

How Much Does \$1.7 Billion Buy?

A Comparison of Scientific Journal Articles to Their Pre-print Versions

Research Team



Peter Broadwell

Sharon E. Farb

Todd R. Grappone

Martin Klein

@peterbroadwell

@farbthink

@liber8er

@mart1nkle1n



Global Trends in Scientific Output

- Global STM publishing market \$25.2 billion
 - 55% of STM comes from USA
 - 40% of STM is for journals (\$10 billion)
 - 68-75% is coming out of libraries' budgets







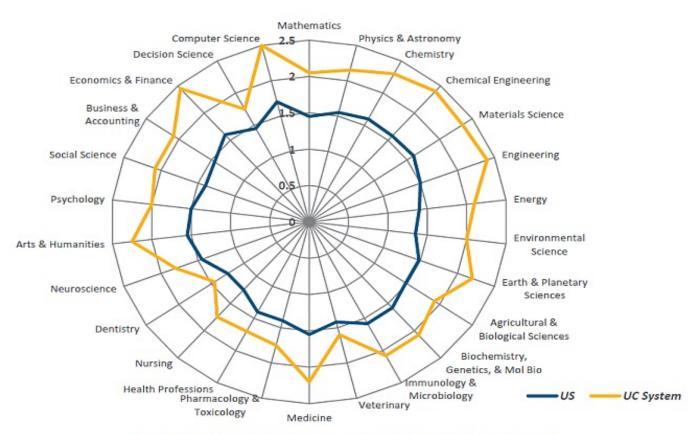


USA is the largest contributor in terms of global output of research papers: 23%





UC Publication Impact

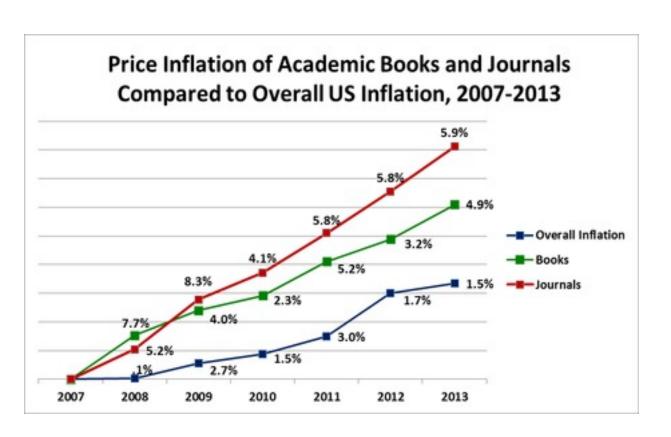


Across all disciplines, the UC FWCI average is 2.15; the U.S. FWCI average is 1.48.





What is the \$ of Knowledge?



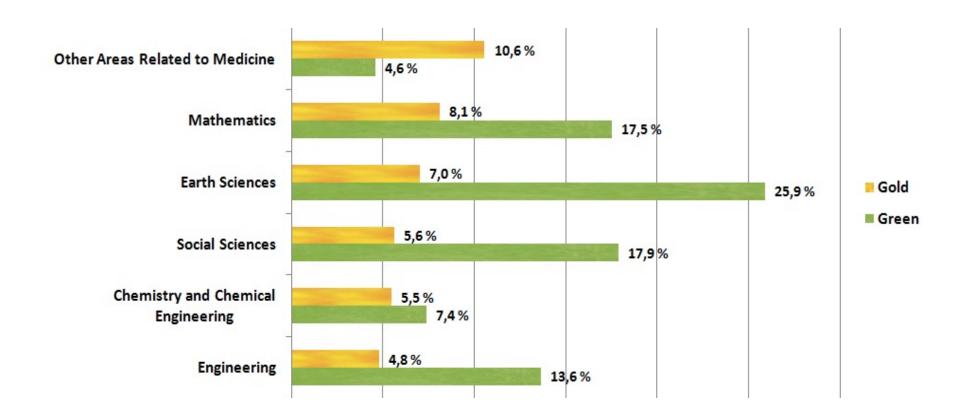
Prices Set by
Profit Maximizing
Publishers are
Determined NOT
by costs, but by
what the market
will bear.

Source: ARL Statistics, Association of Research Libraries, Washington, D.C.





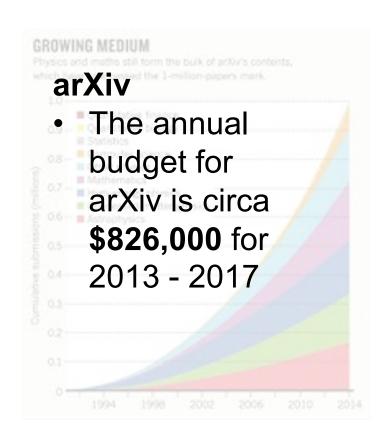
OA by Disciplines







Pre-print v. Final Published?



Final Published

English language
 STM journals:
 \$10 billion in 2013







Role of Publisher*

- Entrepreneur
- Copyediting
- Tagging
- Marketer
- Distributor
- E-Host



*STM Report 2015: An Overview of Scientific and Scholarly Publishing Michael Ware and Michael Mabe





Working Assumptions

- If the publishers' argument is valid, the text of a pre-print paper should vary significantly from its corresponding post-print version
- By applying standard similarity measures, we should be able to detect and quantify such differences.





Data Gathering

Assembling a pre-print corpus

- Source: arXiv.org
 - 1.1 million publication records
 - metadata (typical DC, including DOI) obtained via OAI-PMH interface
 - PDF versions of articles available via Amazon's S3 service (using "requester pays" option)
 - Latest version used if multiple available





Data Gathering

Finding a matching post-print corpus

- Extract DOIs from arXiv metadata
 - 44.5% or articles have DOI
- CrossRef's Metadata Search API
 - Match by DOI: download XML marked up or PDF full-text version of article
 - Metadata included in XML
 - Access based on UCLA's serial subscriptions





Data Processing

Both pre-print and post-print corpora

- Convert PDF to XML where needed
- Extract sections from XML
 - Title, authors, abstract, body, references, publication date
 - Occasionally, sections are missing
- Also try extraction from arXiv's OAI-PMH interface
 - Data provided by authors: abstracts, titles
 - Texts often have markup, complicating comparisons





Text Comparison Methods

- Length ratio
- Levenshtein ratio
- Cosine similarity
- Simhash
- Jaccard coefficient
- Sorensen similarity





Text Comparison Methods 1/3

Length ratio:

Ratio of the shorter text's length to the longer text's length

Example

"Four score and seven" → "Four score and seven years ago"

Four score and seven (20 characters)

Four score and seven years ago (30 characters)

Length ratio: 20/30 = .667



Text Comparison Methods 2/3

Levenshtein edit distance:

Number of operations (insert, delete, substitute) needed to transform one string into the other

Example

"The rose is red" → "The roose is rot"

- rose -> roose (1 insertion)
- red -> rod (1 substitution, 'o' for 'e')
- rod -> rot (1 substitution, 't' for 'd')

Levenshtein distance: 1 + 1 + 1 = 3





Text Comparison Methods 2/3

Levenshtein edit distance:

Number of operations (insert, delete, substitute) needed to transform one string into the other

Example

"The rose is red" → "The roose is rot" (edit distance: 3)

Levenshtein ratio:

<u>length of text 1 + length of text 2 – edit distance</u> length of text 1 + length of text 2





Text Comparison Methods 3/3

Cosine similarity:

Two texts have a high similarity (max value: 1) if they share many of the same words in the same proportions.

In practice, common words ("and," "the," "is") are ignored, and words that are more *characteristic* of a given text have greater importance.

Example

"QCD sum rule approach for scalar mesons as four-quark states"

"QCD sum rule approach for the light scalar mesons as four-quark states"





Text Comparison Methods 3/3

Cosine similarity:

Two texts have a high similarity (max value: 1) if they share many of the same words in the same proportions.

In practice, common words ("and," "the," "is") are ignored, and words that are more *characteristic* of a given text have greater importance.

Example

qcd sum rule approach for scalar mesons as four quark states qcd sum rule approach for the light scalar mesons as four quark states





Text Comparison Methods 3/3

Cosine similarity:

Two texts have a high similarity (max value: 1) if they share many of the same words in the same proportions.

In practice, common words ("and," "the," "is") are ignored, and words that are more *characteristic* of a given text have greater importance.

Example

qcd sum rule approach for scalar mesons as four quark states qcd sum rule approach for the light scalar mesons as four quark states

Cosine similarity: 0.8955





Analyzing News Events in Non-Traditional Digital Library Collections

Martin Klein
University of California Los Angeles
Research Library
Los Angeles, CA, USA
martinklain/Milhyany usla adu

BSTRACT

Digital liberatios are called upon to organias, aggregatio, and stream learn-digital new collections. Rabber than continuously building sides of such non-traditional collections, digital liberation are solding to manage them collections in conjunction with such other in order to provide the most value to the collection of the collection of the collection of the properties the sensitive of a prefutner of the collection of

Categories and Subject Descriptors

1. INTRODUCTION

by expaniants, aggregating, and streaming born-diguial content. As each, many liberies and archives around the worldner building digital collections from a phethern of indepenture of the content of the content of the content of the varying degrees of completeness. Digital liberais are some facing the task of making these "messy" collections as usuful as possible to relabor. This is a challenge ordinavor, and these most types of collections. The collection of and custodisastic power contents of the collection of the content of the content of the collections. The UCLA liberary has assumed as the content of the collection of the collection of the consumbed as when cutty of out-one content is a repre-

in this pure, we focus on the following two:

NewSchaper, Februion nerve coverage of the events conoldered in our experiments was analyzed via the NewSchape, and online arthres of new NewSchales recorded from called a second development of the control of the called his local development of the control of the called a second development of the control of the control of the local development of the control of the control of the American to making date of the control of the control of the date of the control of

ACDC 15, June 21–25, 2015, Karcerille, Tennessee, US Conviols is held by the connectaments. Publication of

Peter Broadwell University of California Los Angeles Research Library Los Angeles, CA, USA hroadwell@library.uda.edu

cola articles (table 1), becoming more previously due to access in technologies for coloring, interlang, and presentations on the coloring in coloring, in presentations are interlanded to the coloring in the present of the coloring in th

we recording are added to the NewSciege such day. Twitter capture The UCL'Al them; jo establishing challenges and the second of the second residual contracts and stated to current world events. This social media captudiffer currently forcess on Twitter data (twois), but an enation to portale such as VorTable and Türk's jo phane has based on available gers source tools even as Social Fe changes (SSM) [2] and tware [3] and provides enhancement to their functionally, Fe example, we have protetype in termentations in place that provide architecture between the contractions of the contraction of the contractions of the contractions in place that provide architecture of the contractions of the contraction of the contraction of the second contraction of the second contraction of the contraction of the contraction of the second contraction of the contraction of the contraction of the second contraction of the contraction of the contraction of the second contraction of the contraction of the contraction of the second contraction of the contraction of the contraction of the second contraction of the contraction of the contraction of the second contraction of the contraction of the contraction of the second contraction of the contraction of the contraction of the second contraction of the contraction of the contraction of the second contraction of the contraction of the contraction of the second contraction of the contraction of the contraction of the contraction of the second contraction of the contraction of the contraction of the second contraction of the contraction of the contraction of the second contraction of the contraction of the contraction of the second contraction of the contraction of the contraction of the second contraction of the second contraction of the contraction of the contraction of the contraction of the con

and visualization features. Uptility illimites have increasingly invested effect in notice the property of the creation of visited" subscrines, so when helding to the creation of visited" subscrines, so when helding and the creation of t

Pinner I. Program of toronto and television months

2. RELATED WORL

The ALL PATRICULAR AND ALL PATRI

3. DATA GATHERING

We collected content about the AirAsia flight 8501 air diseater, beginning on December 27th 2014, the date of the initial disappearance of the flight. To analyze television covtenther's index to search through the time-coded stacking of or per source of the contract of the contract of the code of the

4. RESULTS

We analyze our collections from three different perspectives: we consider the timeline of events, analyze the content of the collections, and approach the notion of causalith between items of both collections by investigating the sequentiality of terms that appear in one or both.

4.1 Time

his event in the two collections own time. For the NewSeys obscilent, this frequency is determined by the number of collection, the frequency is determined by the masher of severed period of time. For the Twitter collection, the first quarray is simply the amount of resets collected by our termservoir, over the same period of time, given that all order to the collection of the collection of the collection of the period of the collection of

Figure 2: Frequency of total tweets and retwee

-axis, shows all mentions or appearances of the terms on election. The Twitter series shows two significant splites, are not at alm on Evolume 28th (270,000 revents) and the second on the merzing of the 38th (180,000). Both splites consider the major seventic the plane was efficially declared sinsing at 3-41m on December 28th, and the first deletiterate resembling parts of a plane were spotted on December 18th. The frequency of tweets drops of after the 36th and, fifter a minor high point on the 31st (500 revents at same,)

forgenery series. This frequency down the reverse until an ACM of the 2 MeV. The count and a first on the 2 MeV. The count of making the count of the count of the count of the count of the 2 MeV. The count of the

who had perturbed bickets [1].

who had perturbed bickets [1].

of the seath terms on theirwise are much boars than
values of twee the despite of "statistical" paid to the scale
terms on a decision now the intrinsic paid to the scale
terms on a decision now the memory of machine
states and the scale of the scale of the scale of the
scale of the scale of the scale of the scale of the
scale of the scale of the scale of the scale of the
scale of the scale of the scale of the
scale of the scale of the scale of the
scale of the scale of the scale of the
scale of the scale of the
scale of the scale of the
scale of the scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of the
scale of

4.2 Content

The total amount of tweets collected around this even (more that his value does not translate to the same number c distinct news items. A big part of Twitter are "retweets" an



Table 1: Terms that occur in both collections an

where the state [5] have indicated but users send reveals are being with a finite of stage over. The present deployed in Figure 2. The blue are too generate the two deployed in Figure 2. The blue are too generate the two contracts of the stage of the stage of the stage of the region of the stage of the region of the regio

frequently in tweets (high Twitter rank) but are sometimed much loss frequently on television (hower TV rank). We observed two patterns here. First, there are terms that are as the exact flight number QUEDO (derived no seen smolly referred to it as fright SUO) and references to journalistic information sorrows such as restrate, Soconday, we found many more emotional supersoine in the Twitter collection, to the two controls of the control of the contro



Figure 3: Segmentiality of terms in tweets and on televisic

3 Segmentiality

series the collection. This delives we is desirely cover be desired to the collection of the legislate of an Hip control in the collection of terms for the legislate of an Hip control in the collection of terms for the manner of terms that the collection of terms for the collection. The collection of the collection of the collection of the legislate of an Hip control in the collection of terms for the manner of terms and the collection of the collection of the terms are mentioned in Texture fixed. The latest proper size and 1 theory, which likely is the to the deep legislate of 1 theory, which likely is that to the deep legislate of 1 theory, which likely is that to the collection of 1 theory, which likely is that to the collection of the deceleration of the collection of the collection of the collection of 1 theory, which likely is that to the collection of the theories of the collection of the collection of the collection of the collection of 1 the collection of the collection of the collection of the collection of 1 the collection of t

CONCLUSIONS

In this paper, we present results of our prelimina analyzing news events in non-traditional collection of their ions were breedenst and the side desired with a control male, garper formers. It had a feel for all the coverage of a record air disouter. Our intensigation procedur includes that this most of collections can vary demandtical theory of the control of the control of the control of the desired of the collection. For example, we showed to disturce other collections. For example, we showed the same center compared to rectifue soles, Sighia in "stream's are seen exampled to rectifue soles, Sighia in "stream's the testings depressing showers." The collection coverage, on the other hand, much to continue at a high breat for the collection of the control of the collection of the processing of the collection of the collection of the garden of the collection of the collection of the garden of the collection of the collection of the garden of the collection of the collection of the garden of g

DEFEDENCES

- Holidays thrown into chaos after AirAsia cancels dis Bali flights.
 http://www.theage.com.au/victoris/holidays-
- bali-flights-20141227-12eac5.html. | Social Feed Manager. http://gwu-
- What happened to AirAsia Flight QZ8501: Your questions answered. http://www.cnn.com/2014/12/28
- world/ania/airania-questions-answers/.
 F. Chierichetti, J. Kleinberg, R. Kumar, M. Mahdian, and S. Pandey. Event detection via communication nottern analysis. In Proceedings of ICWSWILL 2014.
- A. Popany, A. Khan, J. Palou, M. Cawlick, S. Wong, and T. Gallisti. Social Media Analytics Twitter as Newseurce during the Beston Marathen Bornbings. http://dx.doi.org/10.6064/sp.figabare.1299123. F. Steen. The NewsCcape Project: Understanding the
- r. Stein. Lie Newscap Project: Understanding the Media through a Multimodal Perspective. Technical report, 2014. https://idre.ucla.edu/featured/newsmcape-
- W. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yai and X. Li. Comparing twitter and traditional media using topic models. In *Proceedings of ECIR'11*, pages 338–349, 2011.





Title

Analyzing News Events in Non-Traditional **Digital Library Collections**

Martin Klein University of California Los Angeles Research Library Los Angeles, CA, USA martinklein@library.ucla.edu

ABSTRACT

Digital libraries are called upon to organize, aggregate, and steward born-digital news collections. Rather than continuously building silos of such non-traditional collections, digi-tal libraries are seeking to manage these collections in conjunction with each other in order to provide the most value to scholars. We here present the results of a preliminary study analyzing characteristics of items in two collections of digital news media: television broadcasts and social media digital news media, television broadcasts and social media coverage. Our findings indicate a number of factors that similar efforts will need to take into consideration when linking digital "news" collections similar to ours.

Categories and Subject Descriptors

H.3.7 [Digital Libraries]: Collection

1 INTRODUCTION

The role of memory organizations is increasingly defined by organizing, aggregating, and stewarding born-digital con-tent. As such, many libraries and archives around the world are building digital collections from a plethora of indepen dent sources, often diverse in format, disconnected, and of varying degrees of completeness. Digital libraries are now facing the task of making these "messy" collections as useful as possible to scholars. This is a challenging endeavor, and traditional analytical library tools may not be applicable to these novel types of collections. The collection of and custodianship over contemporary digital news content is a repre-sentative example of the issues entailed in the management of such non-traditional collections. The UCLA Library has assembled a wide variety of such non-traditional collections; in this paper, we focus on the following two:

NewsScape: Television news coverage of the events considered in our experiments was analyzed via the NewsScape, an online archive of news broadcasts recorded from cable and broadcast news networks in the United States as well as local television markets and international news sources Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed classroom use is granted without for provided that copies are not made or distributed for profit or commercial advantage and that copies board this notice and the full citation on the fast page. Copyrights for components of this work owned by others than ACM man the honored. Abstracting with refers in permitted. To copy deriveis, or republish, to post on servers or to redistribute to lists, requires prior specific permission andwar a fize. Request permission from memorisment the modern a fize. Request permission from memorisment the modern a fize department of the commercial control and the commercial control and the control and the commercial control and the co

ACM 978-1-4503-3594-2/15/06 ...\$15.00. http://dy.doi.org/10.1145/2756406-2756948

Peter Broadwell University of California Los Angeles Research Library Los Angeles, CA, USA broadwell@library.ucla.edu

[7]. The NewsScape represents an emergent type of multi-media archive that is becoming more prevalent due to advances in technologies for collecting, indexing, and present ing large amounts of multimedia data. As of this writing, the NewsScape stores and provides streaming access for research and instructional purposes to more than 244,000 hours of television news programs from 2005 to the present, recorded from 38 networks. The audio stream is indexed chronologically via its time-coded closed-caption texts, as well as official program transcriptions that are aligned with the caption texts. Words that appear on screen, such as in a storyrelated graphic or a rotating "news crawl" at the bottom of the screen, also are identified by the NewsScape's automatio optical character recognition tool and are tagged with their time of appearance as well as their location on screen. The more than 2.9 billion words thereby extracted from these programs are made searchable via an Apache Solr index, which is augmented with various official and computation ally derived program-related metadata. Approximately 135 new recordings are added to the NewsScape each day. Twitter capture: The UCLA library is establishing a

framework to build collections of social media news coverage related to current world events. This social media capture effort currently focuses on Twitter data (tweets), but an expansion to portals such as YouTube and Flickr is planned for the near future. The social media collection framework is based on available open source tools such as Social Feed Manager (SFM) [2] and twarc [3] and provides enhancements to their functionality. For example, we have prototype implementations in place that provide archival functions to eserve embedded resources, as well as interactive real-time data visualization features.

Digital libraries have increasingly invested effort in avoid ing the creation of "siloed" collections, so when building non-traditional news collections that are focused upon the ame event, advanced analytical methods are required to help recognize, analyze, and (temporally) correlate different accounts of the event. In this paper, we address this chal-lenge and provide evidence of how such analytical methods should work. We do not present mature solutions and do not nary findings which, given the nature of our two collections concern a scenario that to the best of our knowledge has not been considered closely before. We provide a number of indicators of what similar efforts will need to take into consideration when linking news collections that have similar characteristics to those described here.

Analyzing News Events in Non-Traditional **Digital Library Collections**

Martin Klein University of California Los Angeles Research Library Los Angeles, CA, USA martinklein@library.ucla.edu

ABSTRACT

Digital libraries are called upon to organize, aggregate, and steward born-digital news collections. Rather than continuously building silos of such non-traditional collections, digi-tal libraries are seeking to manage these collections in conjunction with each other in order to provide the most value to scholars. We here present the results of a preliminary study analyzing characteristics of items in two collections of digital news media: television broadcasts and social media digital news media, television broadcasts and social media coverage. Our findings indicate a number of factors that similar efforts will need to take into consideration when linking digital "news" collections similar to ours.

Categories and Subject Descriptors

H.3.7 [Digital Libraries]: Collection

1 INTRODUCTION

The role of memory organizations is increasingly defined by organizing, aggregating, and stewarding born-digital con-tent. As such, many libraries and archives around the world are building digital collections from a plethora of independent sources, often diverse in format, disconnected, and of varying degrees of completeness. Digital libraries are now facing the task of making these "messy" collections as useful as possible to scholars. This is a challenging endeavor, and traditional analytical library tools may not be applicable to these novel types of collections. The collection of and custodianship over contemporary digital news content is a representative example of the issues entailed in the management of such non-traditional collections. The UCLA Library has assembled a wide variety of such non-traditional collections; in this paper, we focus on the following two:

NewsScape: Television news coverage of the events considered in our experiments was analyzed via the NewsScape, an online archive of news broadcasts recorded from cable and broadcast news networks in the United States as well as local television markets and international news sources Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the fall cluiton on the first page. Copyrights for components of this work owned by others than ACM must be homeour. Advantage with recedit is permitted. To copy otherwise, or re-publish, to post on servers or to reflar/falter to lists, requires prior specific permissions made or a fee. Recept permissions from permissions 9000 mp permissions 9000 mp. or modern for permissions 9000 mp. or permissio artator a rec. recipies permissions from permissions warm.org. JCDL'15, June 21–25, 2015, Knoxville, Tennessee, USA. Conviright is held by the owner/author(s). Publication rights licensed to ACM.

ACM 978-1-4503-3594-2/15/06 ...\$15.00. http://dy.doi.org/10.1145/2756406-2756948

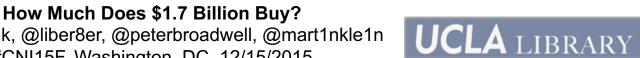
Peter Broadwell University of California Los Angeles Research Library Los Angeles, CA, USA broadwell@library.ucla.edu

The NewsScape represents an emergent type of multi-lia archive that is becoming more prevalent due to adces in technologies for collecting, indexing, and present ing large amounts of multimedia data. As of this writing, the NewsScape stores and provides streaming access for research and instructional purposes to more than 244,000 hours of elevision news programs from 2005 to the present, recorded from 38 networks. The audio stream is indexed chronologially via its time-coded closed-caption texts, as well as official program transcriptions that are aligned with the caption texts. Words that appear on screen, such as in a storydated graphic or a rotating "news crawl" at the bottom of he screen, also are identified by the NewsScape's automatic ptical character recognition tool and are tagged with their time of appearance as well as their location on screen. The more than 2.9 billion words thereby extracted from these programs are made searchable via an Apache Solr index, which is augmented with various official and computation ally derived program-related metadata. Approximately 135 new recordings are added to the NewsScape each day. Twitter capture: The UCLA library is establishing a

framework to build collections of social media news coverage related to current world events. This social media capture effort currently focuses on Twitter data (tweets), but an expansion to portals such as YouTube and Flickr is planned for the near future. The social media collection framework is based on available open source tools such as Social Feed Manager (SFM) [2] and twarc [3] and provides enhancements to their functionality. For example, we have prototype implementations in place that provide archival functions to preserve embedded resources, as well as interactive real-time data visualization features.

Digital libraries have increasingly invested effort in avoid ing the creation of "siloed" collections, so when building non-traditional news collections that are focused upon the same event, advanced analytical methods are required to help recognize, analyze, and (temporally) correlate different counts of the event. In this paper, we address this challenge and provide evidence of how such analytical methods should work. We do not present mature solutions and do not nary findings which, given the nature of our two collections, concern a scenario that to the best of our knowledge has not been considered closely before. We provide a number of indicators of what similar efforts will need to take into consideration when linking news collections that have similar characteristics to those described here.

Abstract

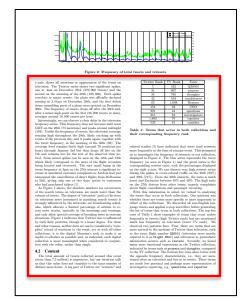


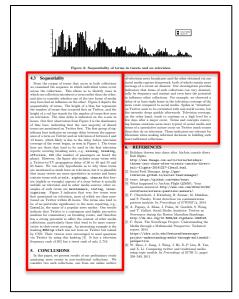


Body













(Authors)

Analyzing News Events in Non-Traditional **Digital Library Collections**

University of California Los Angeles Research Library Los Angeles, CA, USA martinklein@library.ucla.edu

Peter Broadwell University of California Los Angeles Research Library Los Angeles, CA, USA broadwell@library.ucla.edu

ABSTRACT

Digital libraries are called upon to organize, aggregate, and steward born-digital news collections. Rather than continuously building silos of such non-traditional collections, digi-tal libraries are seeking to manage these collections in conjunction with each other in order to provide the most value to scholars. We here present the results of a preliminary study analyzing characteristics of items in two collections of digital news media: television broadcasts and social media coverage. Our findings indicate a number of factors that similar efforts will need to take into consideration when linking digital "news" collections similar to ours.

Categories and Subject Descriptors

H.3.7 [Digital Libraries]: Collection

1 INTRODUCTION

The role of memory organizations is increasingly defined by organizing, aggregating, and stewarding born-digital con-tent. As such, many libraries and archives around the world are building digital collections from a plethora of indepen dent sources, often diverse in format, disconnected, and of varying degrees of completeness. Digital libraries are now facing the task of making these "messy" collections as useful as possible to scholars. This is a challenging endeavor, and traditional analytical library tools may not be applicable to these novel types of collections. The collection of and custodianship over contemporary digital news content is a repre-sentative example of the issues entailed in the management of such non-traditional collections. The UCLA Library has assembled a wide variety of such non-traditional collections; in this paper, we focus on the following two:

NewsScape: Television news coverage of the events considered in our experiments was analyzed via the NewsScape, an online archive of news broadcasts recorded from cable and broadcast news networks in the United States as well as local television markets and international news sources Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear alls notice and the full clinic from four commercial advantage and that copies bear alls notice and the full clinic tion on the first page. Copyrights for components of this work owned by others than CAM must be honored. Abstracting with reclassif in permitted. To copy derwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a face. Request permission from another a face. Request permission from another a face. Request permission from a more interesting that the commercial permission from the commercial permis

ACM 978-1-4503-3594-2/15/06 ...\$15.00. http://dy.doi.org/10.1145/2756406-2756948

[7]. The NewsScape represents an emergent type of multi-media archive that is becoming more prevalent due to advances in technologies for collecting, indexing, and present ing large amounts of multimedia data. As of this writing, the NewsScape stores and provides streaming access for research and instructional purposes to more than 244,000 hours of television news programs from 2005 to the present, recorded from 38 networks. The audio stream is indexed chronologically via its time-coded closed-caption texts, as well as official program transcriptions that are aligned with the caption texts. Words that appear on screen, such as in a storyrelated graphic or a rotating "news crawl" at the bottom of the screen, also are identified by the NewsScape's automatic optical character recognition tool and are tagged with their time of appearance as well as their location on screen. The more than 2.9 billion words thereby extracted from these programs are made searchable via an Apache Solr index, which is augmented with various official and computation

ally derived program-related metadata. Approximately 135 new recordings are added to the NewsScape each day. Twitter capture: The UCLA library is establishing a framework to build collections of social media news coverage related to current world events. This social media capture effort currently focuses on Twitter data (tweets), but an expansion to portals such as YouTube and Flickr is planned for the near future. The social media collection framework is based on available open source tools such as Social Feed Manager (SFM) [2] and twarc [3] and provides enhancements to their functionality. For example, we have prototype implementations in place that provide archival functions to eserve embedded resources, as well as interactive real-time data visualization features.

Digital libraries have increasingly invested effort in avoid ing the creation of "siloed" collections, so when building non-traditional news collections that are focused upon the ame event, advanced analytical methods are required to help recognize, analyze, and (temporally) correlate different accounts of the event. In this paper, we address this chal-lenge and provide evidence of how such analytical methods should work. We do not present mature solutions and do not nary findings which, given the nature of our two collections concern a scenario that to the best of our knowledge has not been considered closely before. We provide a number of indicators of what similar efforts will need to take into consideration when linking news collections that have similar characteristics to those described here.

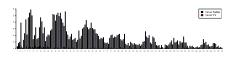


Figure 3: Sequentiality of terms in tweets and on television

4.3 Sequentiality

From the corpus of terms that occur in both collections we examined the sequence in which individual terms occur across the collections. This allows us to identify cases in which one collection introduces a term earlier than the other and also to consider whether one of the two forms of media may have had an influence on the other. Figure 3 depicts the sequentiality of terms. The height of a blue bar represents the number of terms that occurred first on Twitter, and the height of a red bar stands for the number of terms first seen on television. The time delta is indicated on the x-axis in ours. Our first observation from Figure 3 is the dominance of blue bars, indicating that the vast majority of shared terms are mentioned on Twitter first. The first group of significant bars indicates an average delay between the appear ance of a term on Twitter and on television of between 4 and 13 hours, which likely is due to the delay before television coverage of the event began, as seen in Figure 1. The terms here are those that tend to be used in the first television reports covering breaking news, e.g., missing, breaking, officials, 155 (the number of passengers on board the plane). However, the figure also includes many terms with a Twitter-to-TV propagation delay of 30 to 40 and 55 and 65 hours. We can only hypothesize as to why these terms are mentioned so much later on television, but it is plausible that many tweets are more speculative in nature and hence contain terms such as tale, sightings, chances that fore see (rightly or wrongly) aspects of a story before it actually unfolds on television and in other media sources; other ex amples of such terms are maintenance, routing, investigations. Figure 3 indicates that very few terms were first mentioned on television, most of which are then men-tioned on Twitter within 26 hours. The terms also tend to be of no particular significance to the news reporting, e.g. Costello, the name of a popular news anchor. Our results indicate that Twitter is a continuous and highly pro-active medium for commentary on breaking events, and therefor has a strong potential to affect the content of other media collections, particularly those that tend to be more conservative in their event coverage. An interesting example is the hashtag #8501qs which was not born on Twitter but coined by CNN. Their viewers were encouraged to send questions via Twitter by using this hashtag [4]. It has a television frequency rank of 203 but a tweet rank of only 2,716.

5. CONCLUSIONS

In this paper, we present results of our preliminary study analyzing news events in non-traditional collections. We consider two such collections, one from our online archive

social media capture framework, both of which contain news coverage of a recent air disaster. Our investigation provides indicators that items of such collections can vary dramatically by frequency and content and even have the potential to influence other collections. For example, we observed a delay of at least eight hours in the television coverage of the news event compared to social media. Spikes in "attention on Twitter seem to be correlated with real-world events, but the intensity drops quickly afterwards. Television coverage, on the other hand, tends to continue at a high level for a few days after a major event. Terms and concepts convey-ing human emotions seem more typical of social media and terms of a speculative nature occur on Twitter much sooner than they do on television. These indicators are relevant for librarians when making informed decisions in building such

6. REFERENCES

Holidays thrown into chaos after AirAsia cancels direct

http://www.theage.com.au/victoria/holidaysbali-flights-20141227-12eac5.html. Social Feed Manager. http://gwu-

libraries.github.io/social-feed-manager, twarc https://github.com/edsu/twarc

What happened to AirAsia Flight QZ8501: Your questions answered. http://www.cnn.com/2014/12/29/ world/asia/airasia-questions-answers/. F. Chierichetti, J. Kleinberg, R. Kumar, M. Mahdian,

and S. Pandey. Event detection via communipattern analysis. In Proceedings of ICWSM'14, 2014. 6 A. Pogany, A. Khan, J. Palau, M. Gawlick, S. Wong, and T. Gallati. Social Media Analytics: Twitter as

Newsource during the Boston Marathon Bombings. http://dx.doi.org/10.6084/m9.figshare.1298123. F. Steen. The NewsScape Project: Understanding the Media through a Multimodal Perspective. Technical report, 2014.

https://idre.ucla.edu/featured/newsscape perspective.

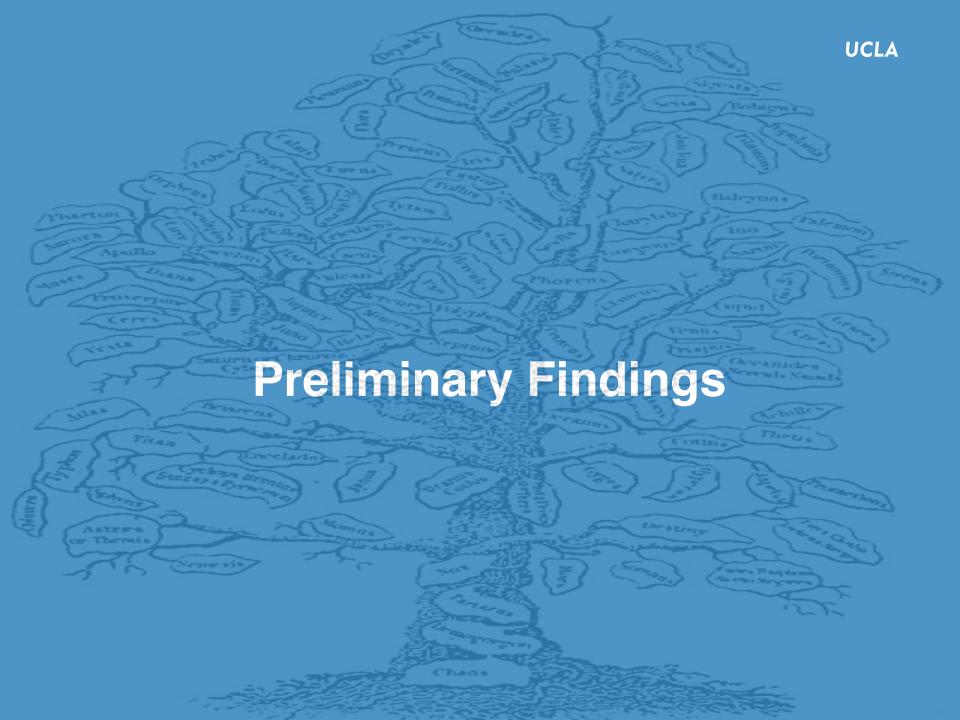
Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li. Comparing twitter and traditional media ising topic models. In Proceedings of ECIR'11, pages

References (future work)









Data Gathering Results

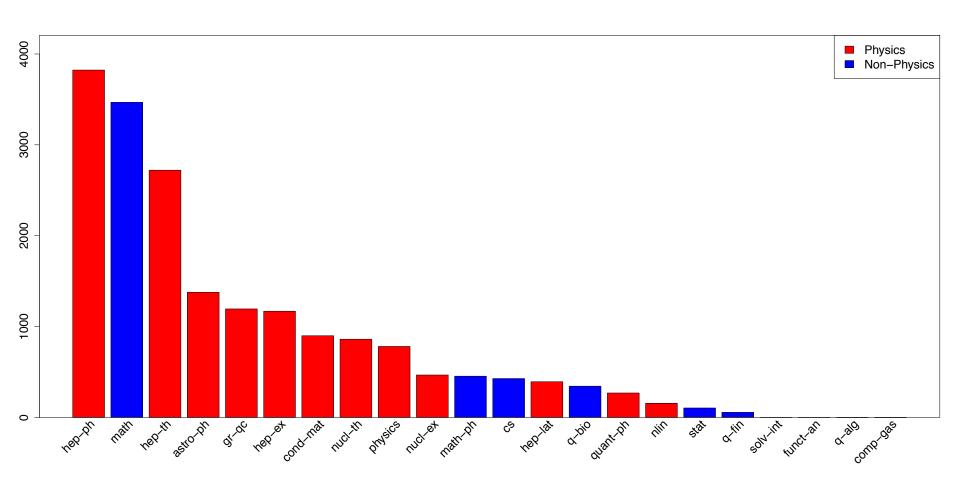
Both pre-print and post-print corpora

- 11,017 full text articles matched by DOI
- Most papers within date range 2003 2015
- 96% of post-prints published by Elsevier
- "Physics Letters B" top journal





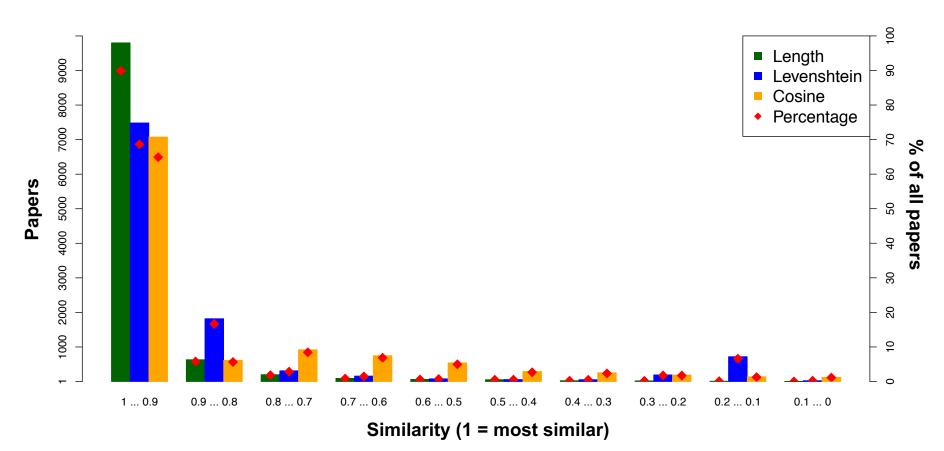
Number of Papers by Category (arXiv)







Title Comparison

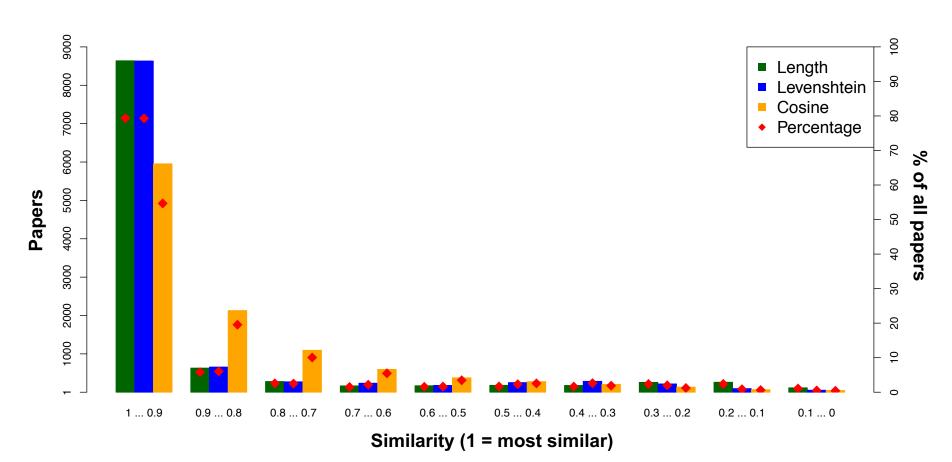


Browse findings at http://sologlo.library.ucla.edu/prepost





Abstract Comparison

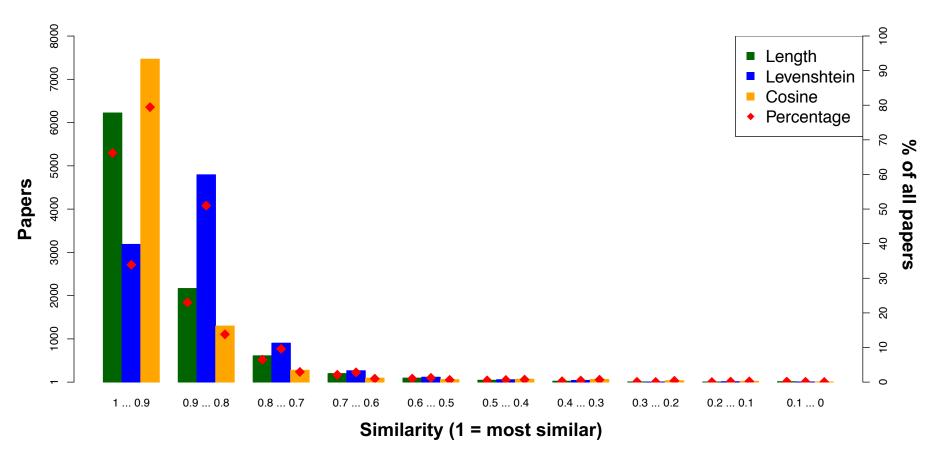


Browse findings at http://sologlo.library.ucla.edu/prepost





Body Comparison



Browse findings at http://sologlo.library.ucla.edu/prepost





10.1016/j.physletb.2006.10.068 Physics Letters B

Saturation physics at HERA and RHIC: An unified description

Saturation physics at HERA and RHIC: An unified description

ABSTRACT

Length sim: 0.9822 | Cosine sim: 0.9808 | Levenshtein ratio: 0.9876 | Levenshtein edit dist: 25 | Jaccard coefficient: 0.9773 | Sorensen sim: 0.9885

One of the frontiers of QCD which are intensely investigated in high energy experiments is the high energy (small x) regime, where we expect to observe the non-linear behavior of the theory. In this regime, the growth of the parton distribution should saturate, forming a Color Glass Condensate (CGC). In fact, signals of parton saturation have already been observed both in ep deep inelastic scattering at HERA and in deuteron-gold collisions at RHIC. Currently, a global description of the existing experimental data is possible considering different phenomenological saturation models for the two processes within the CGC formalism. In this letter we analyze the universality of these dipole cross section parameterizations and verify that they are not able to describe the HERA and RHIC data simultaneously. We analyze possible improvements in the parameterizations and propose a new parametrization for the forward dipole amplitude which allows us to describe quite well the small-x ep HERA data on F 2 structure function as well as the dAu RHIC data on charged hadron spectra. It is an important signature of the universality of the saturation physics.

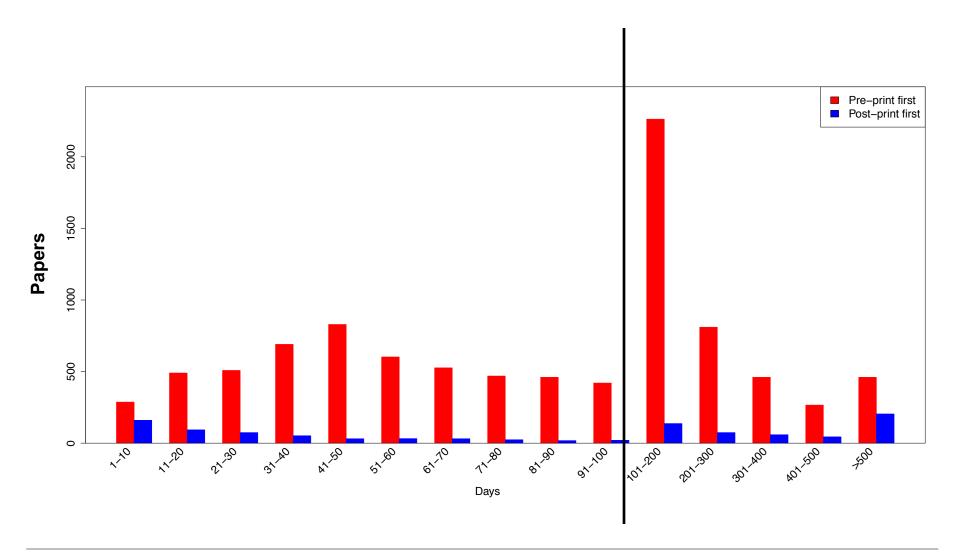
Abstract One of the frontiers of QCD which are intensely investigated in high energy experiments is the high energy (small x) regime, where we expect to observe the non-linear behavior of the theory. In this regime, the growth of the parton distribution should saturate, forming a color glass condensate (CGC). In fact, signals of parton saturation have already been observed both in ep deep inelastic scattering at HERA and in deuterongold collisions at RHIC. Currently, a global description of the existing experimental data is possible considering different phenomenological saturation models for the two processes within the CGC formalism. In this Letter we analyze the universality of these dipole cross section parameterizations and verify that they are not able to describe the HERA and RHIC data simultaneously. We analyze possible improvements in the parameterizations and propose a new parameterization for the forward dipole amplitude which allows us to describe quite well the small- x ep HERA data on F 2 structure function as well as the dAu RHIC data on charged hadron spectra. It is an important signature of the universality of the saturation physics.







Publication Dates







Author Similarity (very preliminary)

- Not trivial to determine due to
 - Different name formatting
 - Varying metadata quality

"Save" statement: **7,731/9,935** articles (**78%**) have identical author lists (number and ranks of authors)









More to Come!

- Refine extraction/comparison of authors and references
- Overlay with ISI Impact factor and usage statistics
- Expand to other disciplines/publishers
 - Social Sciences
 - Humanities
 - Economy
 - Linguistics
- Operate at scale
 - Seeking collaboration
 - Enable other institutions to conduct similar experiments





Discussion

Questions

- Collecting imperative
 - New forms of scholarly communication
 - Developing data collections
- UC faculty is contributing to 1/12 of Elsevier publications, yet we license back 100%
- Open Access
 - Not even APCs added to the calculations
 - Revenue: \$182 million in 2012, growing 30% per year
- What about preservation?







How Much Does \$1.7 Billion Buy?

A Comparison of Scientific Journal Articles to Their Pre-print Versions

Sharon E. Farb
Peter Broadwell
Martin Klein
Todd R. Grappone

@farbthink

@peterbroadwell

@mart1nkle1n

@liber8er

If Harvard Can't Afford It, Who Can?

Harvard University says it can't afford journal publishers' prices

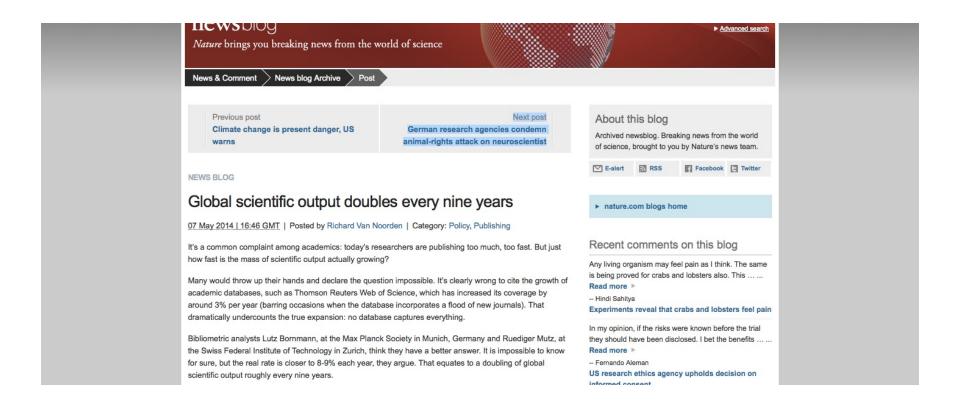
University wants scientists to make their research open access and resign from publications that keep articles behind paywalls



A graduation ceremony at Harvard University. Photograph: Brooks Kraft/Corbis



Increase in Global Scientific Output



Maybe It's Time for an Open Access Model

OPINION

If Harvard Can't Afford Academic Journal Subscriptions, Maybe It's Time for an Open Access Model

By Keith Wagstaff @kwagstaff | April 26, 2012 | Add a Comment









Read Later

Last week, Harvard's Faculty Advisory Council revealed that the school now spends \$3.75 million annually on academic journal subscriptions. Why so much? According to a memo the council sent out, some journals cost the school up to \$40,000 every year, with the two top publishers increasing the price of content 145% over the last six years.

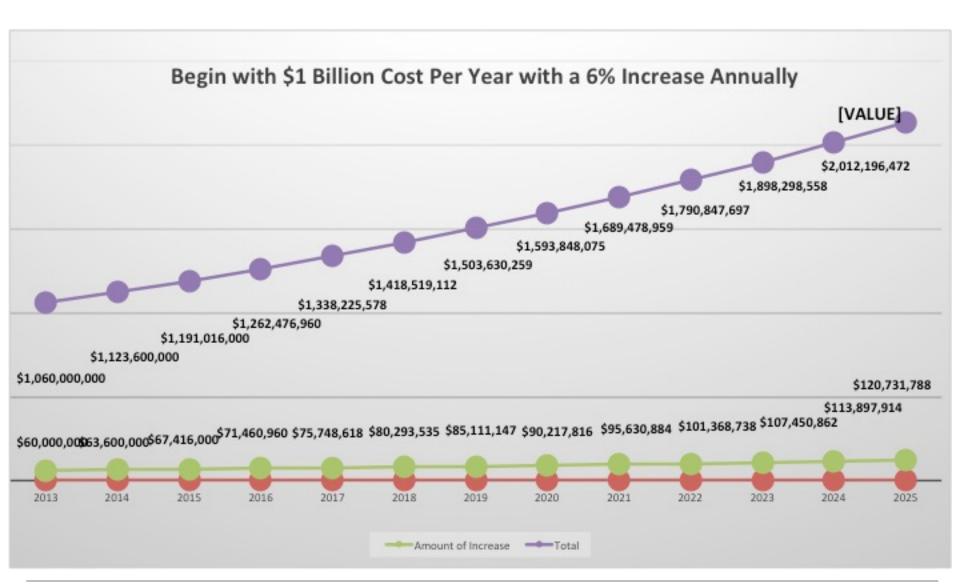
This is troubling for a number of reasons. First, in an age where the public can browse nearly 4 million articles for free on Wikipedia, a curious person looking to read up on the latest scientific research can expect to spend nearly \$30 to \$40 for a single paper from publishers such as Elsevier and Springer.



Steve Durwell / Getty Images







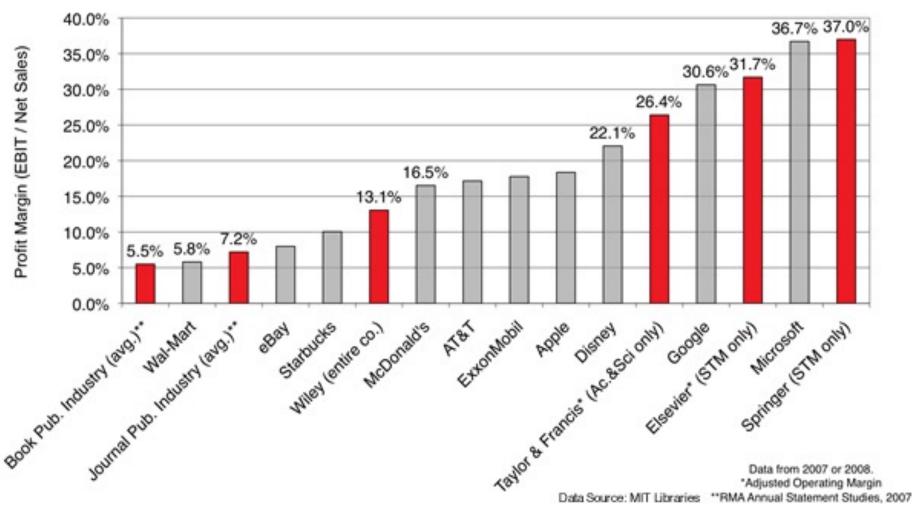


How Much Does \$1.7 Billion Buy?

@farbthink, @liber8er, @peterbroadwell, @mart1nkle1n #CNI15F, Washington, DC, 12/15/2015



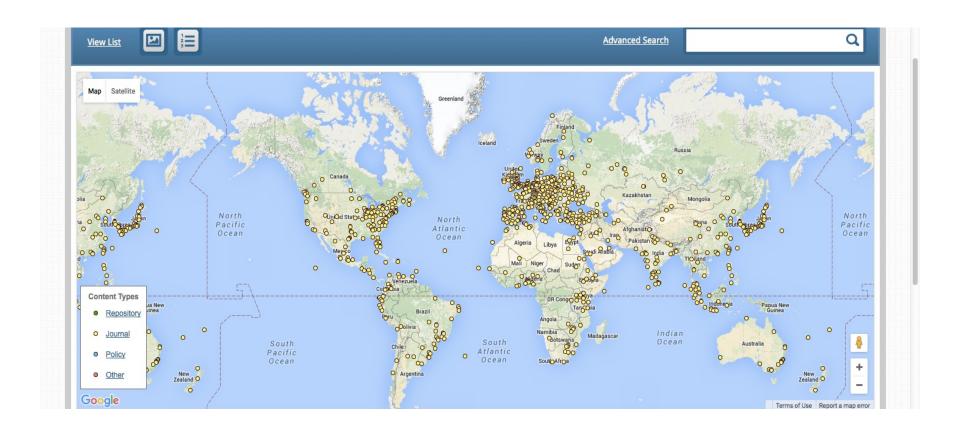
Profit Margin







http://www.openaccessmap.org/







The Rise of OA "Overlay" Journals?



Leading mathematician launches arXiv 'overlay' journal

Journal that reviews papers from preprint server aims to return publishing to the hands of academics.

Philip Ball

15 September 2015



New journals spring up with overwhelming, almost tiresome, frequency these days. But *Discrete Analysis* is different. This journal is online only — but it will contain no papers. Rather, it will provide links to mathematics papers hosted on the preprint server arXiv. Researchers will submit their papers directly from arXiv to the journal, which will evaluate them by conventional peer review.



