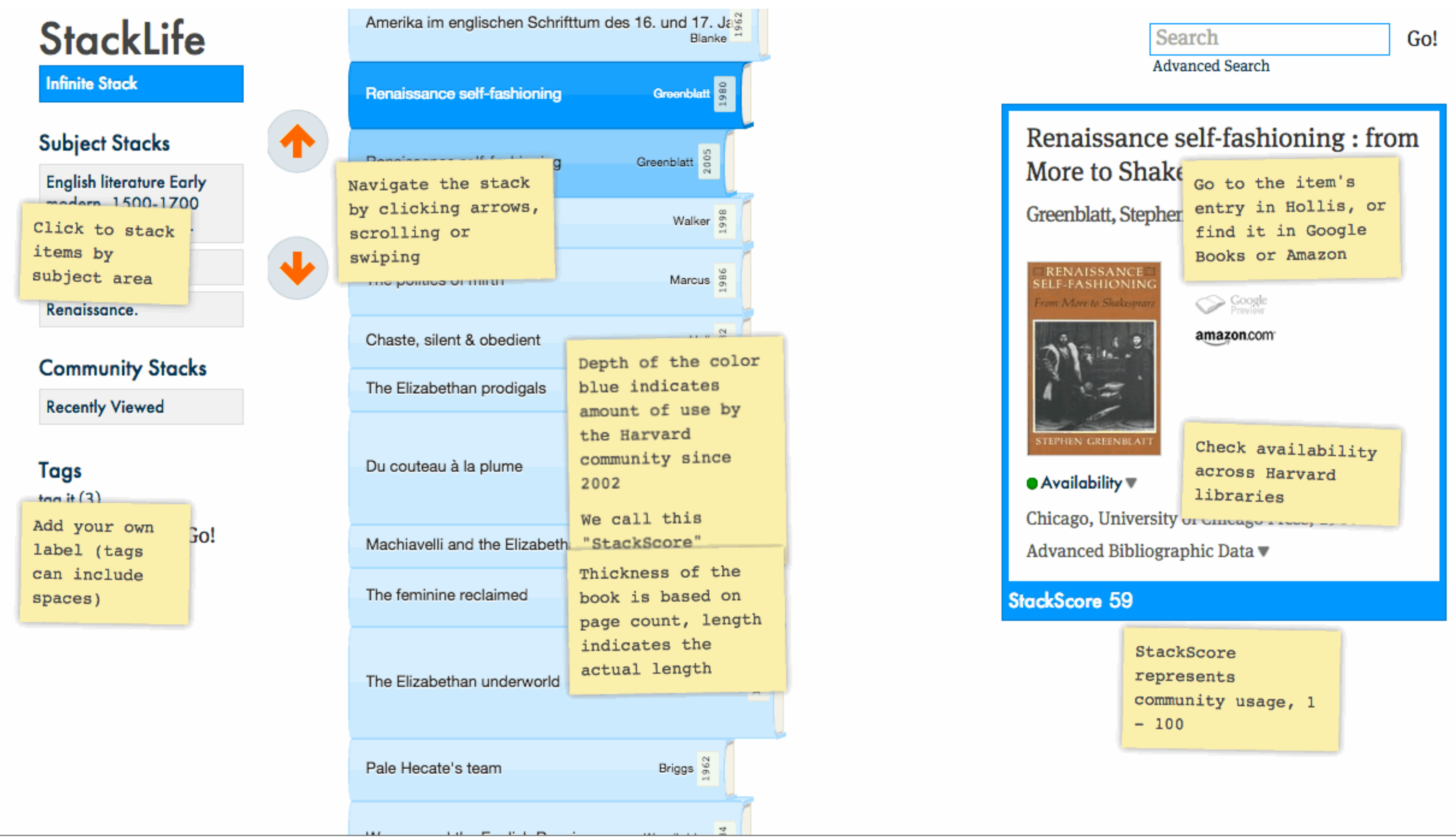
Example story: As a librarian, I would like help building my collection by seeing what is being used by students and faculty.

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


StackLife
Infinite Stack
Subject Stacks
English literature Early modern, 1500-1700
Renaissance.
Community Stacks
Recently Viewed
Tags
Click to stack items by subject area
Add your own label (tags can include spaces)
Go!
Amerika im englischen Schrifttum des 16. und 17. Ja
Blanke
1962
Renaissance self-fashioning
Greenblatt
1980
Greenblatt
2005
Walker
1998
Marcus
1986
Navigate the stack by clicking arrows, scrolling or swiping
Chaste, silent & obedient
The Elizabethan prodigals
Du couteau à la plume
Machiavelli and the Elizabeth
The feminine reclaimed
The Elizabethan underworld
Pale Hecate's team
Briggs
1962
Depth of the color blue indicates amount of use by the Harvard community since 2002
We call this "StackScore"
Thickness of the book is based on page count, length indicates the actual length
Search
Go!
Advanced Search
Renaissance self-fashioning : from More to Shake
Greenblatt, Stephen
RENAISSANCE SELF-FASHIONING
From More to Shakespeare
STEPHEN GREENBLATT
Go to the item's entry in Hollis, or find it in Google Books or Amazon
Google Preview
amazon.com
Availability
Check availability across Harvard libraries
Chicago, University
Advanced Bibliographic Data
StackScore 59
StackScore represents community usage, 1 - 100

StackLife

on the origin of species Go!
Advanced Search

Refine by StackScore
1 20 40 60 80 100

Format
- Book (350)
- Book Part (27)
- Video/Film (8)
- Manuscript (5)
- Other (3)

Library
- MCZ (144)
- WID (111)
- FIG (43)
- MED (43)
- CAB (26)
- GRA (25)
- HOU (23)
- ARN (21)
- LAM (20)
- TOZ (20)
- more

Subject

Creator

Year

Language

Showing 0 to 15 of 393 results for "on the origin of species"

Title	Author	Year	StackScore
The annotated Origin	Darwin, Charles, 1809-1882.	2009	53
On the origin of species	Darwin, Charles, 1809-1882.	1964	52
Darwin's Origin of species	Browne, E. J. (E. Janet), 1950-	2006	33
The origin of species by means of natural selection	Darwin, Charles, 1809-1882.	1936	15
Darwinism, war, and history	Crook, D. P. (David Paul).	1994	11
Darwin's Origin of species	Browne, E. J. (E. Janet), 1950-	2006	09
Genetics and the origin of species	Dobzhansky, Theodosius Grigorievich, 1900-1975.	1951	08
On the origin of species by means of natural selection	Darwin, Charles, 1809-1882.	2003	07
Darwin's Origin of species	Browne, E. J. (E. Janet), 1950-	2006	07
Systematics and the origin of species from the viewpoint of a zoologist	Mayr, Ernst, 1904-2005.	1999	06
The origin of species and the voyage of the beagle	Darwin, Charles, 1809-1882.	2003	05
The origin of species by means of natural selection, or, The preservation of favoured races in the struggle for life	Darwin, Charles, 1809-1882.	1982	05
Metaphysics and the origin of species	Ghiselin, Michael T., 1939-	1997	05

StackLife

Infinite Stack

Subject Stacks

All editions

Evolution (Biology).

Natural selection.

Community Stacks

Recently Viewed

tag it Go!

51 items

Title	Author	Year
The annotated Origin	Darwin	2009
The origin of species by means of natural s	Darwin	1936
On the origin of species by means of natural s	Darwin	2003
Darwin's Origin of species	Browne	2006
The origin of species	Darwin	1959
Negotiating Darwin	Artigas	2006
The origin of species : complete and fully illustrate	Darwin	1979
The origin of species	Darwin	1975
The origin of species by means of natural selection	Darwin	1928

Search Go!

Advanced Search

The annotated Origin : a facsimile of the first edition of On the origin of species

Darwin, Charles, 1809-1882. | Costa, James T., 1963-

HOLLIS

Google Preview

amazon.com

Availability ▼

Cambridge, Mass., Belknap Press of Harvard University Press, 2009

Advanced Bibliographic Data ▼

StackScore 53

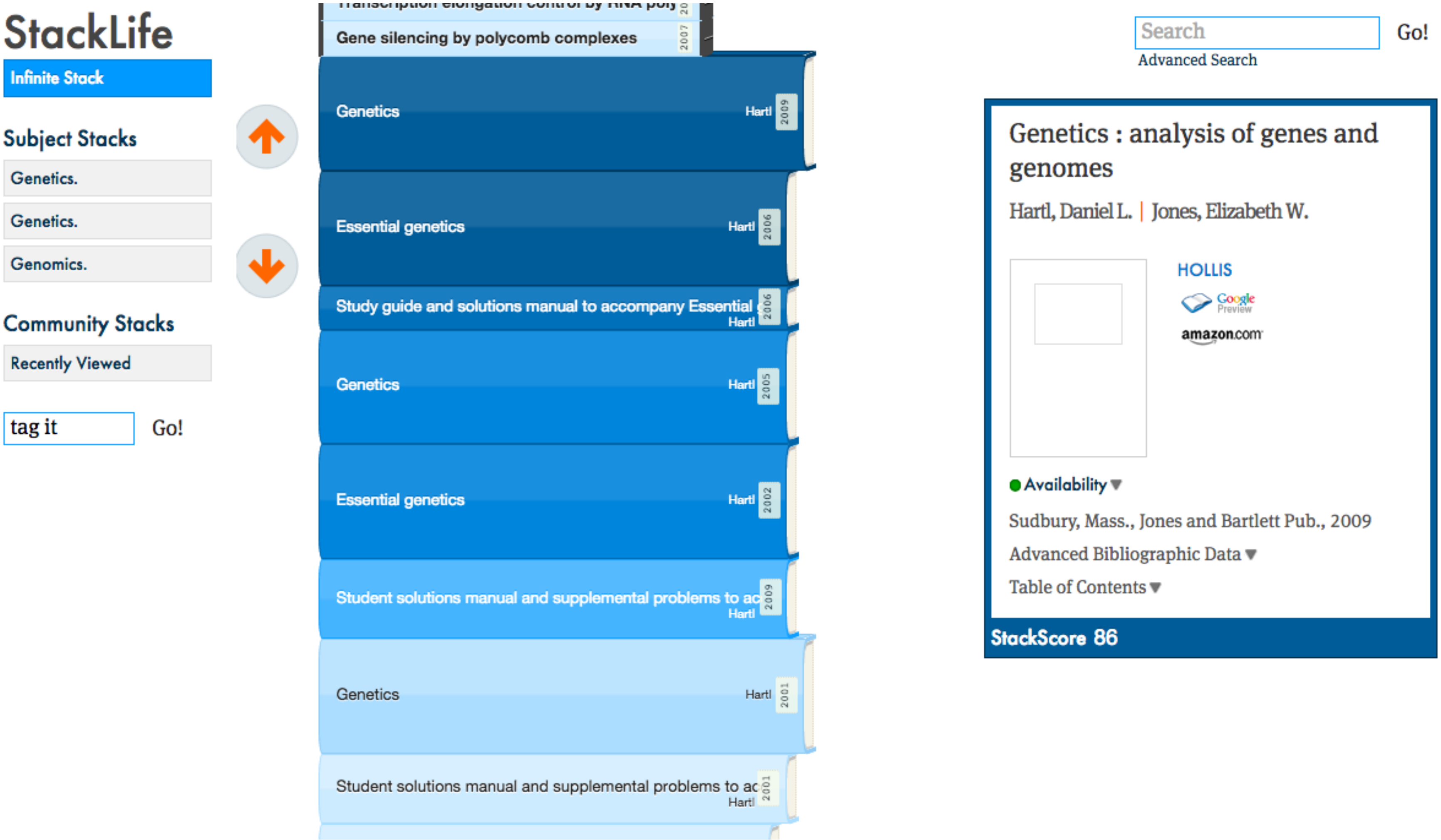
Availability

Cambridge, Mass., Belknap Press of Harvard University Press, 2009

Advanced Bibliographic Data

StackScore 53

Faculty checkouts: 6
Undergrad checkouts: 105
Graduate checkouts: 1
Holding libraries: 6


StackLife
Infinite Stack
Subject Stacks
Genetics.
Genetics.
Genomics.
Community Stacks
Recently Viewed
tag it
Go!
Gene silencing by polycomb complexes
2007
Genetics
Hartl
2009
Essential genetics
Hartl
2006
Study guide and solutions manual to accompany Essential
Hartl
2006
Genetics
Hartl
2005
Essential genetics
Hartl
2002
Student solutions manual and supplemental problems to ac
Hartl
2009
Genetics
Hartl
2001
Student solutions manual and supplemental problems to ac
Hartl
2001
Search
Go!
Advanced Search
Genetics : analysis of genes and genomes
Hartl, Daniel L. | Jones, Elizabeth W.
HOLLIS
Google Preview
amazon.com
Availability
Sudbury, Mass., Jones and Bartlett Pub., 2009
Advanced Bibliographic Data
Table of Contents
StackScore 86

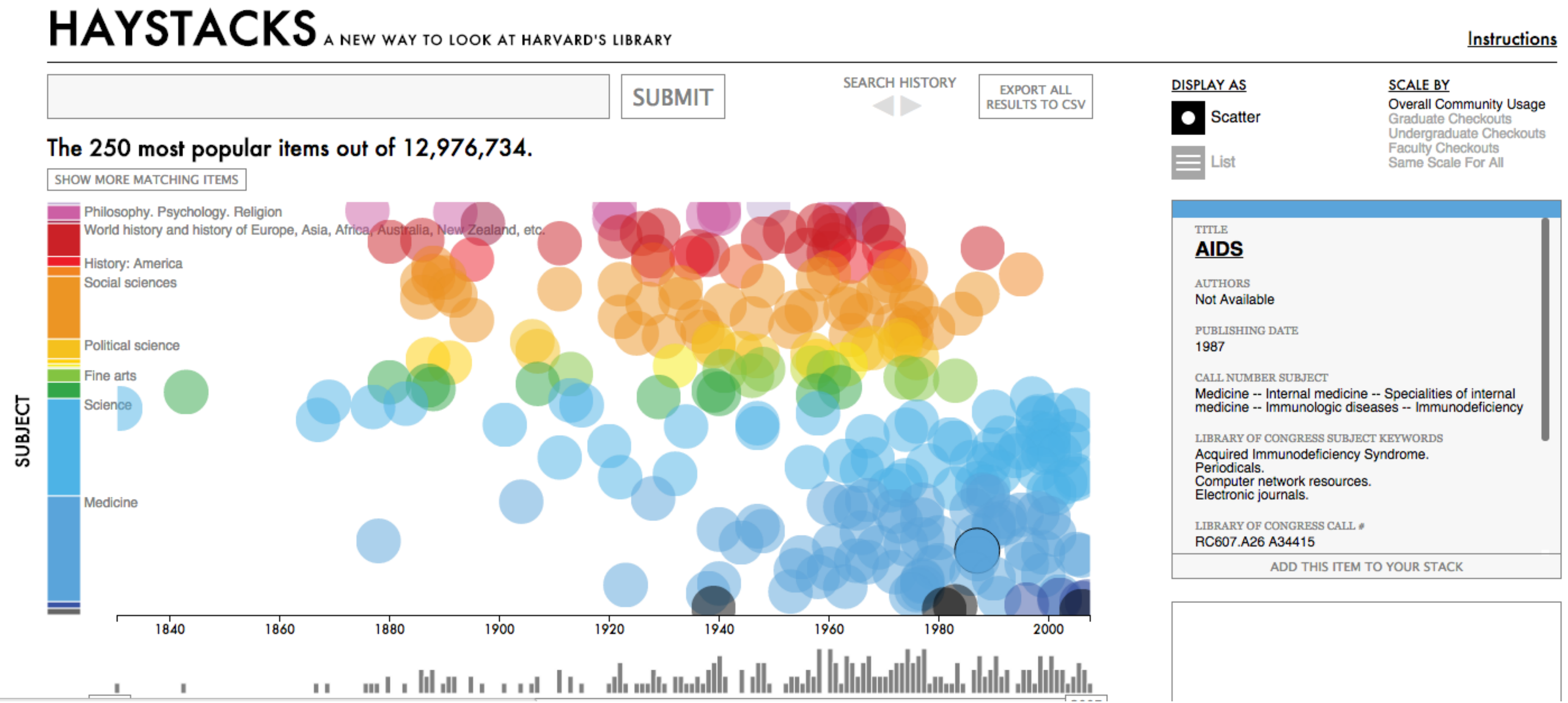
Sort by: Relevance

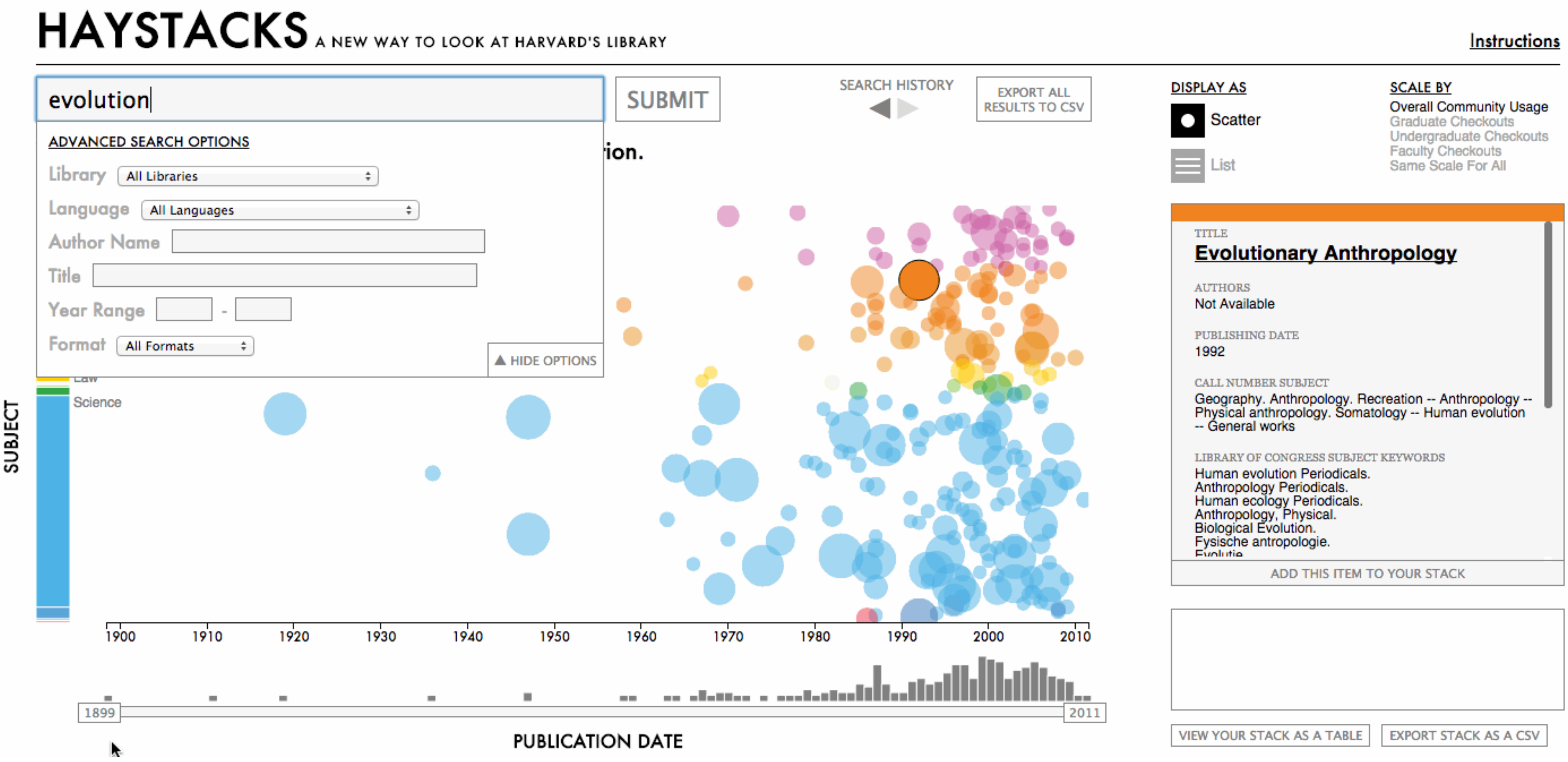
About the DPLA Bookshelf

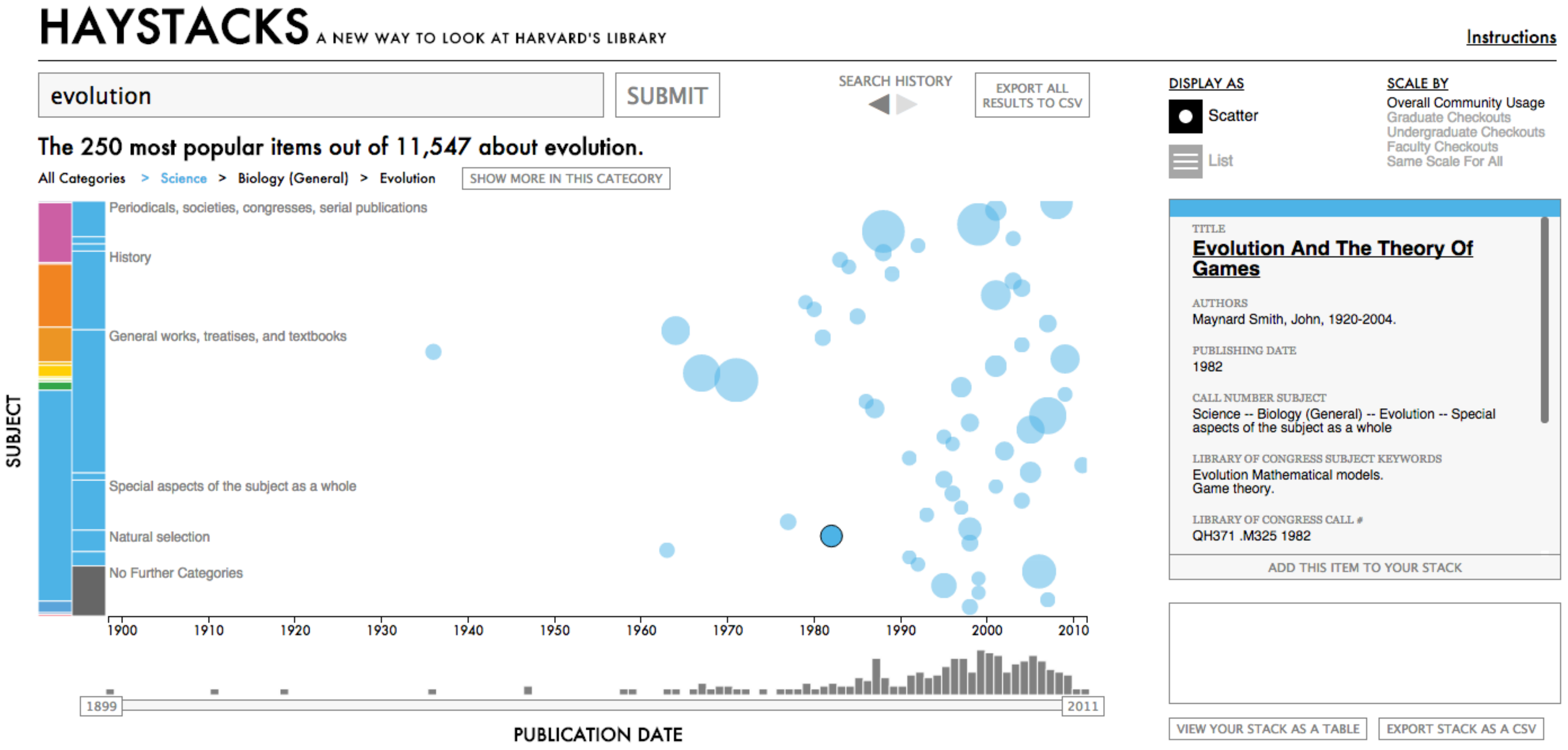
The bookshelf is an easy way to search DPLA's books, serials, and journals. The darker the shade of blue, the more relevant the results. Click on a spine for details and related images. Book thickness indicates the page count, and the horizontal length reflects the book's actual height.

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
Instructions
SUBMIT
SEARCH HISTORY
EXPORT ALL RESULTS TO CSV
DISPLAY AS
Scatter
List
SCALE BY
Overall Community Usage
Graduate Checkouts
Undergraduate Checkouts
Faculty Checkouts
Same Scale For All
The 250 most popular items out of 12,976,734.
SHOW MORE MATCHING ITEMS
SUBJECT
Philosophy. Psychology. Religion
World history and history of Europe, Asia, Africa, Australia, New Zealand, etc.
History: America
Social sciences
Political science
Fine arts
Science
Medicine
1840
1860
1880
1900
1920
1940
1960
1980
2000
TITLE
AIDS
AUTHORS
Not Available
PUBLISHING DATE
1987
CALL NUMBER SUBJECT
Medicine -- Internal medicine -- Specialities of internal medicine -- Immunologic diseases -- Immunodeficiency
LIBRARY OF CONGRESS SUBJECT KEYWORDS
Acquired Immunodeficiency Syndrome.
Periodicals.
Computer network resources.
Electronic journals.
LIBRARY OF CONGRESS CALL #
RC607.A26 A34415
ADD THIS ITEM TO YOUR STACK


HAYSTACKS A NEW WAY TO LOOK AT HARVARD'S LIBRARY
Instructions
evolution
SUBMIT
SEARCH HISTORY
EXPORT ALL RESULTS TO CSV
DISPLAY AS
Scatter
List
SCALE BY
Overall Community Usage
Graduate Checkouts
Undergraduate Checkouts
Faculty Checkouts
Same Scale For All
ADVANCED SEARCH OPTIONS
Library
All Libraries
Language
All Languages
Author Name
Title
Year Range
Format
All Formats
HIDE OPTIONS
Science
SUBJECT
1900
1910
1920
1930
1940
1950
1960
1970
1980
1990
2000
2010
1899
2011
PUBLICATION DATE
TITLE
Evolutionary Anthropology
AUTHORS
Not Available
PUBLISHING DATE
1992
CALL NUMBER SUBJECT
Geography. Anthropology. Recreation -- Anthropology -- Physical anthropology. Somatology -- Human evolution -- General works
LIBRARY OF CONGRESS SUBJECT KEYWORDS
Human evolution Periodicals.
Anthropology Periodicals.
Human ecology Periodicals.
Anthropology, Physical.
Biological Evolution.
Fysische antropologie.
ADD THIS ITEM TO YOUR STACK
VIEW YOUR STACK AS A TABLE
EXPORT STACK AS A CSV


HAYSTACKS A NEW WAY TO LOOK AT HARVARD'S LIBRARY
Instructions
evolution
SUBMIT
SEARCH HISTORY
EXPORT ALL RESULTS TO CSV
DISPLAY AS
Scatter
List
SCALE BY
Overall Community Usage
Graduate Checkouts
Undergraduate Checkouts
Faculty Checkouts
Same Scale For All
The 250 most popular items out of 11,547 about evolution.
All Categories > Science > Biology (General) > Evolution
SHOW MORE IN THIS CATEGORY
SUBJECT
Periodicals, societies, congresses, serial publications
History
General works, treatises, and textbooks
Special aspects of the subject as a whole
Natural selection
No Further Categories
1900
1910
1920
1930
1940
1950
1960
1970
1980
1990
2000
2010
1899
2011
PUBLICATION DATE
TITLE
Evolution And The Theory Of Games
AUTHORS
Maynard Smith, John, 1920-2004.
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
ADD THIS ITEM TO YOUR STACK
VIEW YOUR STACK AS A TABLE
EXPORT STACK AS A CSV

HAYSTACKS A NEW WAY TO LOOK AT HARVARD'S LIBRARY

Instructions

evolution

SUBMIT

SEARCH HISTORY

EXPORT ALL RESULTS TO CSV

DISPLAY AS

Scatter

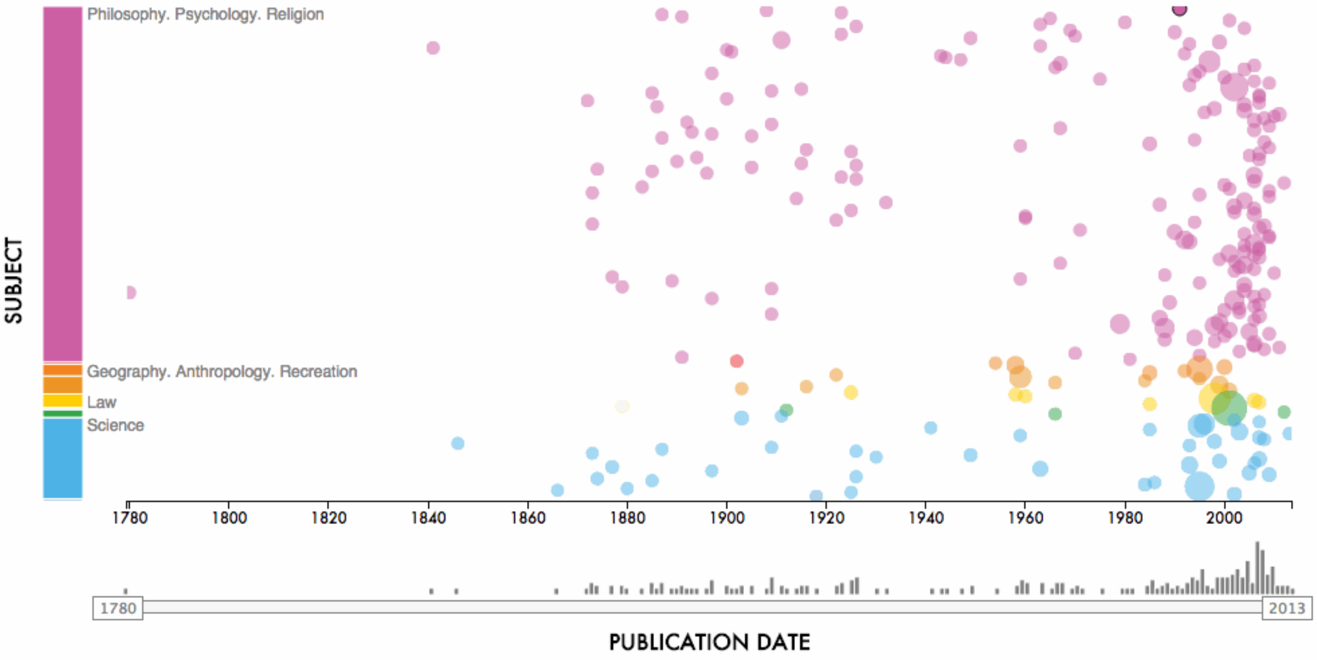
List

SCALE BY

Overall Community Usage
Graduate Checkouts
Undergraduate Checkouts
Faculty Checkouts
Same Scale For All

The 250 most popular items out of 942 about evolution, found in the Andover-Harv. Theol collection.

SHOW MORE MATCHING ITEMS



TITLE

A Blessed Rage For Order

AUTHORS

Argyros, Alex.

PUBLISHING DATE

1991

CALL NUMBER SUBJECT

Philosophy. Psychology. Religion -- Philosophy (General) -- Modern (1450/1600-) -- Special topics and schools of philosophy

LIBRARY OF CONGRESS SUBJECT KEYWORDS

Deconstruction.
Cosmology.
Evolution.
Change.
Culture Philosophy.

ADD THIS ITEM TO YOUR STACK

VIEW YOUR STACK AS A TABLE

EXPORT STACK AS A CSV

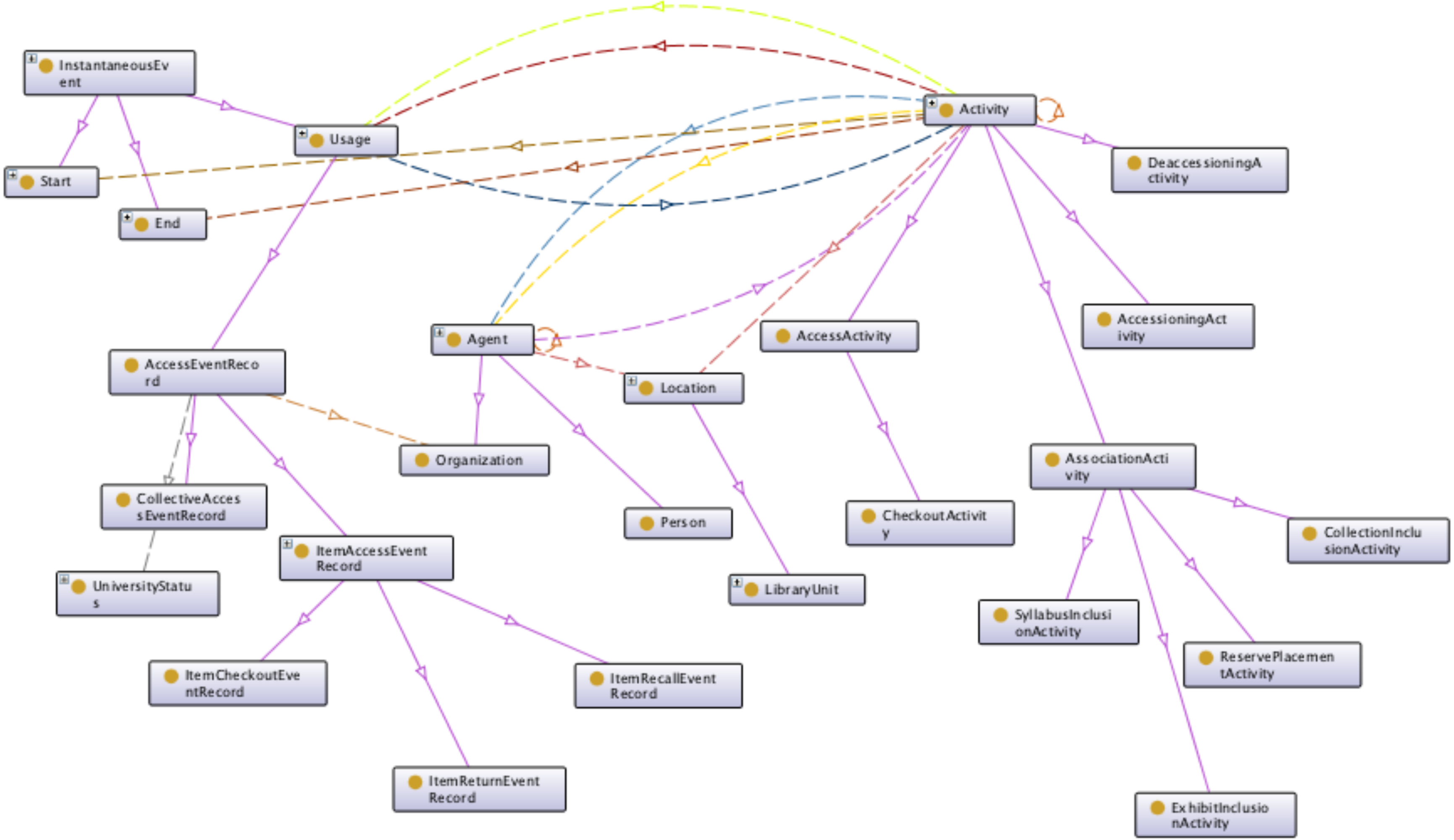
Usage Data Modeling

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

UsageData.owl ×

```
<?xml version="1.0" encoding="UTF-8"?><rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:usage="http://ld4l.org/ontology/usage#"
    xmlns:afn="http://jena.hpl.hp.com/ARQ/function#"
    xmlns:obo="http://purl.obolibrary.org/obo/"
    xmlns:owl="http://www.w3.org/2002/07/owl#"
    xmlns:dc="http://purl.org/dc/elements/1.1/"
    xmlns:provo="http://www.w3.org/ns/prov#"
    xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
    xmlns:vitro="http://vitro.mannlib.cornell.edu/ns/vitro/0.7#"
    xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#" >
```



InstantaneousEvent
Usage
Activity
Start
End
DeaccessioningActivity
Agent
AccessEventRecord
Location
AccessActivity
AccessioningActivity
Organization
CollectiveAccessEventRecord
Person
CheckoutActivity
AssociationActivity
CollectionInclusionActivity
ItemAccessEventRecord
UniversityStatus
LibraryUnit
SyllabusInclusionActivity
ItemCheckoutEventRecord
ItemRecallEventRecord
ReservePlacementActivity
ItemReturnEventRecord
ExhibitInclusionActivity


Object property hierarchy: topObjectProperty
topObjectProperty
alternateOf
specializationOf
associatedAgent
associatedWithActivity
atLocation
atLocation
downloadLocation
hadActivity
hadGeneration
hadPlan
hadRole
hadUsage
heldEntityRecord
heldItem
holdingAccessed
holdingLocation
holdingLocationRecord
image
influenced
generated
invalidated
influencer
activity
agent
entity
itemAccessRecord
locationOf
mainImage
patronAffiliation
patronUniversityStatus
qualifiedInfluence
relatedEntityInfluence
rootTab
thumbnailImage
usageRecordInActivity
wasInfluencedBy

October 7, 2014 by David Weinberger

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?

Links

1. http://stacklife.harvard.edu
2. http://haystacks.law.harvard.edu (u:haystacks, p:needles)
3. http://chronicle.com/blogs/conversation/2014/10/07/a-good-dumb-way-to-learn-from-libraries/