PageOneX lines of work

Development

Scraping

The user would be able to select from different data sources when starting a thread.

Print front pages
We are currently scraping from
- [kiosko.net](#)

Future steps would include other data sources:
- newseum (http://www.newseum.org/todaysfrontpages/) they have an archive of images since approx. 2002 (stored in CD's!). Possible collaboration with Paul Sparow.
- newspapers print front pages archives:
    - set up scrapers for the different newspapers front pages in pdf([gdoc that list different newspaper front page data to help bulid a script](#))
    - rescue the scrapers already built (earlier commits)
    - pdf to img
- local files: images uploaded by users

An interesting related project is Xed http://diuf.unifr.ch/main/diva/research/research-projects/xed "a new tool for extracting hidden structures from electronic documents. *Document Image Analysis for Libraries, 2004. Proceedings. First International Workshop on*. IEEE."

Online home page
Make preliminary analysis: frequency of scraper, images, html to enable queries
- Example and database: http://pastpages.com/ Only iages

Thread creation / edit

Better newspaper selection / order
Enable thread and topic forking (wikiscraper model)..

Coding

- Fix slow behavior when many images or areas in the thread with:

- ○ ajax
    - ○ resize of images to show smaller images in the display view
- Allow multicoder (sharing thread ownership),
- Allow simultaneous coding
- Add annotation features (for days, for images), annotation standards?
- Capability for magnet areas to avoid overlapping of areas
- Zooming in images to read small texts
- Facilitate the coding process enabling search of certain words that would pre-seslect front pages (OCR, or read text from PDF should be activated, external data bases: API New York Times, Lexis Nexis...)

- Allow connection of areas to form one "area-article", this will lead to be able have the "count" of articles per day/newspapers in one topic http://numeroteca.org/wp-content/uploads/2012/10/syria-20112-07.png
- Allow multiple taxonomies (Ex: one taxonomy for theme + taxonomy for frame analysis http://numeroteca.org/2013/02/06/3-steps-to-measure-the-corruption-coverage-in-spain/)

Export
- Already works ods + json
- Image .png
- Enable embed option: image? svg?
- Thread: all related data

Display
Test D3 to allow richer visualizations and direct svg integration (Ed is working in this):
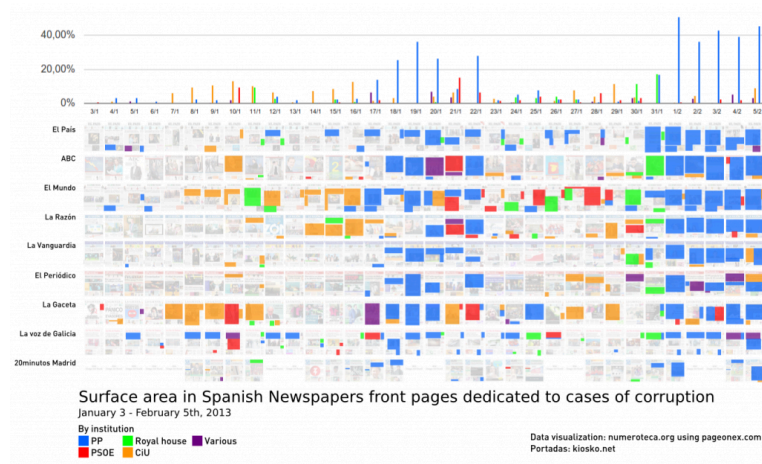- bar char split by newspapers
- live selection of newspapers to modify on the fly which newspapers show in the bar chart
- array of newspapers in svg

Enable the selection of only certain days for the display view (different form the original time frame of the thread). Very long threads have difficulties for being displayed. In the past, a year long study implied the creation of 12 threads. This new feature would enable the creation of 1 thread and then visualize only certain parts of it.
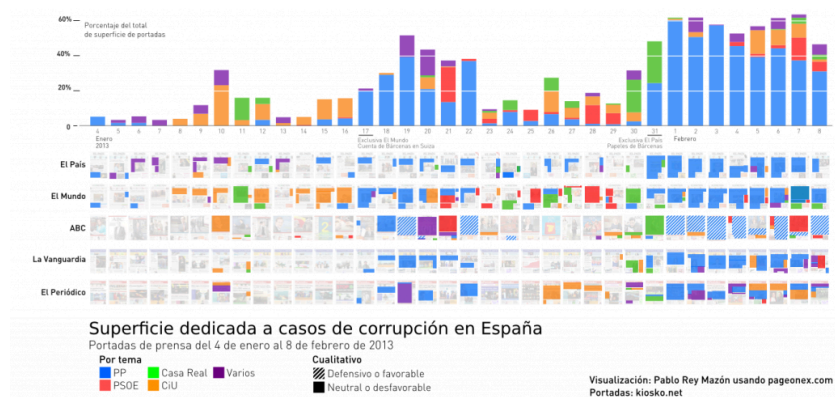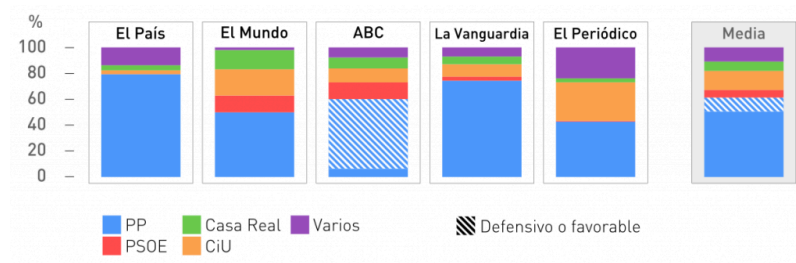
Exploring new ways to show the data:
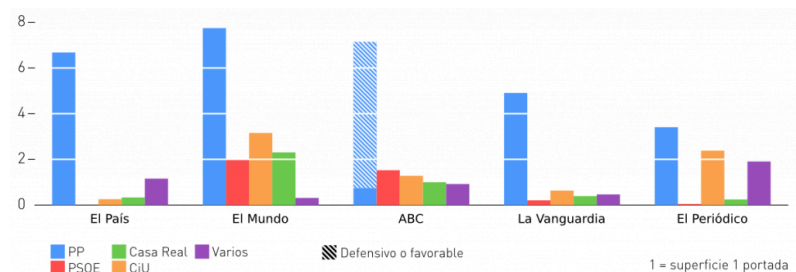
Bar chart:
Multiple column

Surface area in Spanish Newspapers front pages dedicated to cases of corruption
January 3 - February 5th, 2013
By institution
PP   Royal house   Various
PSOE   CiU
Data visualization: numeroteca.org using pageonex.com
Portadas: kiosko.net

1 column

Superficie dedicada a casos de corrupción en España
Portadas de prensa del 4 de enero al 8 de febrero de 2013
Por tema
PP   Casa Real   Varios
PSOE   CiU
Cualitativo
Defensivo o favorable
Neutral o desfavorable
Visualización: Pablo Rey Mazón usando pageonex.com
Portadas: kiosko.net

Comparison among newspapers: % to each topic

PP   Casa Real   Varios   Defensivo o favorable
PSOE   CiU

Total space (in number of front pages) for each topic

PP   Casa Real   Varios   Defensivo o favorable
PSOE   CiU   1 = superficie 1 portada

Platform

- Users pages
- Mix with other media data: TV from archive.org, Twitter data...

- Voting/curation of threads to give visibility to other threads (not just the last created)
- Show preview of the thread in thread list

Outreach / Community building

Tutorials
Screencast
Case studies

User driven design

Contact existing organizations that make similar work
- Women in Journalism. Report: http://womeninjournalism.co.uk/wp-content/uploads/2012/10/Seen_but_not_heard.pdf
- FAIR: Fairness & Accuracy In Reporting http://fair.org/