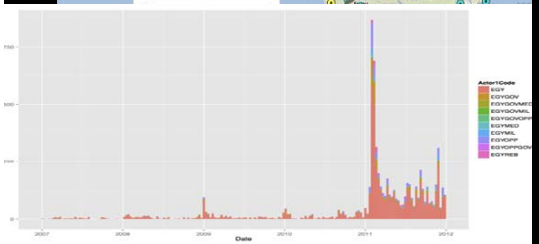
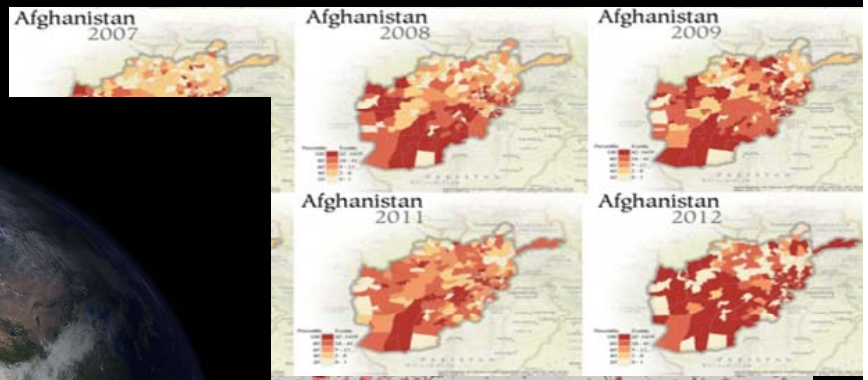
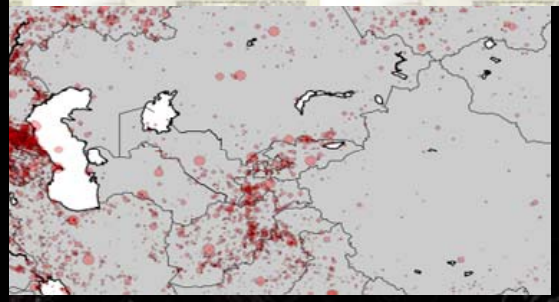
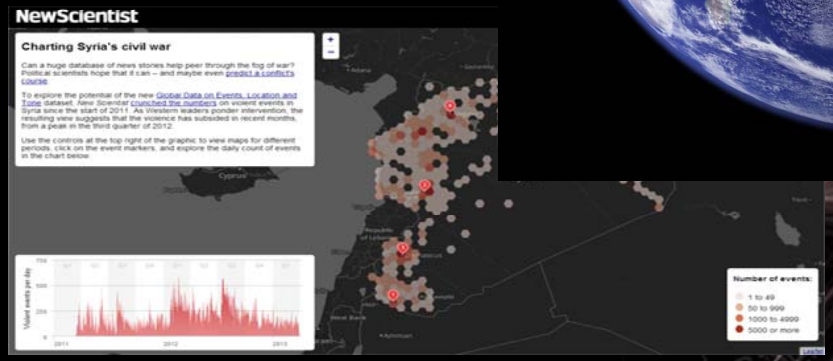
Pioneers in Mining Electronic News for Research

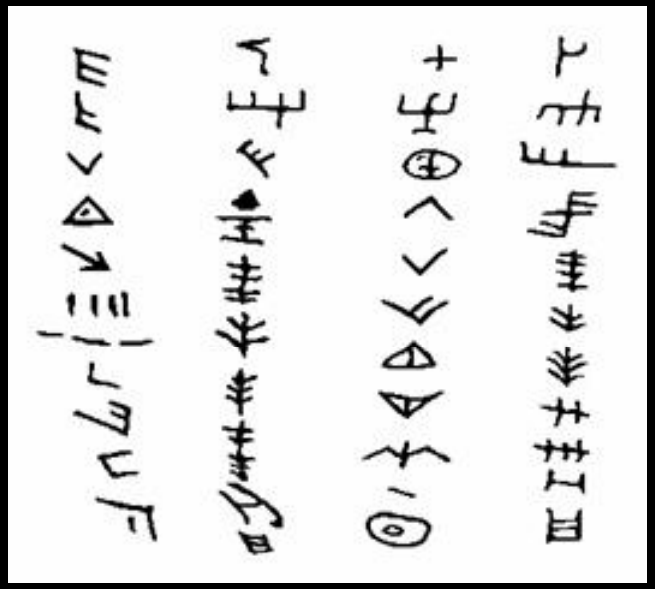
Kalev Leetaru
University of Illinois
http://www.kalevleetaru.com/

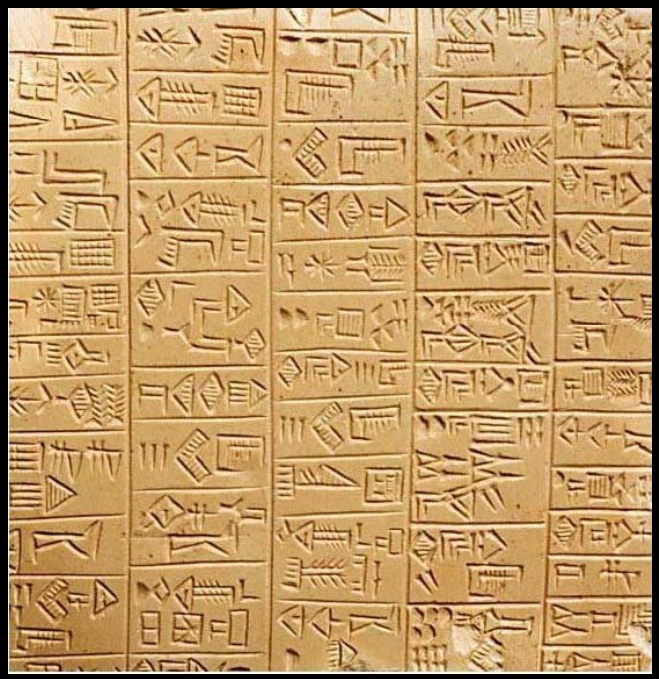


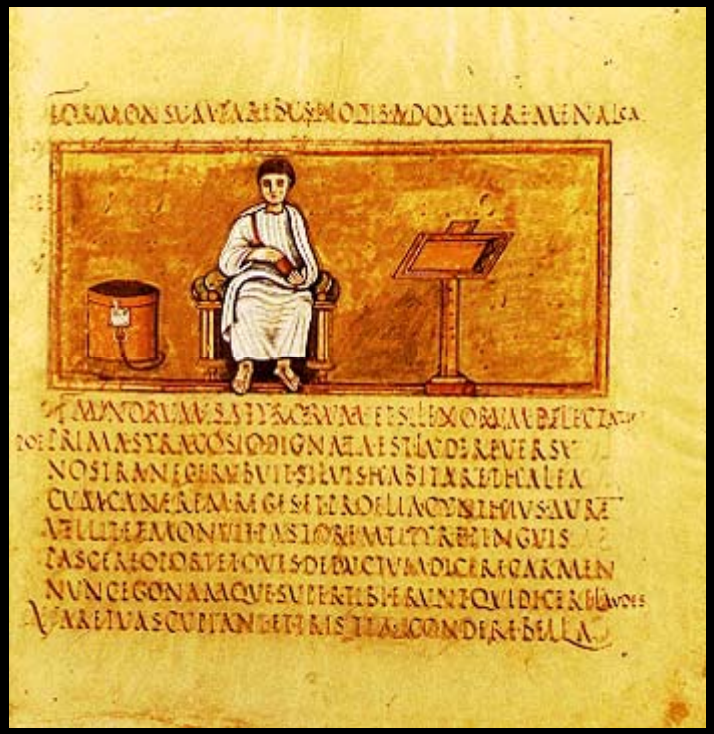


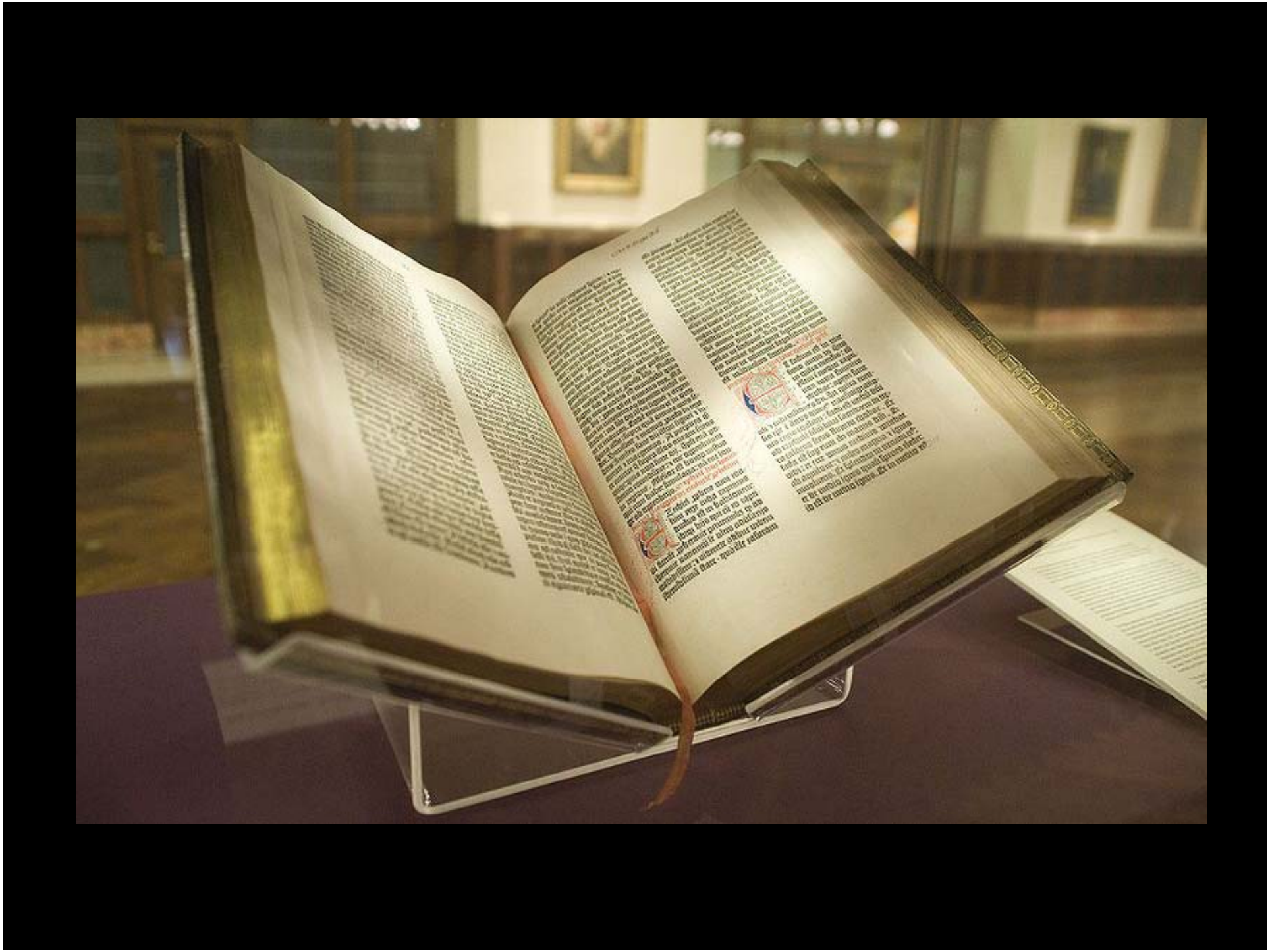


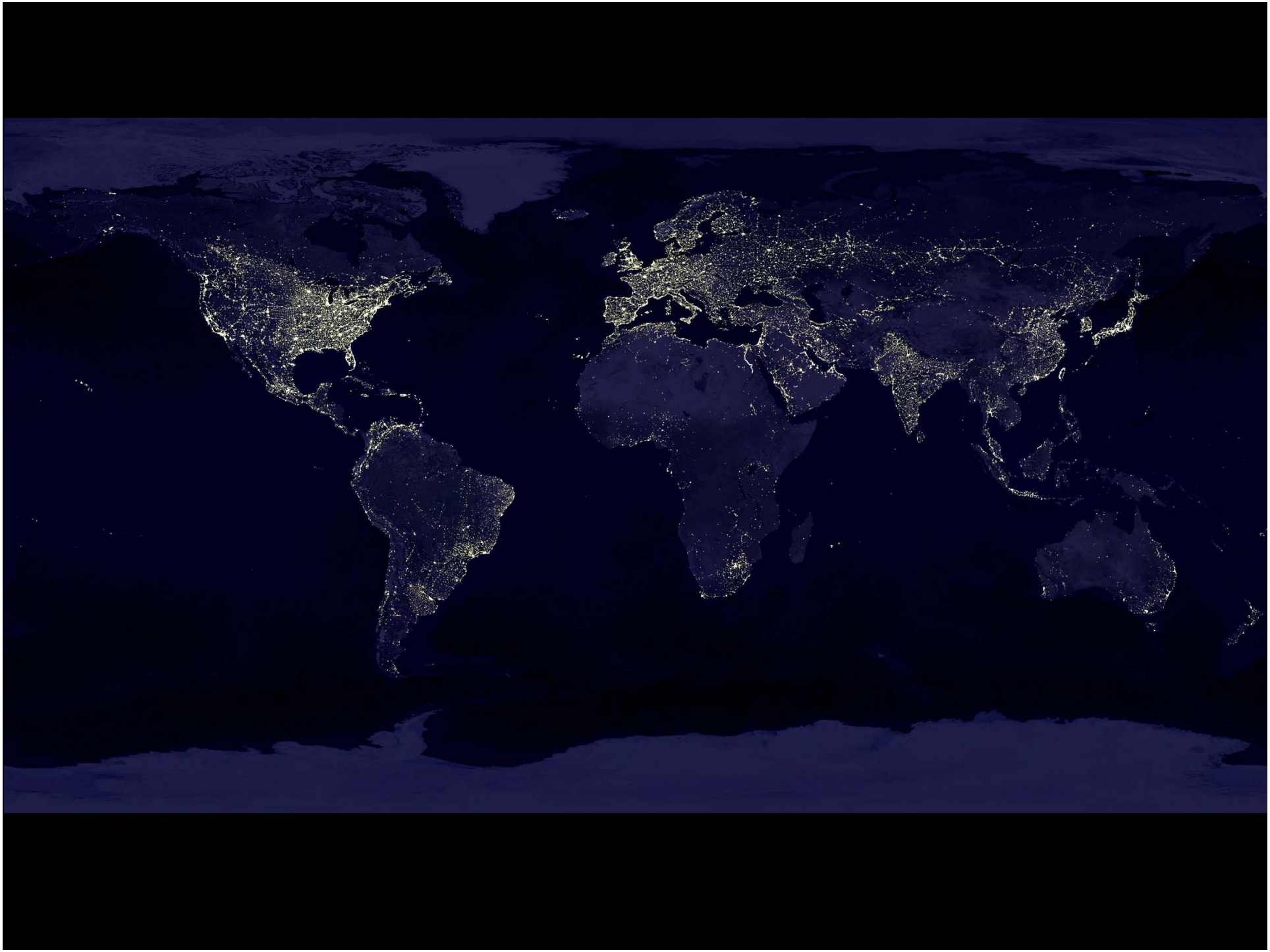












facebook
December 2010

Our Digital World

- 1/3 global population online
- As many cell phones as people on earth
- Facebook alone has:
  - 240 billion photographs (35% of all online photos)
  - 1 billion members with 1 trillion connections

Every Year

- 6.1 trillion text messages
- 2.2 trillion cell minutes in the US alone
- 107 trillion emails
- 1.6 million days worth of video uploaded to YouTube

Every Day

- 2.5 billion new items added to Facebook
- 300 million photos posted to Facebook
- 500TB of new data about society's innermost thoughts posted to Facebook
- As many words posted to Twitter every day as the entire New York Times in the last half-century
- 100 billion+ social media actions taken

Every Minute

- 600 new websites created
- 204 million emails sent
- 700,000 shares on Facebook
- 200,000 photos posted to Facebook
- 277,000 tweets sent

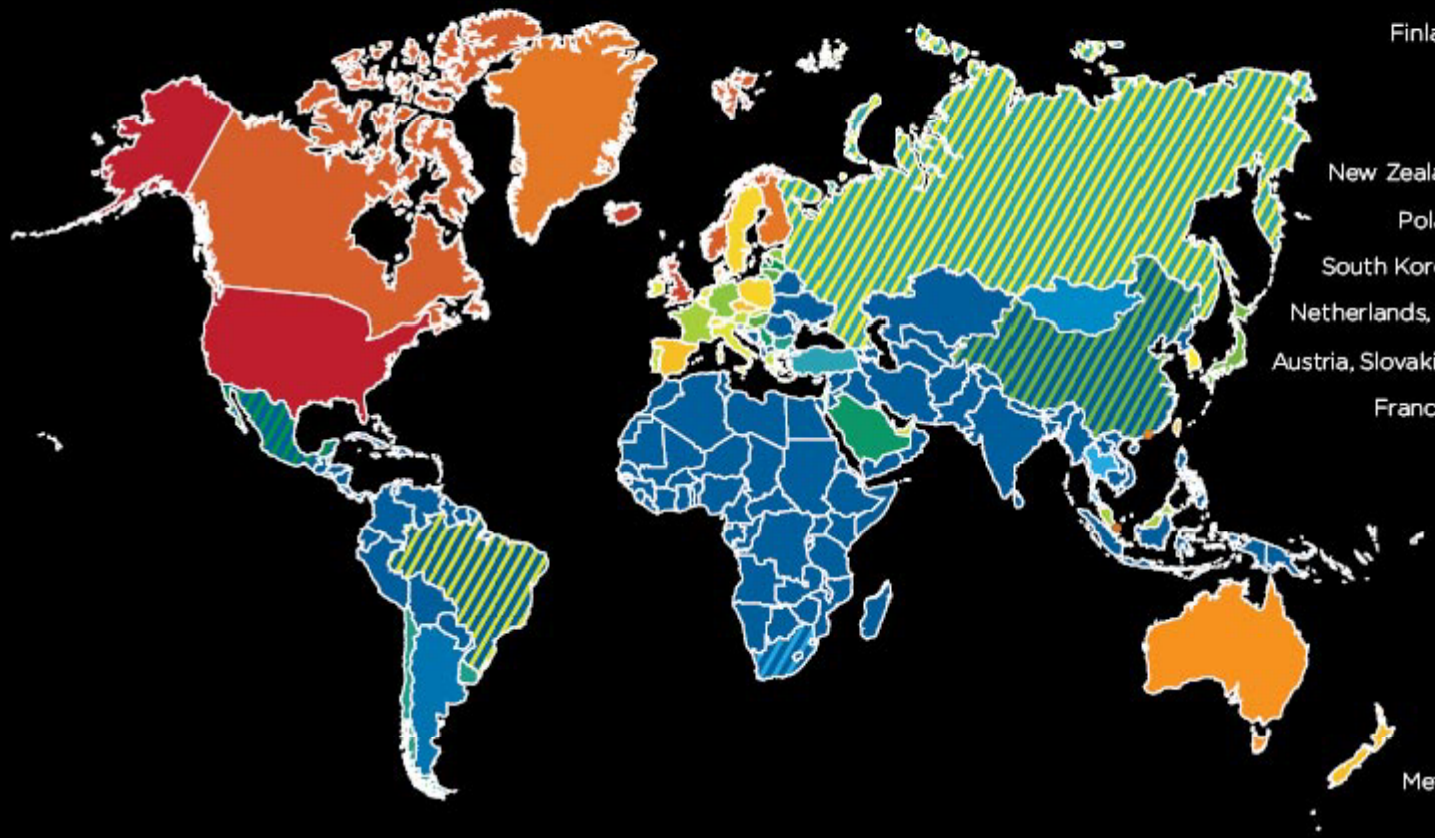
sgi
sgi
sgi
sgi
sgi
sgi
sgi
sgi

2005
Luca Bruno / AP

2013
NBC NEWS
Michael Sohn / AP

Newspaper extinction timeline

When newspapers in their current form will become insignificant*



USA **2017**
UK, Iceland **2019**
Canada, Norway **2020**
Finland, Singapore, Greenland **2021**
Australia, Hong Kong **2022**
Denmark **2023**
New Zealand, Spain, Czech, Taiwan **2024**
Poland, Sweden, Switzerland **2025**
South Korea, Metro Russia, Belgium **2026**
Netherlands, Ireland, Metro Brazil, Italy **2027**
Austria, Slovakia, Greece, Portugal, U.A.E. **2028**
France, Israel, Malaysia, Croatia **2029**
Germany, Estonia **2030**
Japan, Metro China **2031**
Hungary, Lithuania **2032**
Latvia, Metro Mexico **2033**
Serbia, Saudi Arabia **2034**
Bulgaria, Chile, Uruguay **2035**
Russia, Turkey **2036**
Metro South Africa, Thailand **2037**
Mongolia **2038**
Argentina **2039**
Rest of the world **2040+**

— stripes indicate that newspapers will be extinct in metropolitan areas before regional areas

FUTURE EXPLORATION NETWORK

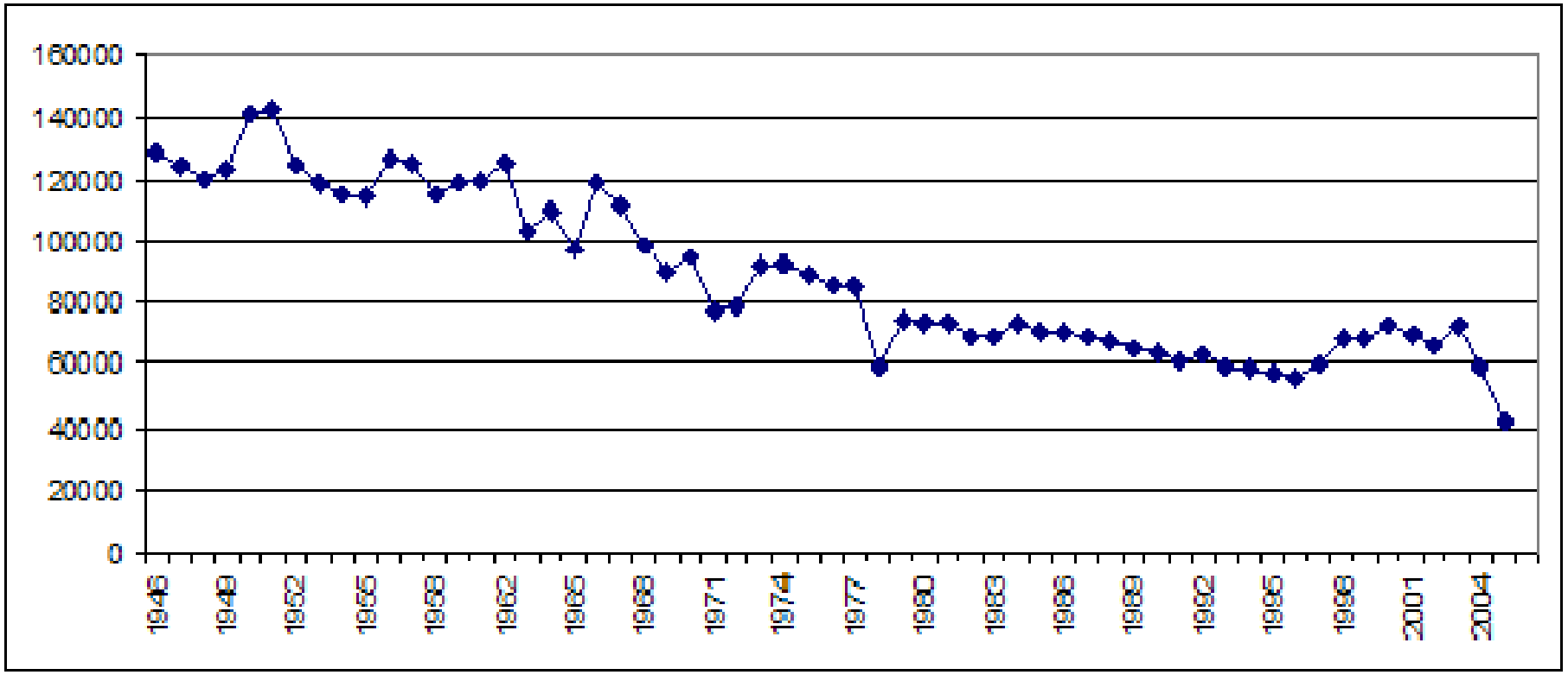
- **Media strategy**
- **Thought leadership content**
- **Scenario planning**

www.futureexploration.net

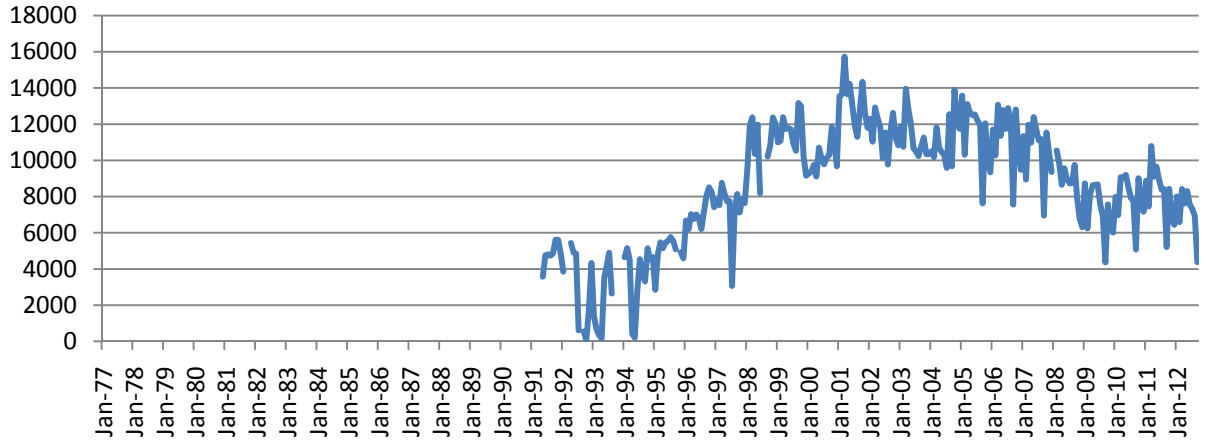
Created by Ross Dawson

Published under a Creative Commons Attribution-ShareAlike 2.5 licence

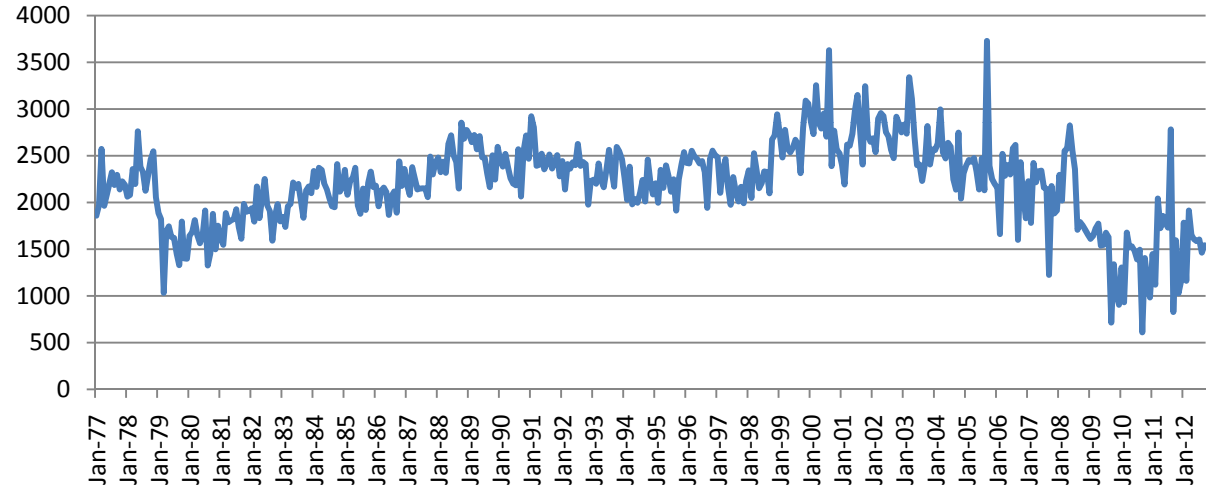
The Shrinking Newshole



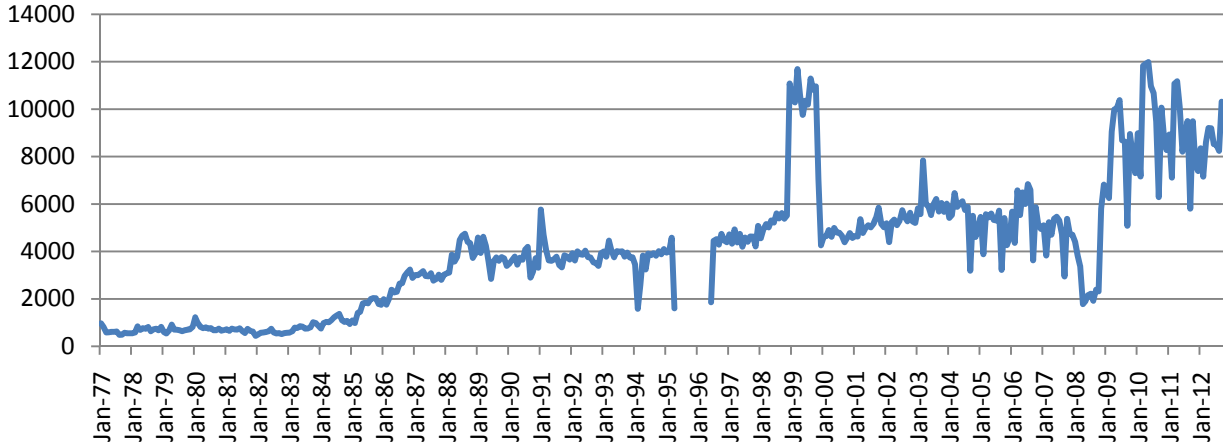
New York Times number articles per year (Proquest)



Articles per month published by Agence France Presse (International coverage)

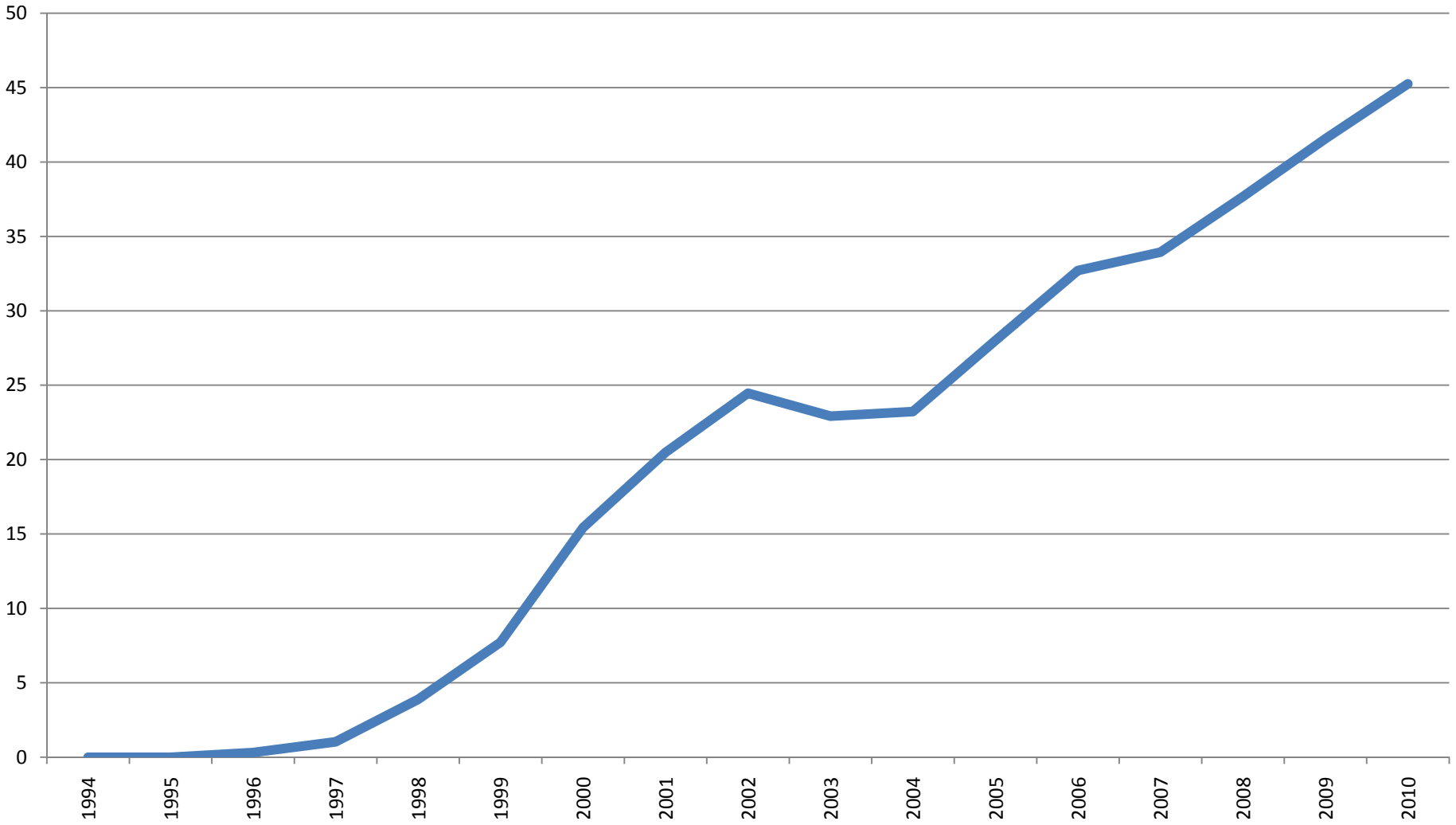


Articles per month published by Associated Press (International coverage)

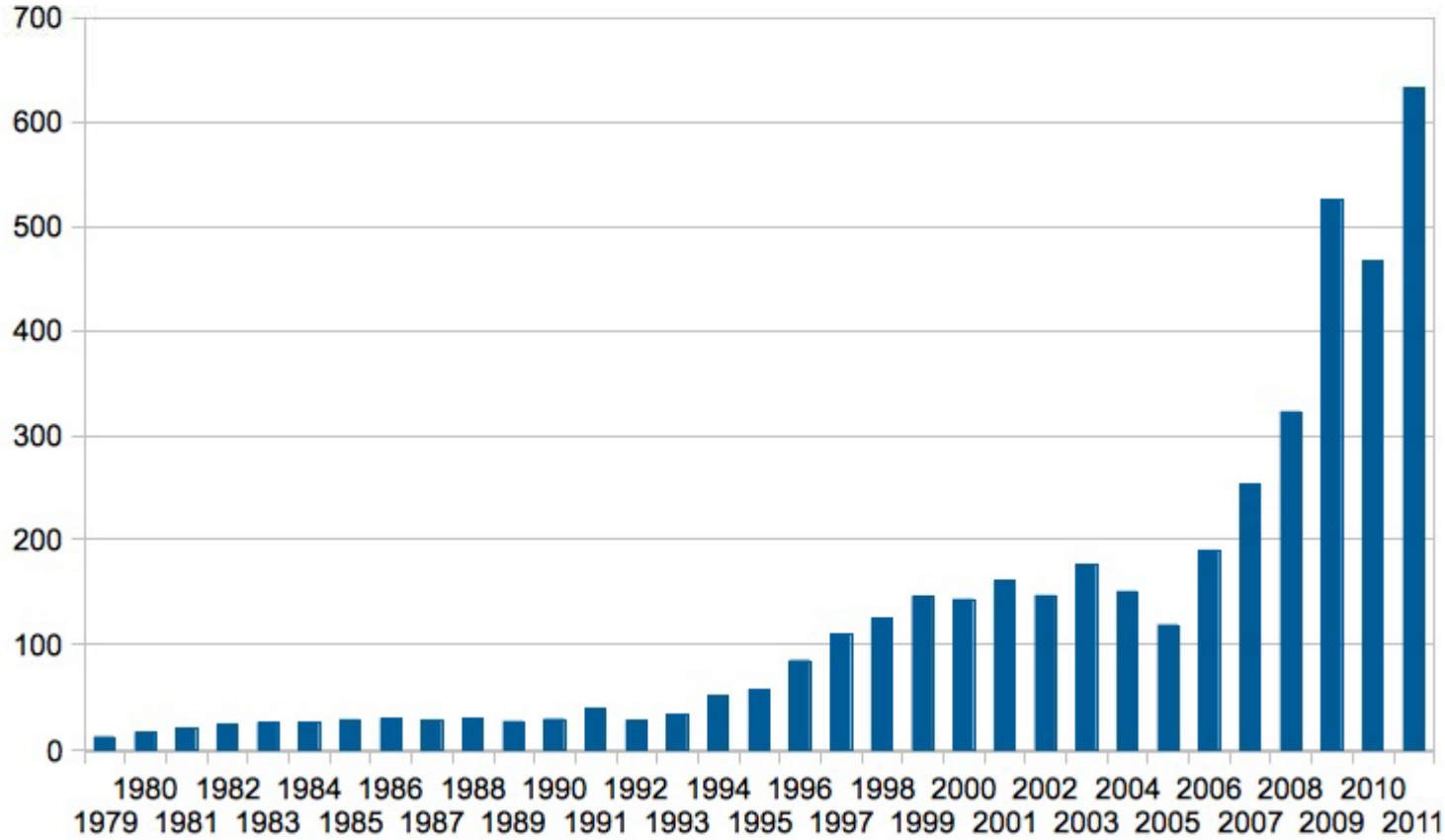


Articles per month published by Xinhua (International coverage)

The Rise of Web News



The Impact of Web News



72261 Results * Search within

Create alert | Create RSS feed | Save search

0 Selected items [Clear] | Save to My Research | Email | Print | Cite | Export/Save

Select 1-20 Brief view | Detailed view

1 Research Libraries To Close Saturdays As a Money-Saver — Preview
New York Times (1923-Current file) [New York, N.Y] 09 May 1975: 37.
...Library have decided to close the system's research libraries on Saturdays,
Citation/Abstract Full text - PDF (30 KB)

2 Research Libraries Map ER Large-Scale Cooperation: Area of Cooperation — Preview
By MICHAEL KNIGHT. **New York Times (1923-Current file)** [New York, N.Y] 15 May 1973: 42.
...Public Library and the Harvard, Yale ard Columbia University libraries, feeling
Citation/Abstract Full text - PDF

3 Four Major Libraries Combine Research Operations: Criticism in Book World — Preview
By ERIC PACE. **New York Times (1923-Current file)** [New York, N.Y] 24 Mar 1974: 59.
...Public Library has joined with the Columbia, Yale and Harvard libraries in a
Cited by (1)
Citation/Abstract Full text - PDF (181 KB)

4 Offers New Plan For All Libraries To Aid Research: Cooperation on National Basis Is Urged by Princeton Report Urges End of "Fallacy" Reviews Comparative Growths — Preview
Special to THE NEW YORK TIMES.. **New York Times (1923-Current file)** [New York, N.Y] 10 Nov 1940: 66.
...university and research libraries to strengthen the bibliographical resources
Citation/Abstract Full text - PDF

5 LIBRARIES EXPAND RESEARCH IMPORTS: Cooperative Plan to Acquire New Books Is Extended to 6 More Countries — Preview
New York Times (1923-Current file) [New York, N.Y] 21 Dec 1948: 23.
...to arrive from abroad within the coming year for American libraries under the
Citation/Abstract Full text - PDF (56 KB)

6 Research Libraries To Close Saturdays As a Money-Saver — Preview
New York Times (1923-Current file) [New York, N.Y] 09 May 1975: 75.
...Library have decided to close the system's research libraries on Saturdays,
Citation/Abstract Full text - PDF (32 KB)

7 Research Libraries Get $1.6 Million Grant — Preview
New York Times (1923-Current file) [New York, N.Y] 07 Oct 1980: C6.
... The four research libraries of the New York Public Library will be the
...million raised by private donations, said a library spokesman.
Citation/Abstract Full text - PDF (17 KB)

Sort results by:

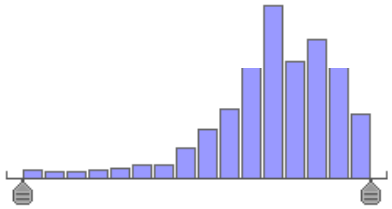
Relevance

Sort

Narrow results by

- Publication title
- Document type
- Subject
- Company/organization
- Location
- Person
- Publication date

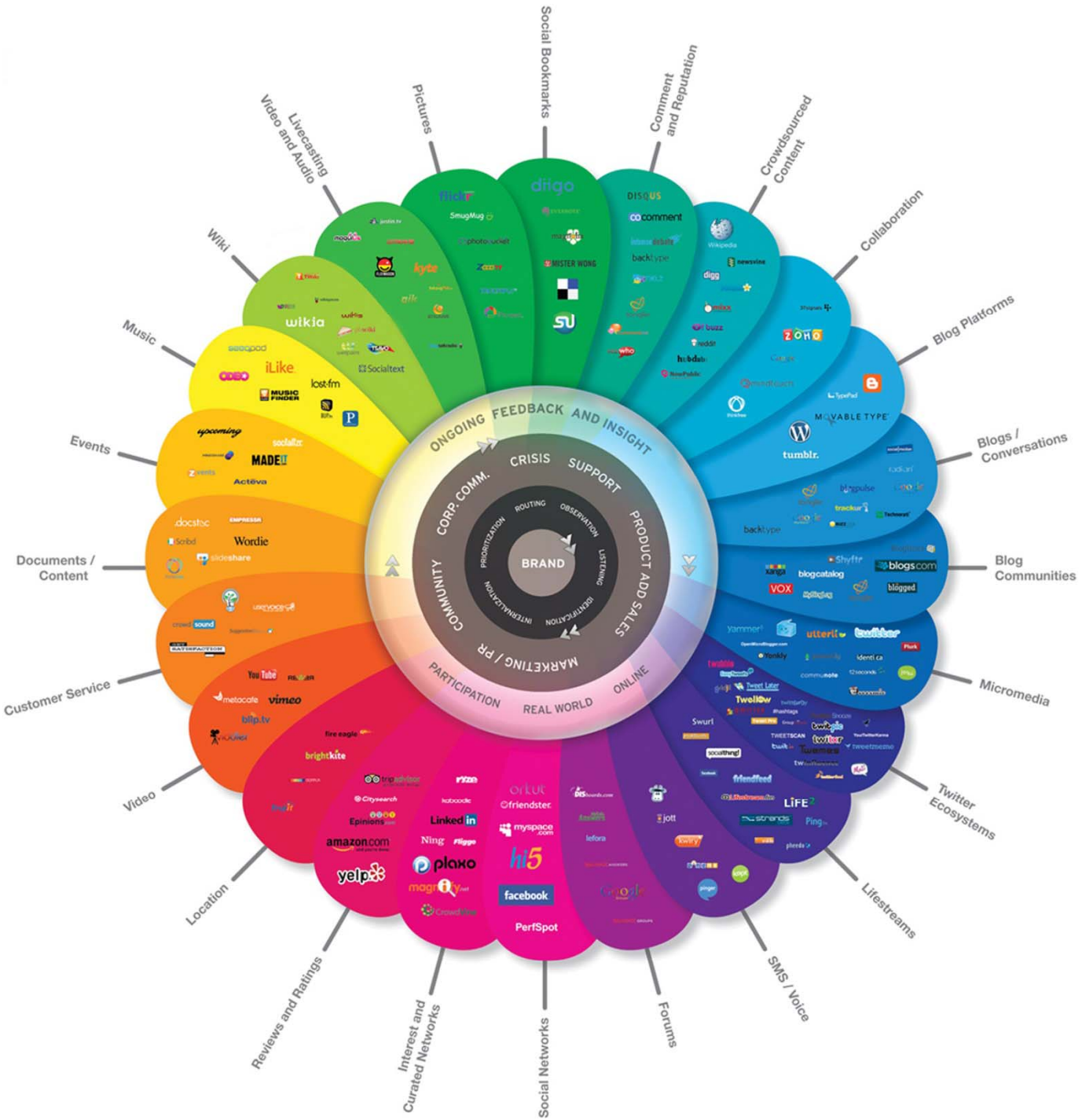
1851 - 2009 (decades)

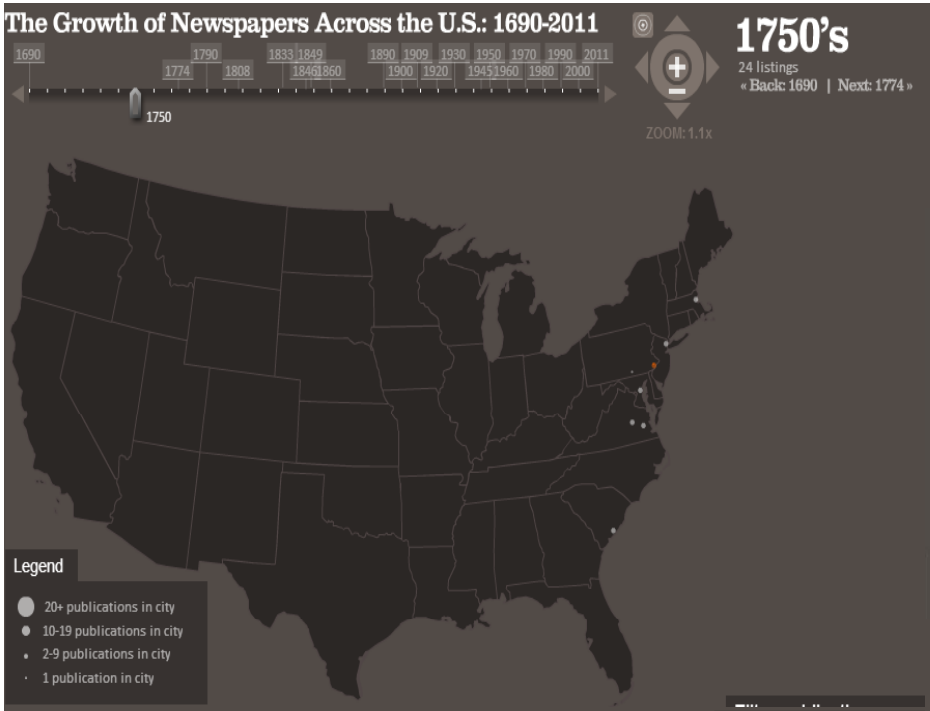


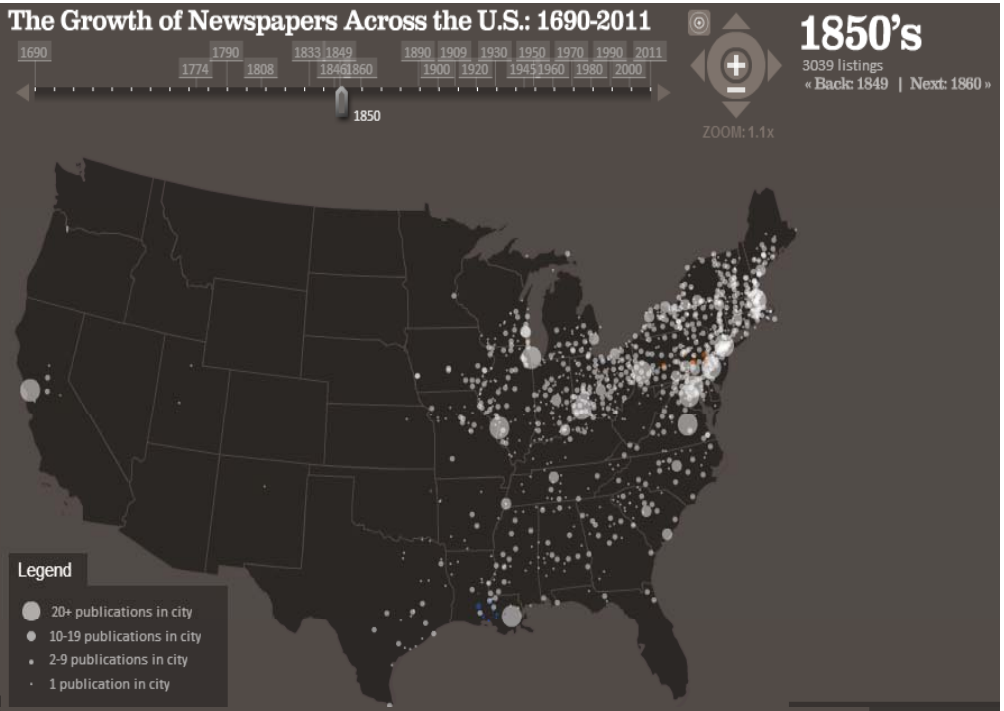
Update

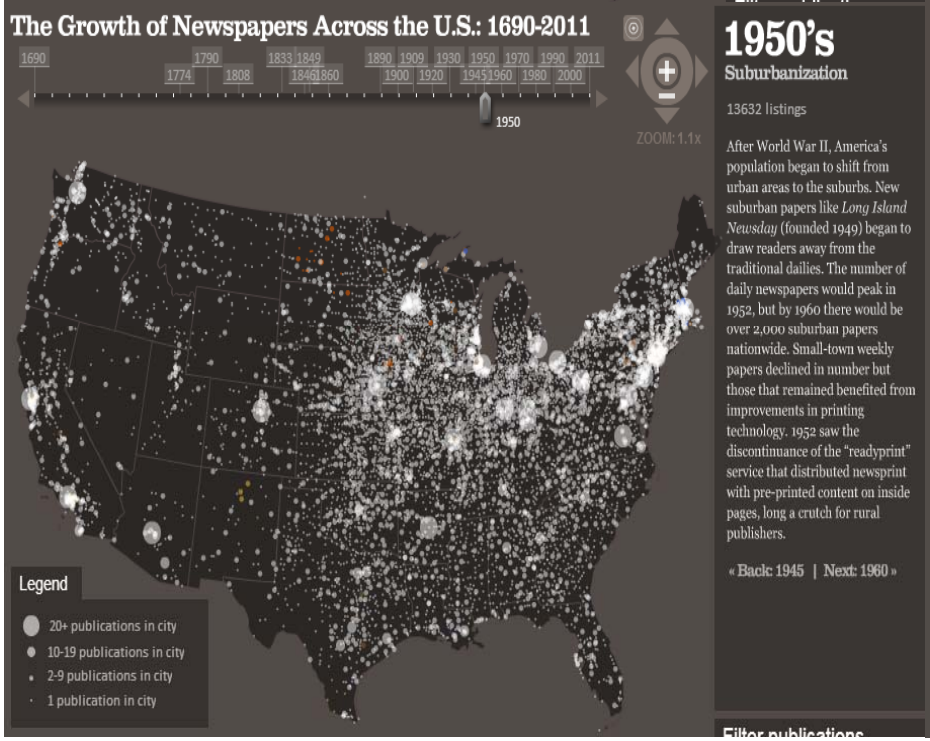
Enter a specific date range

How do we use the news?


Social Bookmarks
Comment and Reputation
Crowdsourced Content
Collaboration
Blog Platforms
Blogs / Conversations
Blog Communities
Micromedia
Twitter Ecosystems
Lifestreams
SMS / Voice
Forums
Social Networks
Interest and Curated Networks
Reviews and Ratings
Location
Video
Customer Service
Documents / Content
Events
Music
Wiki
Livecasting Video and Audio
Pictures
ONGOING FEEDBACK AND INSIGHT
PRODUCT ADD SALES
MARKETING / PR
COMMUNITY
CORP. COMM.
CRISIS
SUPPORT
ROUTING
OBSERVATION
LISTENING
IDENTIFICATION
INTERNALIZATION
PRIORITIZATION
BRAND
PARTICIPATION
REAL WORLD
ONLINE


The Growth of Newspapers Across the U.S.: 1690-2011
1690 1774 1790 1808 1833 1849 1846 1860 1890 1900 1909 1920 1930 1945 1950 1960 1970 1980 1990 2000 2011
1750
ZOOM: 1.1x
1750's
24 listings
« Back: 1690 | Next: 1774 »
Legend
20+ publications in city
10-19 publications in city
2-9 publications in city
1 publication in city


The Growth of Newspapers Across the U.S.: 1690-2011
1690 1774 1790 1808 1833 1849 1846 1860 1890 1900 1909 1920 1930 1945 1950 1960 1970 1980 1990 2000 2011
1850
ZOOM: 1.1x
1850's
3039 listings
« Back: 1849 | Next: 1860 »
Legend
20+ publications in city
10-19 publications in city
2-9 publications in city
1 publication in city


The Growth of Newspapers Across the U.S.: 1690-2011
1690 1774 1790 1808 1833 1849 1846 1860 1890 1900 1909 1920 1930 1945 1950 1960 1970 1980 1990 2000 2011
1950
ZOOM: 1.1x
1950's
Suburbanization
13632 listings
After World War II, America's population began to shift from urban areas to the suburbs. New suburban papers like Long Island Newsday (founded 1949) began to draw readers away from the traditional dailies. The number of daily newspapers would peak in 1952, but by 1960 there would be over 2,000 suburban papers nationwide. Small-town weekly papers declined in number but those that remained benefited from improvements in printing technology. 1952 saw the discontinuance of the "readyprint" service that distributed newsprint with pre-printed content on inside pages, long a crutch for rural publishers.
« Back: 1945 | Next: 1960 »
Legend
20+ publications in city
10-19 publications in city
2-9 publications in city
1 publication in city

The Growth of Newspapers Across the U.S.: 1690-2011
1690 1774 1790 1808 1833 1849 1846 1860 1890 1900 1909 1920 1930 1945 1950 1960 1970 1980 1990 2000 2011
2011
ZOOM: 1.1x
2011
13670 listings
« Back: 2000
Legend
20+ publications in city
10-19 publications in city
2-9 publications in city
1 publication in city

“Japanese radio intensifies still further its defiant hostile tone; in contrast to its behavior during earlier periods of Pacific tension, Radio Tokyo makes no peace appeals. Comment on the United States is bitter and increased.”

December 6, 1941

Communications Analysis

- “Mass Communications”
- Late 1800’s rise of formalized study of the press – how was press changing? (topics, sensationalism, distortion, etc)
- DeWeese 1977 – estimate $3M per 1 billion words digitized (vs ~3.6 billion words per day on Twitter)

Communications Analysis

Five Stages of Textual News Analysis (Van Cuilenburg, 1991):

- Frequency Analysis (-1950's): counts of words and themes (back again via ngrams)
- Valence Analysis (1950's-): positive/negative words
- Intensity Analysis (1950's-): how positive/negative each word is
- Contingency Analysis (1960's-): moving from counts to associations and patterns (descriptive to predictive)
- Computational Analysis (1960's-): General Inquirer, etc (rise of digital news content)

Communications Analysis

- Visual Presentation: subject/presentation of figures/photos used, layout of text, etc (still mostly human – requires page image ala PressDisplay)
- Layout: above/under fold, page number, section organization and structure (human or machine – page image or XML structural info)
- Content: just need text-only (LexisNexis and other text-only archives)

Communications Analysis

- Human analysis often needs page-view access – in web era, often studying context, navbars, clickstream – need rich preservation of original HTML and visual layout
- Increasing shift towards large-scale computational analysis – almost exclusively textual – “chrome extraction” (extracting news article body text from the navbars, ads, template, etc).

Political Communication

- Political discourse – discussion of candidates and political themes. Very similar to overall Mass Communication usage. Often more of a focus on the content, but imagery, especially campaign photos, and positioning with respect to other articles often key.

History

- Mostly focused on digitized historical content (ie, Proquest Historical Newspapers, etc).
- Big focus on presentation – for example, portrayals of minorities and women in advertisements. Visual often critical (historical images as context).
- Early digitization efforts often discarded advertisements.
- Emphasis on visual presentation, but increasingly digital humanities relying on the text.

Political Science / Sociology

- Early quantitative “event databases” in the late 1970’s – large teams of humans reading news articles and compiling lists of “events”
- Codify a textual description of a riot into a spreadsheet entry recording where and when it happened and who was involved
- Increasingly automated – relies just on textual article content
- Wall Street pushing these techniques – new dedicated newswires designed for machine-only consumption

Computer Science

- Little interest in the content of the news, just using news as a source of textual input for algorithm development
- One of the biggest consumers of digital and digitized news content
- Uses whatever is easiest to get: moving more towards Wikipedia and Twitter now, but still big focus on news because of legacy collections and gold standards

Computer Science

- Almost exclusively textual news
- Small set of gold standard collections – most work focuses on new algorithms and improvements to processing those collections for faster or more accurate (better part-of-speech tagging; better topic extraction, etc)
- Little interest in results as they apply to understanding the news – instead, focus is on comparing to past work to demonstrate faster/better/etc.
- Means must be able to run on the EXACT same text as other projects – share massive volumes of copyrighted content.

Computer Science Collections

- Gigaword: 26GB / 4M articles: AFP/NYT/Xinhua/LATimes/WashPost/Bloomberg/etc 1990's-2010 (traditional wire/print news media)
- ICWSM: 3TB / 300M+ items / 14M news articles: "includes the syndicated text, its original HTML as found on the web, annotations and metadata (e.g., author information, time of publication and source URL), and boilerplate/chrome extracted content" (web news)
- LDC Archive >150 archives (http://www.ldc.upenn.edu/Catalog/byType.jsp)

Computer Science Collections

- Focused on diverse collections of content for replication, not archival. Doesn't always contain 100% of content.
- Web content often provided from commercial web aggregators similar to Google News: crawling open web and archiving
- Worked through all of the legal issues to allow long-term archival (decades already for some collections) and unlimited redistribution of TB's of news content for non-commercial academic research.

Full Content Access

- A common theme among all of these disciplines is the need to access the full contents of the articles, not just “ngrams” or proxy derivatives – even in the digital humanities, most focus is on relationships and context
- Provides unique challenges with respect to copyright and access

Dual Archiving: Presentation + Content

- Humanities and some social sciences often want access to the visual presentation of the news
- Other social sciences and the computer sciences largely treat presentation as noise and just want the text
- Means we need to have two copies of each page: a visual snapshot of its presentation ala PressDisplay and a text-only version with all ads and navbars, etc, extracted to contain just the core text

What is “All of the News”?

- LexisNexis has just a fraction of news sources and not all content from those sources (blackouts, licensing restrictions, etc)
- By virtue of being so ubiquitous on campuses, it has become “all the news” for many fields – ok to say “I used all news on topic X from LexisNexis Newswires file”.
- Google News via RSS increasingly becoming this way – gives a common definition, but no replication – content isn’t archived long-term

Archiving Web News

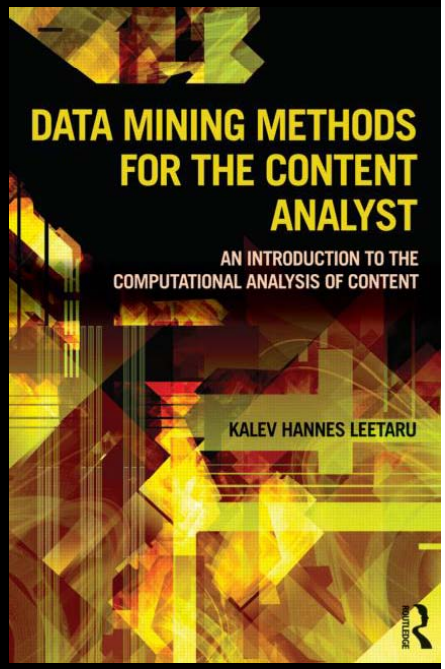
- News websites make heavy use of dynamic customization today. Used to be “morning” and “evening” editions of a paper – today every single visitor has their own custom-tailored copy, at least in the advertisements, but increasingly across navbars and content ranking, and that changes moment-by-moment
- Discussion sections at the bottom of articles
- Even major papers like New York Times are updating and EDITING articles days or weeks later
- Dynamic tracking URLs means solid URLs are fading
- Articles expire after 24 hours, 1 week, 1 month, etc
- Wire stories may be merged together and presented in a single page

Why Preserve News?

- Why preserve news in the first place?
- Humanities and social sciences often need long time horizons – need historical backfiles ("time machine" to turn back the clock) – historically just needed human access to retrieve small portions of it – completeness critical, and need visual presentation
- Computer science needs replication – be able to repeat a study using the exact same source material – need to be able to bulk-download TB's of data and keep for long periods of times for massive projects - just give them a ZIP/XML bulk download feature – they don't care about completeness, just access, and text is preferred

What's "Good Enough?"

- The Computer science communities have well-established history of using partial collections – a "Google News-like" system that permanently archived all web news it found would work fantastically
- Humanities and social sciences need to understand "completeness" – "what % of all of CNN.com is in here?" – however, current standards like LexisNexis aren't complete either, so a widely-accessible archive might become that new standard even if it isn't complete

Pioneers in Mining Electronic News for Research

Kalev Leetaru
University of Illinois
http://www.kalevleetaru.com/