PrEV

Information Retriever @ Tsinghua University

NUS - Extreme - Tsinghua

Preservation Explorer and Vault for Web 2.0 User-Generated Content

http://han.ddns.comp.nus.edu.sg/prev/

Anqi Cui[1], Liner Yang[1], Dejun Hou[2],
Min-Yen Kan[2], Yiqun Liu[1], Min Zhang[1], Shaoping Ma[1]
[1]Tsinghua University, National University of Singapore[2]

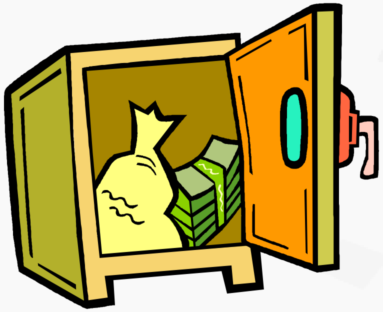
NUS National University of Singapore

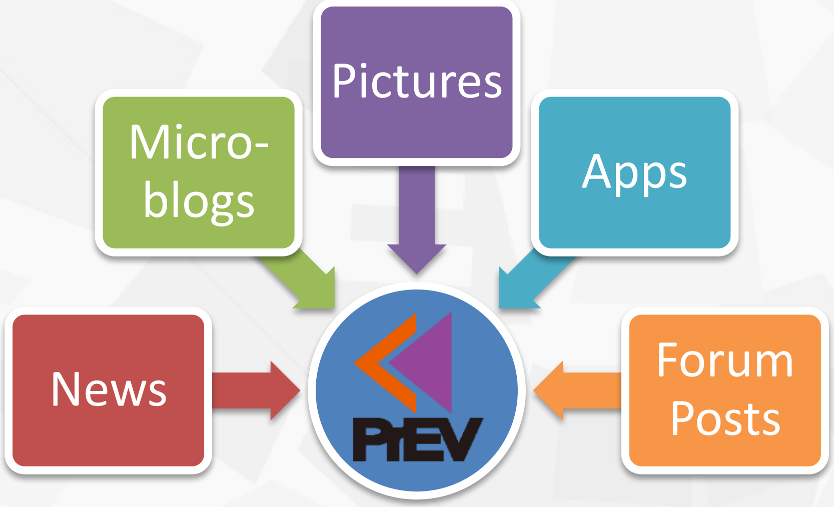
清華大學 Tsinghua University

mda Media Development Authority Singapore

PrEV and NExT

- *Preserve* the past and today's Web 2.0 User-Generated Content (UGC) as a *Vault*, to help future researchers –
- – *Explore* what was happening in our lives





Pictures

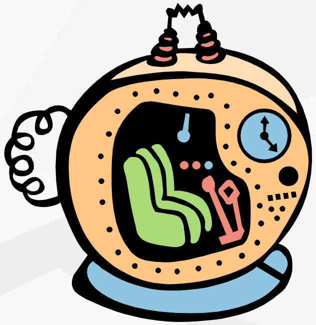
Micro-blogs

Apps

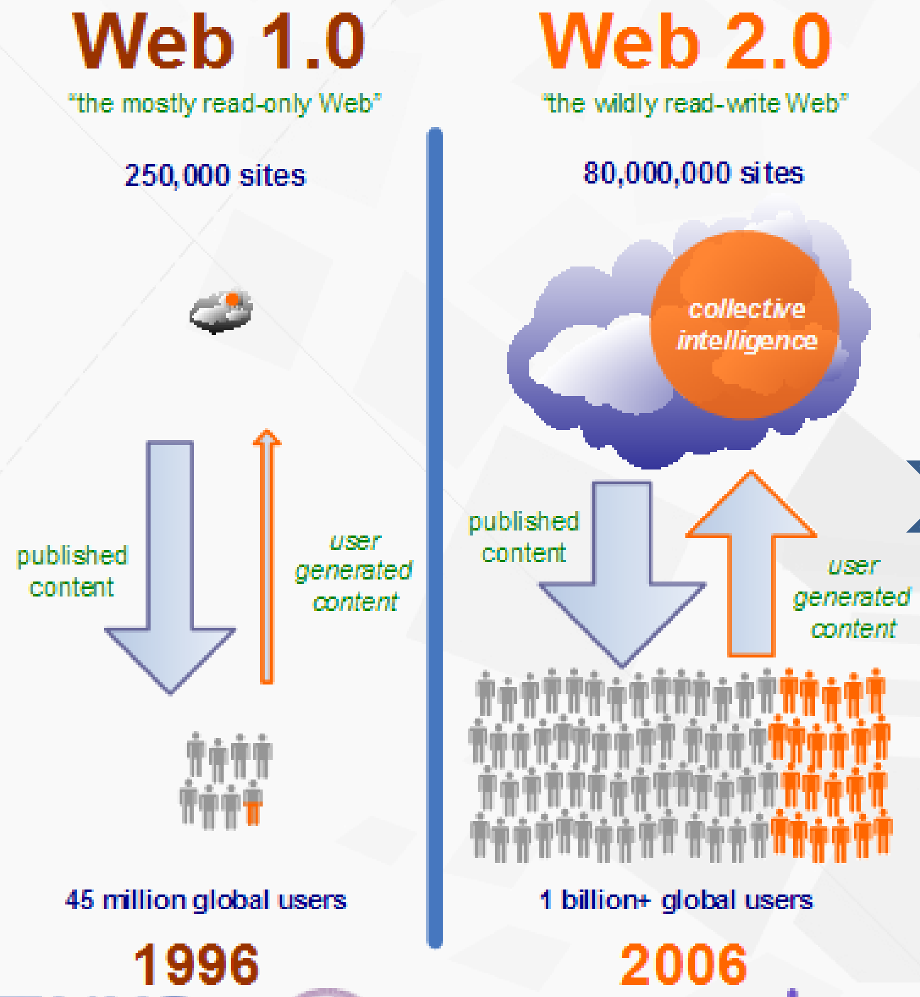
News

Forum Posts

PrEV



Motivation: The Web



Web 1.0

"the mostly read-only Web"

250,000 sites

45 million global users

1996

Web 2.0

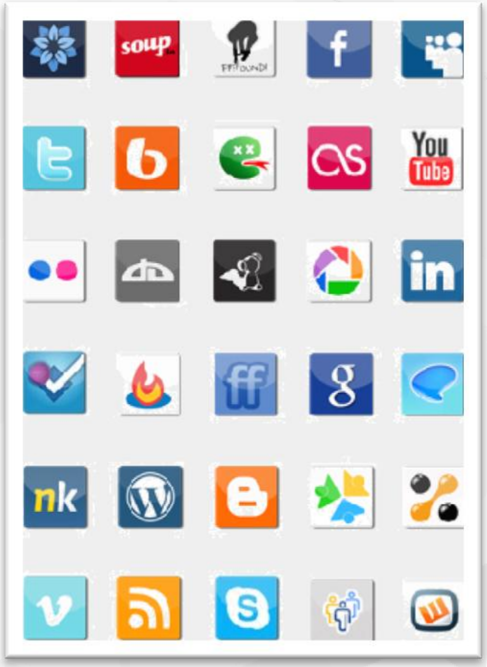
"the wildly read-write Web"

80,000,000 sites

1 billion+ global users

2006

> 600,000,000 sites



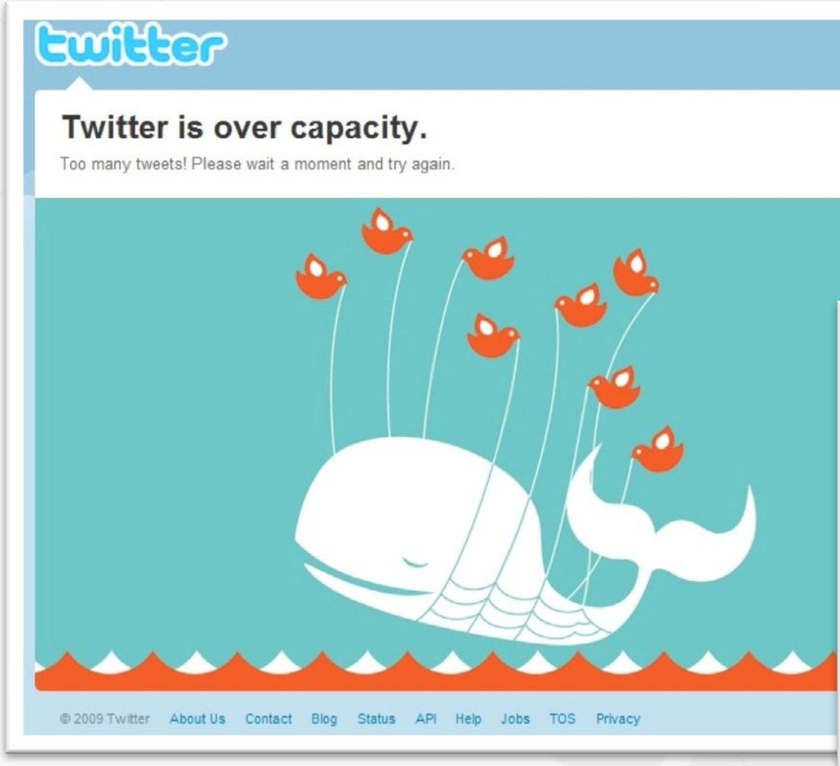
2.2 billion global users

2012

?

Reliable Services?

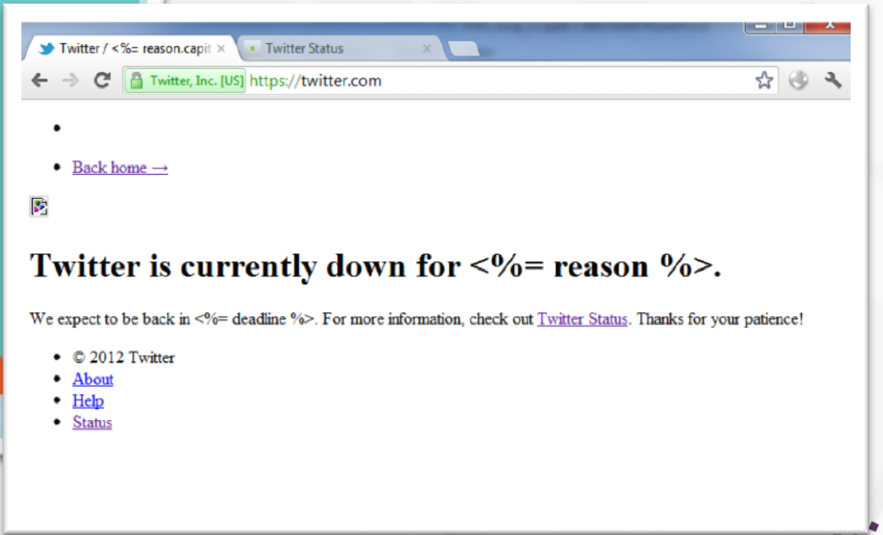
- The famous Twitter’s Fail Whale:



Twitter is over capacity.

Too many tweets! Please wait a moment and try again.

Below: The Fail Whale failed on July 27th 4am UTC



Twitter / <%= reason.capit
Twitter Status
Twitter, Inc. [US] https://twitter.com

-
- Back home →

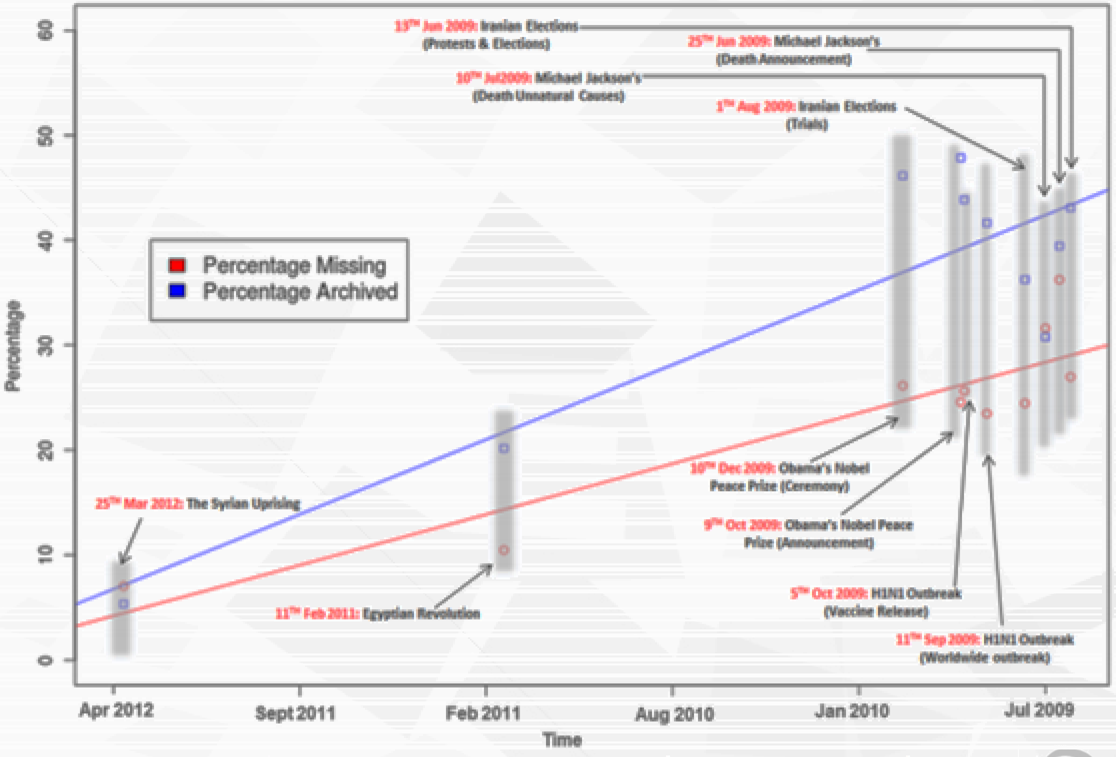
Twitter is currently down for <%= reason %>.

We expect to be back in <%= deadline %>. For more information, check out Twitter Status. Thanks for your patience!

- © 2012 Twitter
- About
- Help
- Status

The Vanishing Web Contents

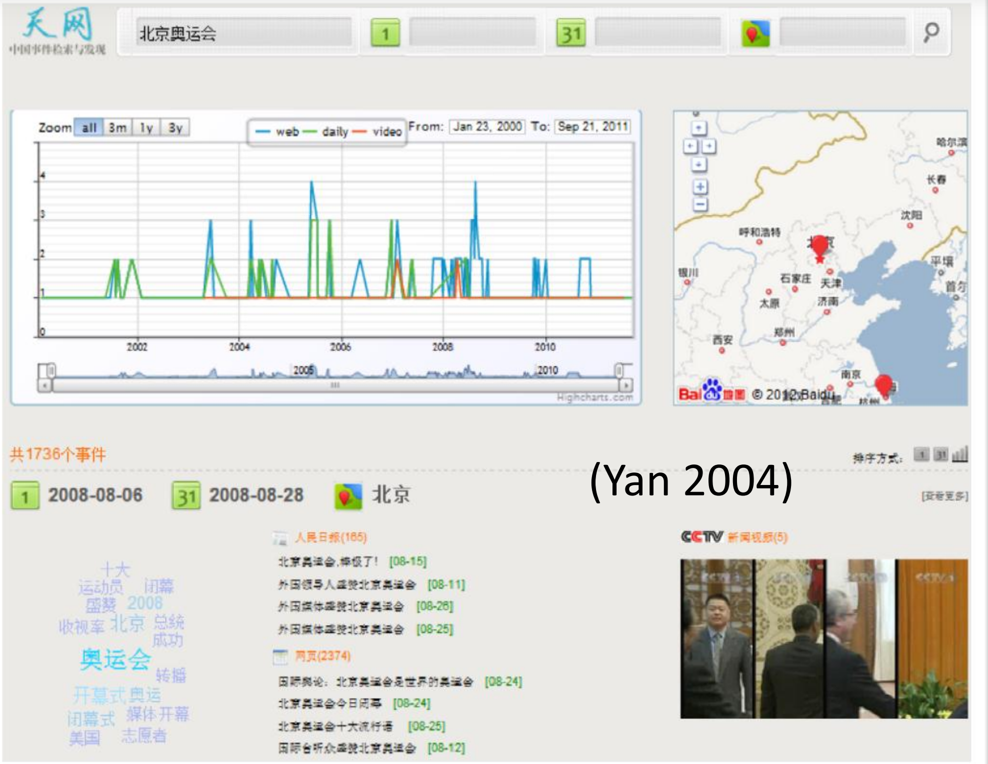
- Almost 30% of recorded history shared over social media has disappeared. (SalahEldeen 2012)



Preservation: Web 1.0 => 2.0

(Kahle 1997)

(Albertsen 2003)



共1736个事件

1 2008-08-06　31 2008-08-28　北京

人民日报(165)

北京奥运会,棒极了！ [08-15]

外国领导人盛赞北京奥运会 [08-11]

外国媒体盛赞北京奥运会 [08-26]

外国媒体盛赞北京奥运会 [08-25]

网页(2374)

国际舆论：北京奥运会是世界的奥运会 [08-24]

北京奥运会今日闭幕 [08-24]

北京奥运会十大流行语 [08-25]

国际台听众盛赞北京奥运会 [08-12]

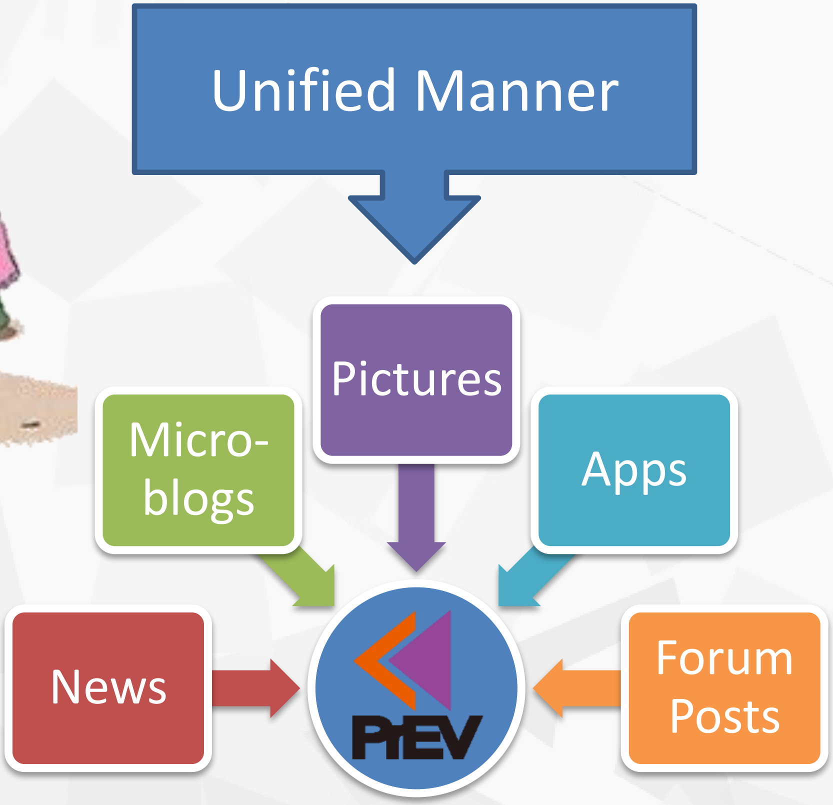
(Yan 2004)

Above: Internet Archive
Right: Country-wide:
Norway’s Paradigma
China’s Infomall

Piecemeal => A Unified Manner

Piecemeal

(Campbell 2009, Hockx-Yu 2011)



Unified Manner

Pictures

Micro-blogs

Apps

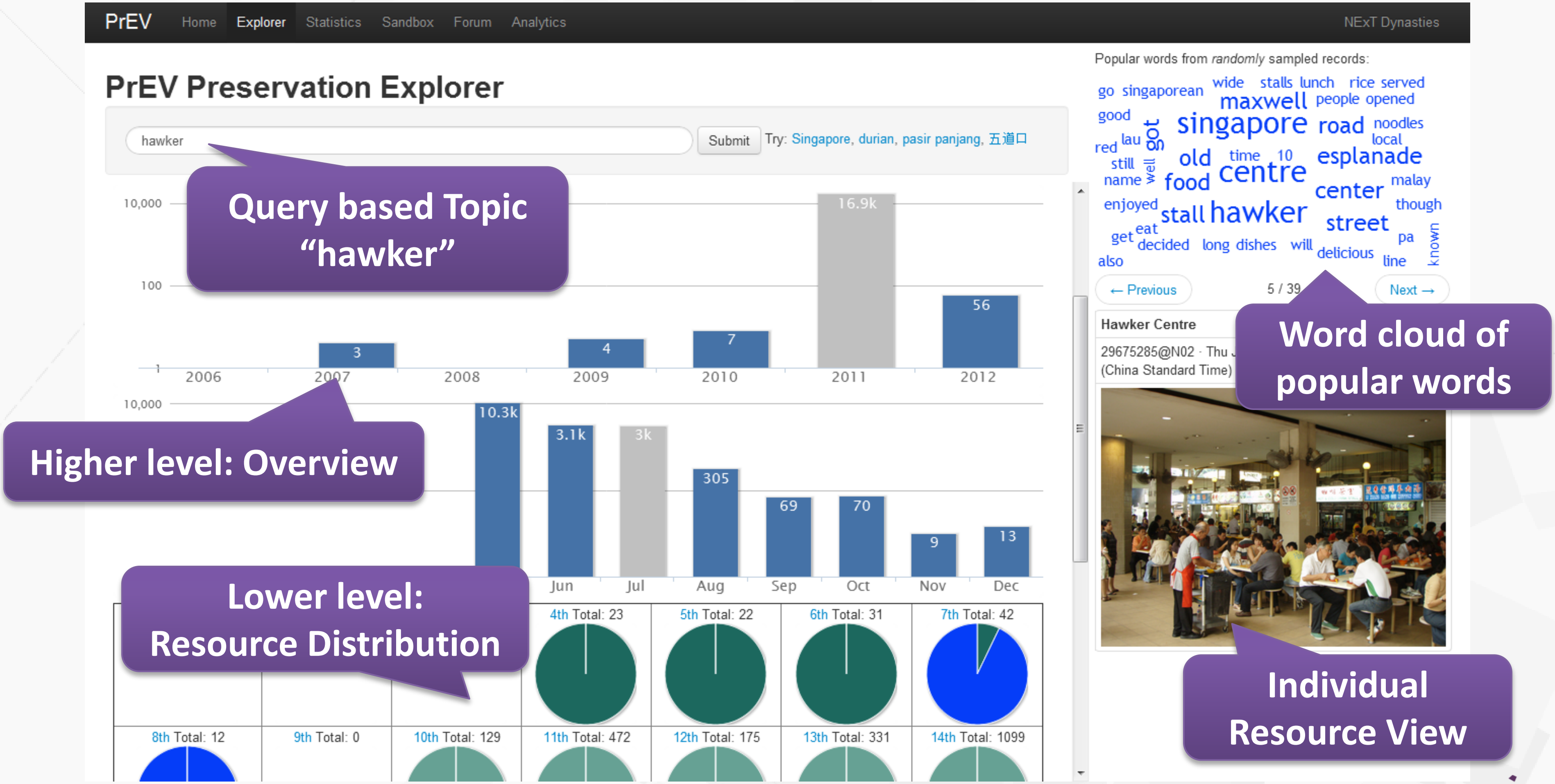
News

Forum Posts

Scenario 1: Summarizing the Data

- Ryan’s course project in 2022: Singaporean hawker center (food court) history

Scenario 1: Query-based View



PrEV Preservation Explorer

Try: Singapore, durian, pasir panjang, 五道口

Popular words from *randomly* sampled records:

go singaporean wide stalls lunch rice served
maxwell people opened
good got singapore road noodles local
red lau
still well old time 10 esplanade
name food centre malay
center
enjoyed stall hawker though
eat street
get decided long dishes will delicious pa known
also line
← Previous
5 / 39
Next →

Hawker Centre

29675285@N02 · Thu
(China Standard Time)

Query based Topic "hawker"

Higher level: Overview

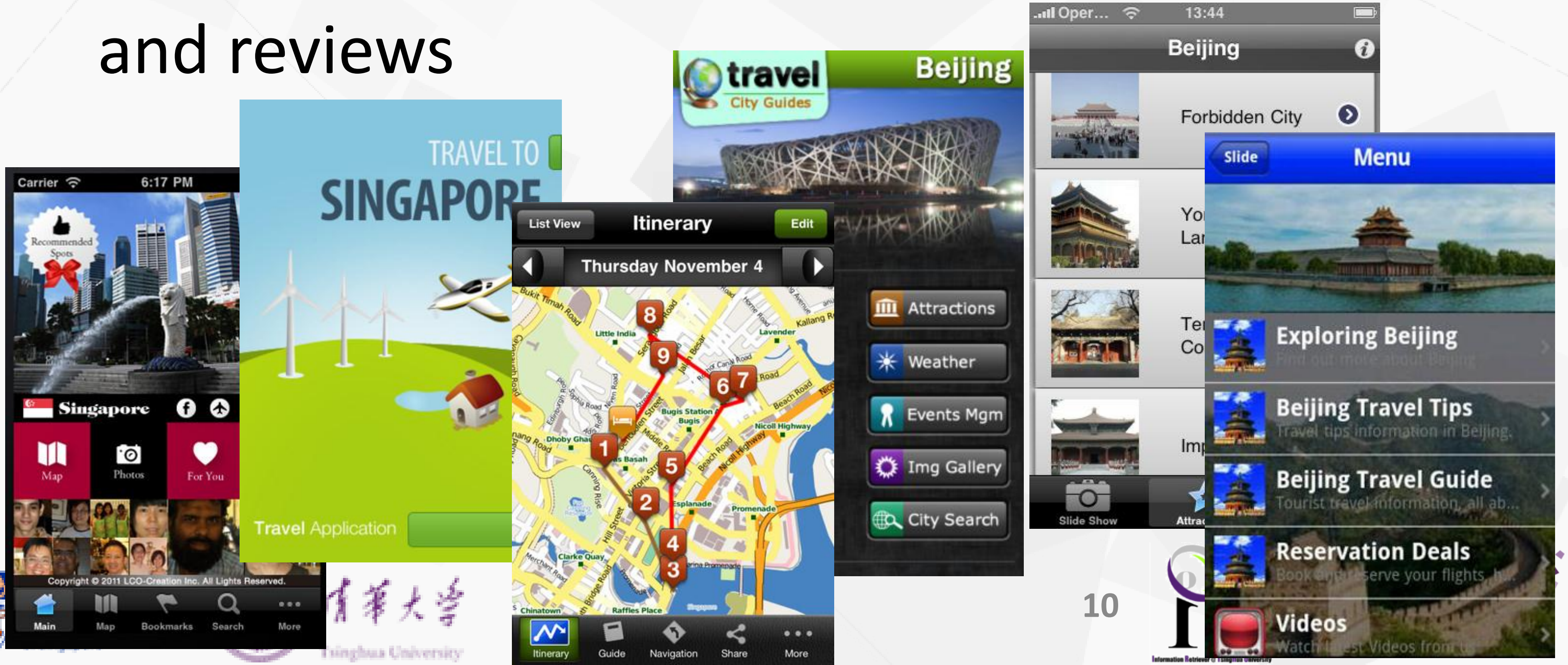
Lower level: Resource Distribution

Word cloud of popular words

Individual Resource View

Scenario 2: Historical Data Archive

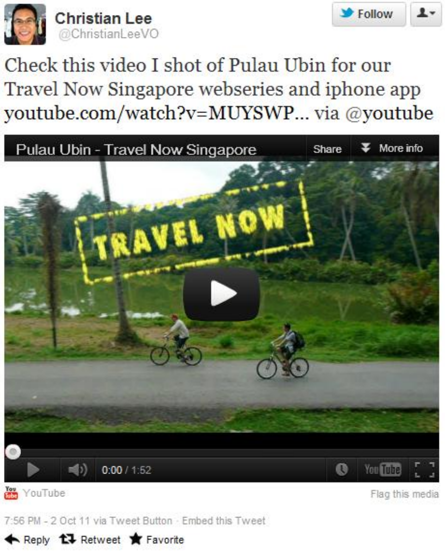
- Blueberry Inc.: Develops “Follow Me” iPhone travel guide app for BJ & SG
- Review existing competitor apps’ descriptions and reviews



Scenario 2: API for Large-scale Access

```
{ "total":86, "count":10, "totalpage":9, "page":0, "data":[
  { "crawlresource":"twitter", "encoding":"en", "tweetcreatedat":"Sun Oct 02 19:56:51 +0800 2011",
1   "url":"http://twitter.com/#!/ChristianLeeVO/status/120467276104339457",
    "maincontent":"Check this video I shot of Pulau Ubin for our Travel Now Singapore webseries and iphone
app http://t.co/mvfgJDqP via @youtube" },
2 { "crawlresource":"weibo", "encoding":"zh", "weibocreatedat":"Sat Jan 14 19:13:27 +0800 2012",
    "url":"http://www.weibo.com/1910529591/y0Lf5mFAk",
    "maincontent":"#App推荐# 旅程规划: Routes. Planning your journeys【出行必备】, iPad/ iPhone通用。这款应用
可让你规划旅游景点，像是到某景点去拜访或是去某家大卖场购物等等。它会算出需要多远的距离以及所需的时间，就像...
http://t.cn/z0gtfEr （分享自 @App每日推送）" },
3 { "crawlresource":"sgbjapps", "encoding": "others", "crawltime":"Wed Dec 23 00:00:00 +0800 2009",
    "maincontent":"Do You Love Travel ? If Yes, You Should Not Miss This App. Updated For Now! Download
this app to your iPhone to enjoy these beautiful scenery anywhere you go! These pictures are HD Photo You
can download the image to your iPhone or iPod and make it to wallpaper. No Ad No Wifi!",
    "name":"A Tourist Paradise <Singapore>" }, ... ] }
```

1



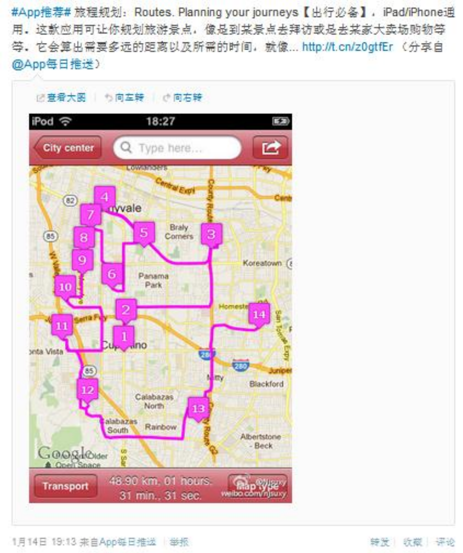
Christian Lee
@ChristianLeeVO

Check this video I shot of Pulau Ubin for our Travel Now Singapore webseries and iphone app youtube.com/watch?v=MUYSWP... via @youtube

Pulau Ubin - Travel Now Singapore

7:56 PM - 2 Oct 11 via Tweet Button · Embed this Tweet

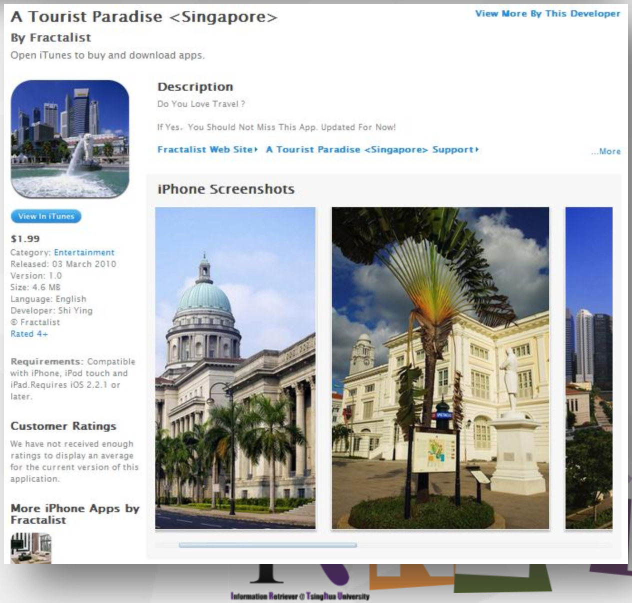
2



#App推荐# 旅程规划：Routes. Planning your journeys【出行必备】，iPad/iPhone通用。这款应用可让你规划旅游景点，像是到某景点去拜访或是去某家大卖场购物等等。它会算出需要多远的距离以及所需的时间，就像... http://t.cn/z0gtfEr （分享自 @App每日推送）

1月14日 19:13 来自App每日推送 | 举报

3



A Tourist Paradise <Singapore>

By Fractalist

Open iTunes to buy and download apps.

Description

Do You Love Travel ?

If Yes, You Should Not Miss This App. Updated For Now!

Fractalist Web Site ▸ A Tourist Paradise <Singapore> Support ▸ ...More

$1.99
Category: Entertainment
Released: 03 March 2010
Version: 1.0
Size: 4.6 MB
Language: English
Developer: Shi Ying
© Fractalist
Rated 4+

Requirements: Compatible with iPhone, iPod touch and iPad.Requires iOS 2.2.1 or later.

Customer Ratings

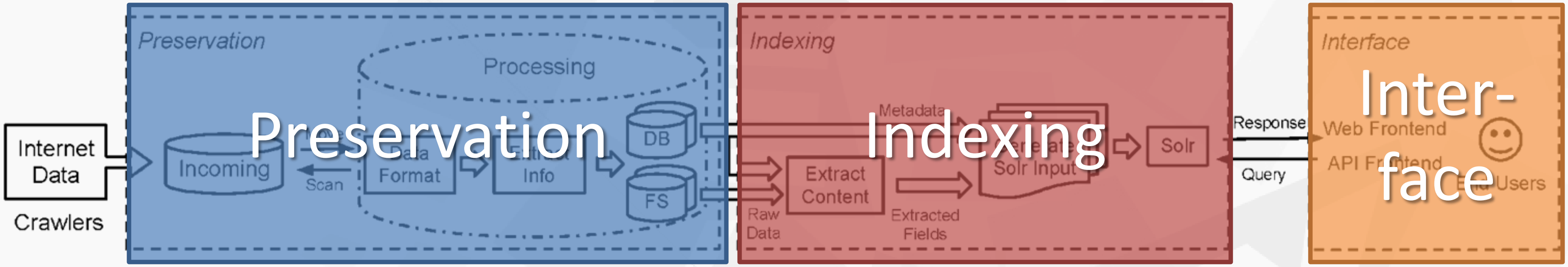
We have not received enough ratings to display an average for the current version of this application.

More iPhone Apps by Fractalist

iPhone Screenshots

System Architecture

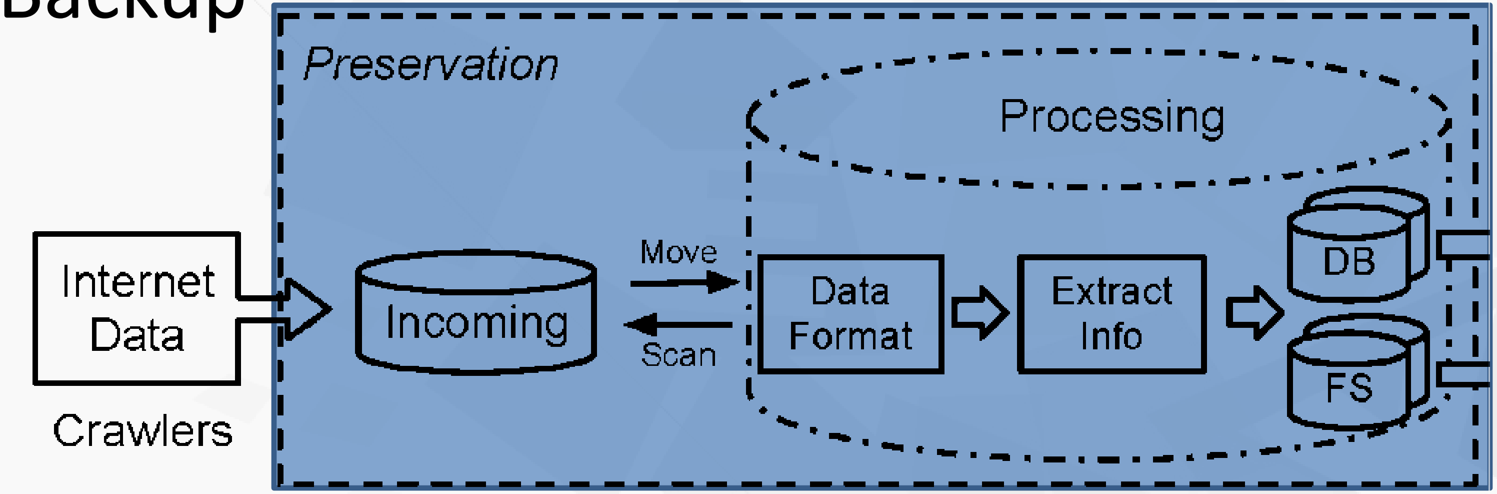
- Three layers, loosely coupled
  - Preservation
  - Indexing
  - Interface



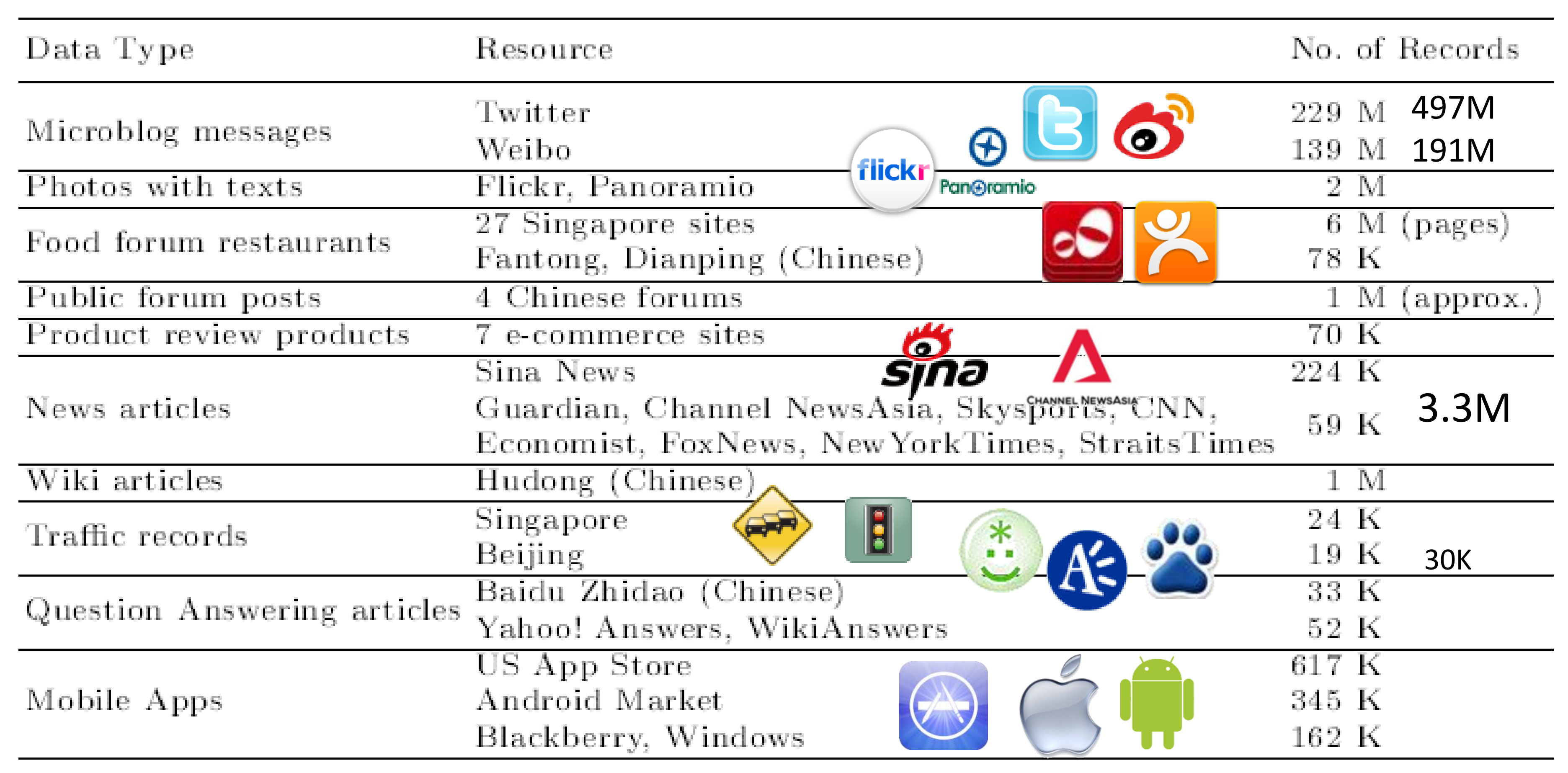
1. Preservation Layer

Preservation | Indexing | Inter-face

- Incoming Data Detection
- Data Format Recognition
- Record Storage
- Backup



1. Collected Resources (as of May 2012)

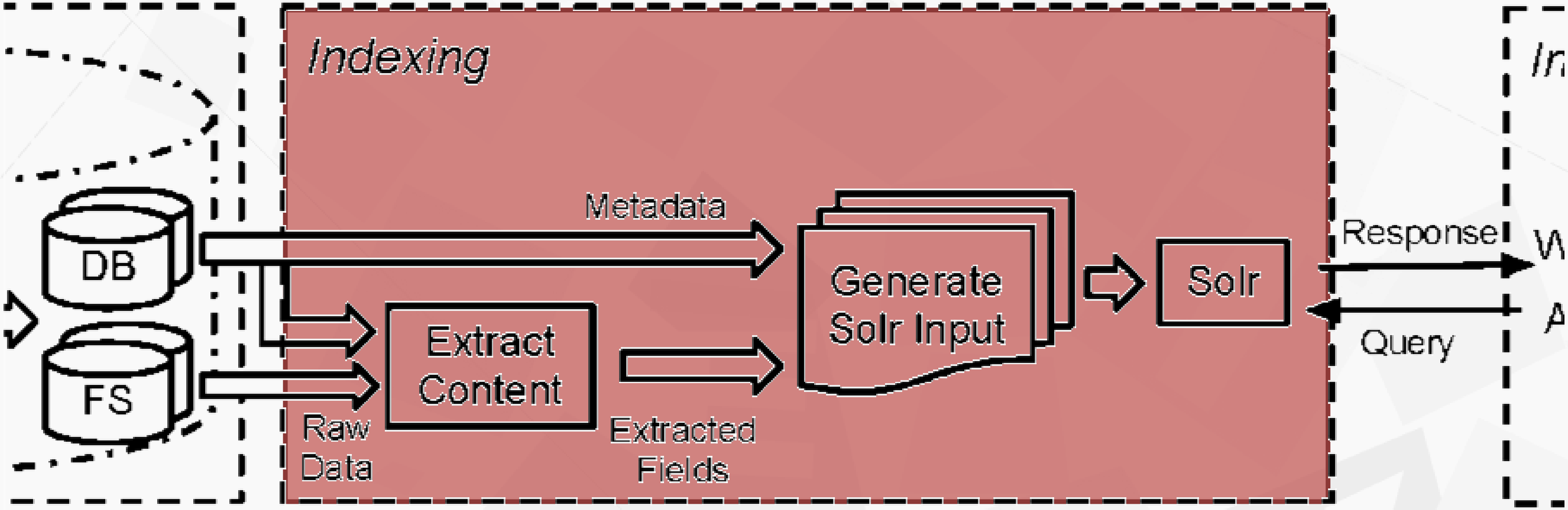


Data Type	Resource	No. of Records	
Microblog messages	Twitter	229 M	497M
	Weibo	139 M	191M
Photos with texts	Flickr, Panoramio	2 M	
Food forum restaurants	27 Singapore sites	6 M (pages)	
	Fantong, Dianping (Chinese)	78 K	
Public forum posts	4 Chinese forums	1 M (approx.)	
Product review products	7 e-commerce sites	70 K	
News articles	Sina News	224 K	
	Guardian, Channel NewsAsia, Skysports, CNN, Economist, FoxNews, New YorkTimes, StraitsTimes	59 K	3.3M
Wiki articles	Hudong (Chinese)	1 M	
Traffic records	Singapore	24 K	
	Beijing	19 K	30K
Question Answering articles	Baidu Zhidao (Chinese)	33 K	
	Yahoo! Answers, WikiAnswers	52 K	
Mobile Apps	US App Store	617 K	
	Android Market	345 K	
	Blackberry, Windows	162 K	

*Updates are of Sep 2012. Please refer to the website for an up-to-date statistics.

2. Indexing Layer

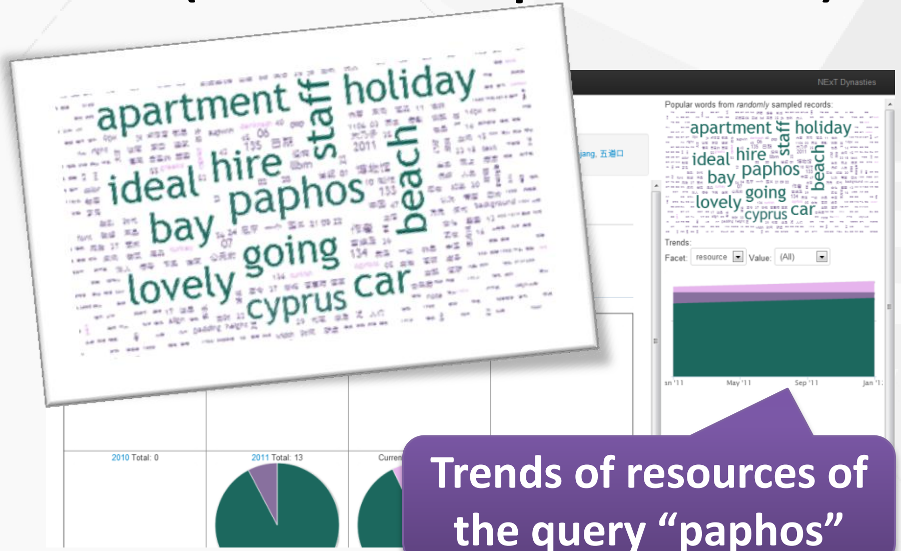
Preservation | Indexing | Inter-face



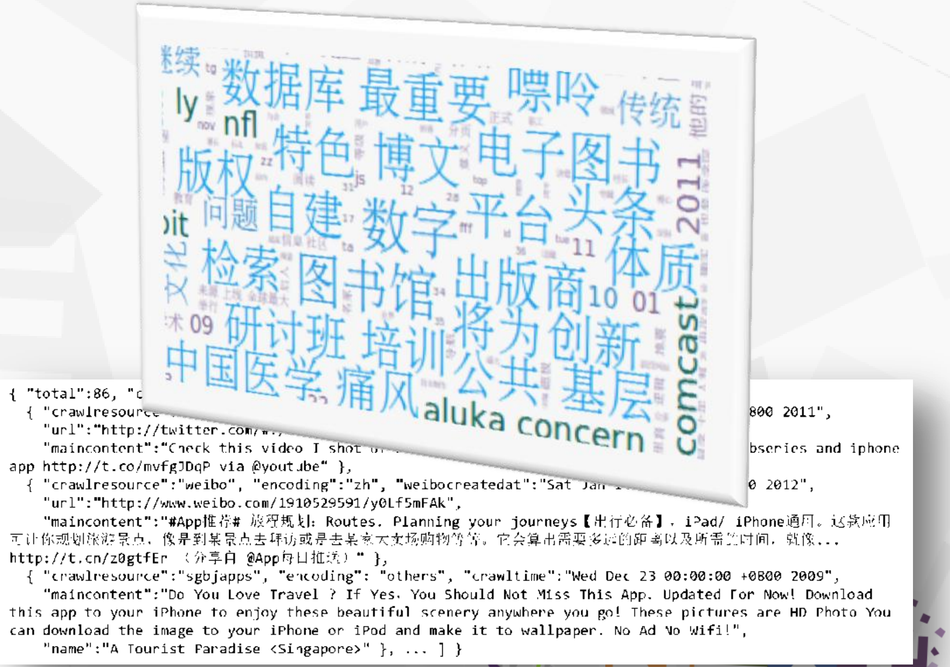
3. Interface Layer

Preservation | Indexing | Inter-face

- Web frontend: Calendar view, word cloud, trends view, individual view
- API Frontend: Authentication, rate limiting (sandbox provided)



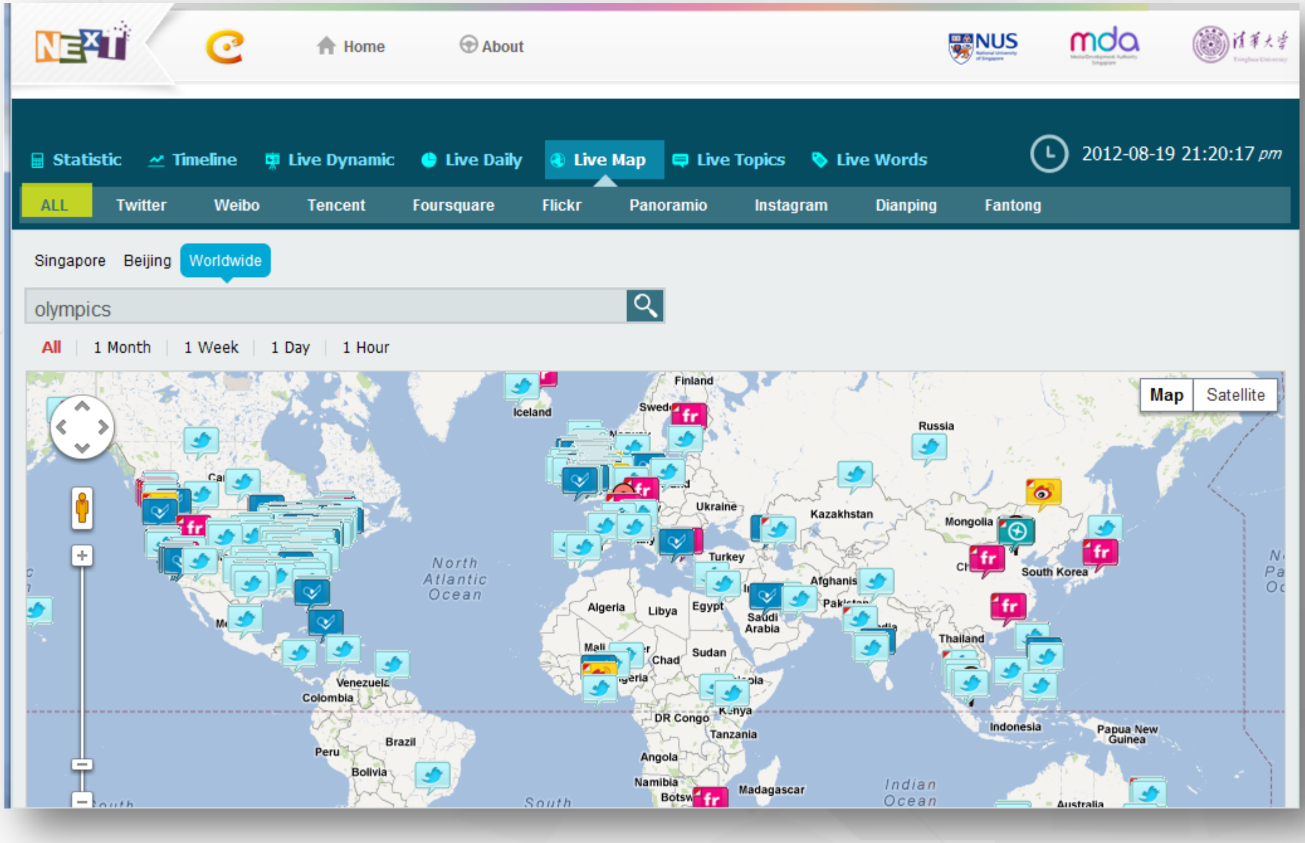
Trends of resources of the query "paphos"

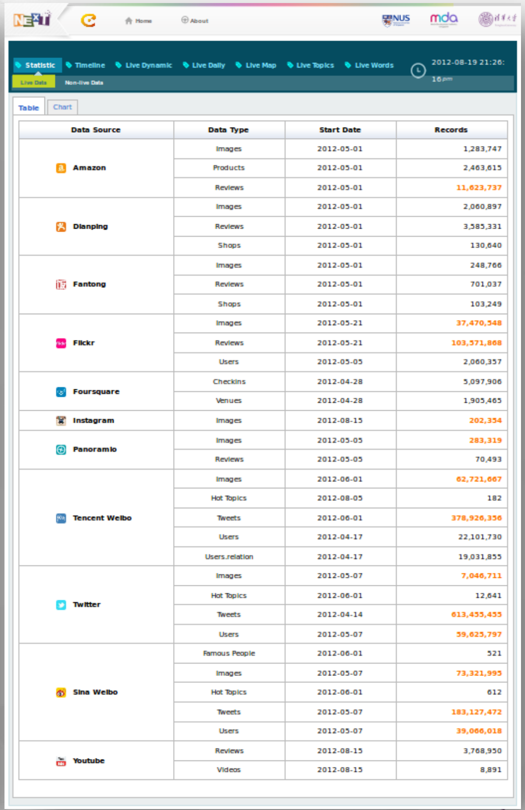


```
{ "total":86, "c
  { "crawlresource                                                                        800 2011",
    "url":"http://twitter.com/#!/
    "maincontent":"Check this video I shot o                                              bseries and iphone
app http://t.co/mvfgJDqP via @youtube" },
  { "crawlresource":"weibo", "encoding":"zh", "weibocreatedat":"Sat Jan 1                 0 2012",
    "url":"http://www.weibo.com/1910529591/y0Lf5mFAk",
    "maincontent":"#App推荐# 旅程规划: Routes. Planning your journeys【出行必备】, iPad/ iPhone通用。这款应用
可让你规划旅游景点，像是到某景点去拜访或是去某家大卖场购物等等。它会算出需要移动的距离以及所需的时间，就像...
http://t.cn/z0gtfEr （分享自 @App每日推送）" },
  { "crawlresource":"sgbjapps", "encoding": "others", "crawltime":"Wed Dec 23 00:00:00 +0800 2009",
    "maincontent":"Do You Love Travel ? If Yes, You Should Not Miss This App. Updated For Now! Download
this app to your iPhone to enjoy these beautiful scenery anywhere you go! These pictures are HD Photo You
can download the image to your iPhone or iPod and make it to wallpaper. No Ad No Wifi!",
    "name":"A Tourist Paradise <Singapore>" }, ... ] }
```

NExT Live UGC Web Portal

- http://137.132.145.151:8080/ugcp/





Data Source	Data Type	Start Date	Records
Amazon	Images	2012-05-01	1,283,747
	Products	2012-05-01	2,463,615
	Reviews	2012-05-01	11,623,737
Dianping	Images	2012-05-01	2,060,897
	Reviews	2012-05-01	3,585,331
	Shops	2012-05-01	130,640
Fantong	Images	2012-05-01	248,766
	Reviews	2012-05-01	701,037
	Shops	2012-05-01	103,249
Flickr	Images	2012-05-21	37,470,548
	Reviews	2012-05-21	103,571,868
	Users	2012-05-05	2,060,357
Foursquare	Checkins	2012-04-28	5,097,906
	Venues	2012-04-28	1,905,465
Instagram	Images	2012-08-15	202,354
Panoramio	Images	2012-05-05	283,319
	Reviews	2012-05-05	70,493
Tencent Weibo	Images	2012-06-01	62,721,667
	Hot Topics	2012-08-05	182
	Tweets	2012-06-01	378,926,356
	Users	2012-04-17	22,101,730
	Users.relation	2012-04-17	19,031,855
Twitter	Images	2012-05-07	7,046,711
	Hot Topics	2012-06-01	12,641
	Tweets	2012-04-14	613,455,455
	Users	2012-05-07	59,625,797
Sina Weibo	Famous People	2012-06-01	521
	Images	2012-05-07	73,321,995
	Hot Topics	2012-06-01	612
	Tweets	2012-05-07	183,127,472
	Users	2012-05-07	39,066,018
Youtube	Reviews	2012-08-15	3,768,950
	Videos	2012-08-15	8,891

Conclusion & Future Work

- What is PrEV?
  - PrEV: City-centric archiving system
  - Archive & unify multilingual Web 2.0 data
- Whom is it for?
  - Individual users: Discover the old good days
  - Enterprise-level use: Programmatically access a large amount of data for business and scientific research
- How is PrEV built?
  - Preservation layer: Collect data from different groups
  - Indexing layer: Faceted search
  - Interface layer: Flexible for different needs
- Future: System performance, user interfaces (visualization)
  - StrmWrd: The visualization tool https://github.com/THUNUS/StrmWrd

Questions?

- Website:

  http://han.ddns.comp.nus.edu.sg/prev/ or http://tinyurl.com/prevweb
  (mobile devices supported)

- Contact:
  - Anqi Cui (Google+), @CAQ9 (Twitter)

About NExT

- Crawling & mining UGCs in SG and BJ in:
  - Location-oriented: shared photos and check-in venues;
  - Topic-oriented: forums, question-answering, tweets;
  - Application-oriented: mobile applications and associated information and discussions; and
  - structured: factual, cultural and historical information.
- Carrying out research projects in the areas of: Extreme Database, Live Event Capturing and Sharing, Live Media Processing, **Live Text Search**, Live City
- http://next.comp.nus.edu.sg/

Extreme Text Search Group

- Extreme search in text: real-time search + faceted search
  - Interesting research directions
  - Commercial and industrial applications for the next generation web
- Topics to be explored includes:
  - Twitter sentiment analysis
  - Mobile app ranking
  - Social and differential news analysis