



# ArcLink

## Additional API support for Wayback Machines

Ahmed AlSum
PhD Candidate
Old Dominion University

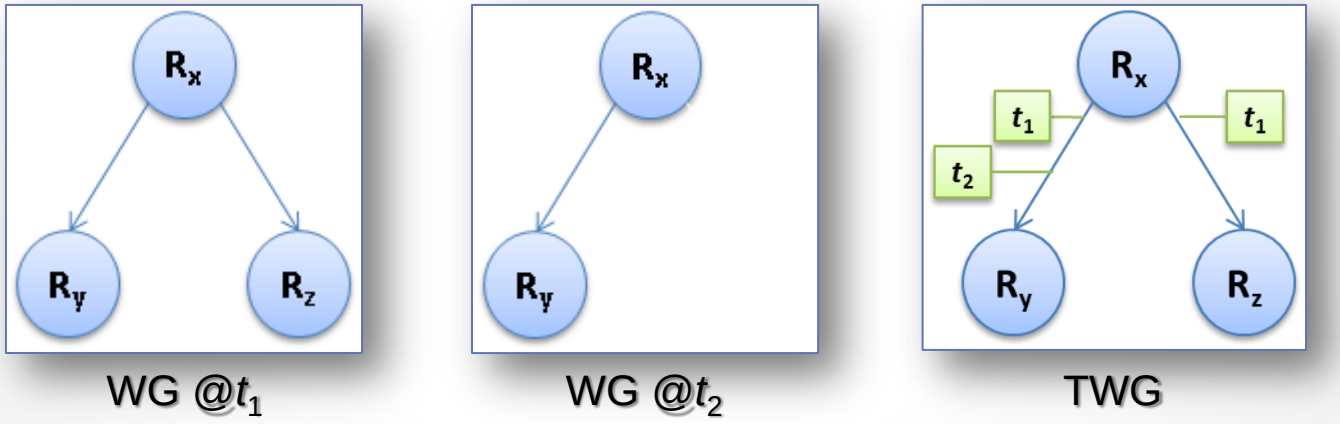
# Introduction

## What is ArcLink?

- ArcLink is a complete system to Extract, Preserve, and Access to Temporal Web Graph.

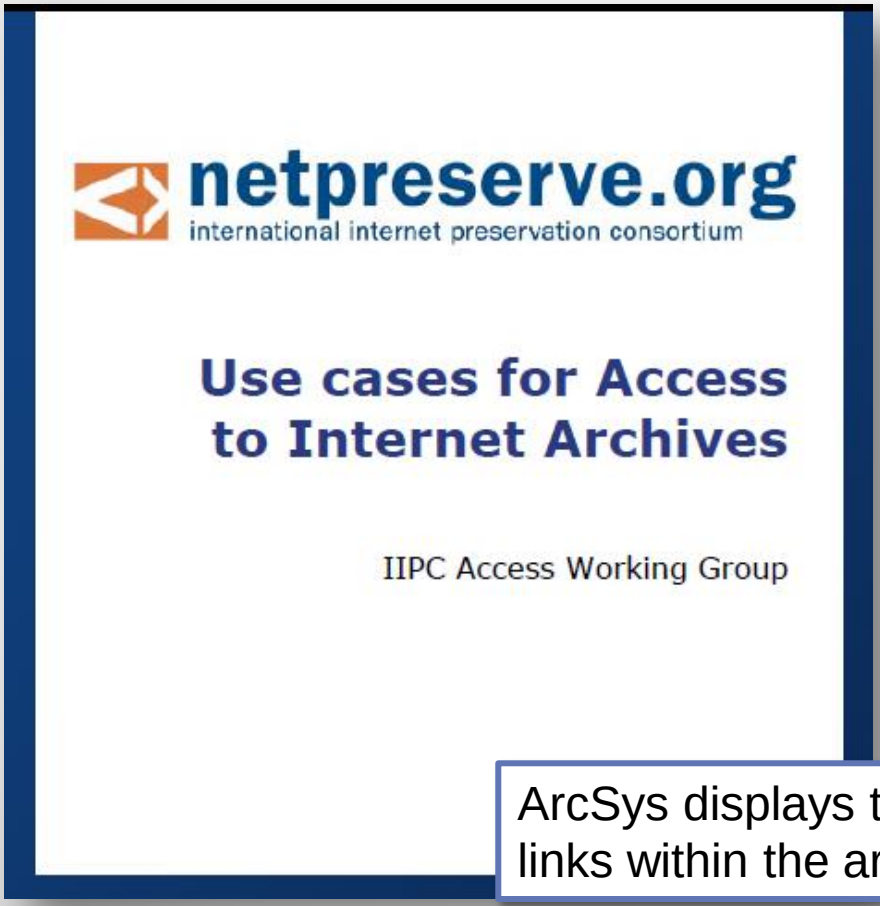
## What is Temporal Web Graph?

- Link structure through the time, including inlinks and outlinks.



# Motivations

# IIPC Use-cases




### 4.3 Advanced Searching with version comparison, linking information

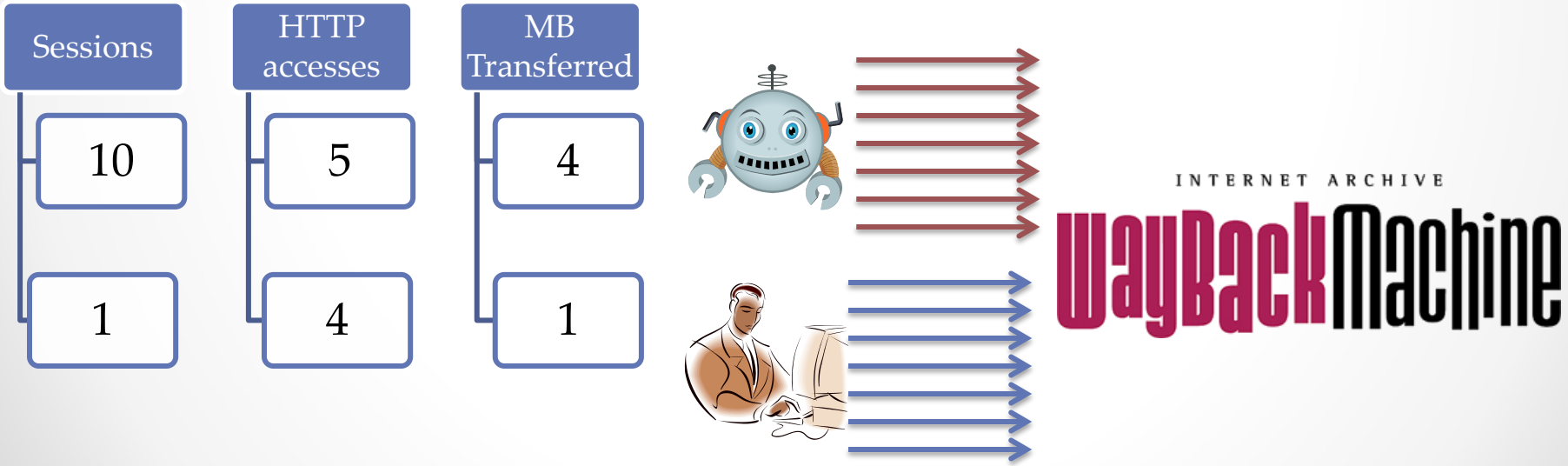
Jane Jones is studying public web sites of different municipalities between 2004-2006 in order to determine how the municipality presents local information to the community. She has narrowed her study to tourism information and local real estate information. Furthermore she also wants to study how single documents have evolved through time (version control, difference detection). Also one aspect is to study how different documents relate to one another (linking information). The ArcSys should be able to provide some statistical data about these issues.

- She uses the UI functions to narrow the results down by date.
- She also narrows it by to cover only 25 different municipalities (narrow by site/host/domain).
- She performs the query.
- She gets the result list but she wants to change the narrow-down parameters. She clicks the "modify search" button and does the changes.
- She gets the results she wants and she selects one page relevant to her study. She views it and then returns to the result list and selects "show the versions".
- She gets the list for all the versions of that particular page. She selects two of those pages and asks the ArcSys to determine the differences between these two pages.
- ArcSys displays the difference results (percentage, word count, link count, image count).

ArcSys displays the linking information (known incoming links within the archive, outgoing links, internal links, etc.).

# Serving Robots!

- Alnoamany[1] reported access to IA wayback machine as **R**obots outnumber **H**umans
  - **10:1** in terms of sessions,
  - **5:4** in terms of raw HTTP accesses
  - **4:1** in terms of megabytes transferred.




[1]ALNOAMANY, Y., WEIGLE, M.C. AND NELSON, M.L., 2013. Access Patterns for Robots and Humans in Web Archives. In *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital Libraries*. JCDL '13.

# Solved Questions

**What are the titles for www.vancouver2010.com through time?**

- Get the TimeMap
- Do page-scraping for each memento
- Extract the Title header

# Solved Questions, *but hard*

| Date | Title |
|---|---|
| 02 Dec 1998 | 2010 Olympic Bid |
| 05 Apr 2001 | Welcome to the Vancouver - Whistler 2010 Bid Corporation Website |
| 18 Jan 2002 | Welcome to the Vancouver/Whistler Winter 2010 Bid Corporation Website |
| 31 Mar 2002 | Welcome to the Vancouver 2010 Bid Corporation Website |
| 23 Sep 2002 | The official site of the Vancouver 2010 Bid Corporation |
| 04 Feb 2006 | Vancouver 2010 - Welcome |
| 30 Apr 2009 | Olympics | 2010 Vancouver Olympic Games Medals Results Schedule Sports |
| 03 Nov 2009 | 2010 Vancouver Olympic Games Medals Results Schedule Sports : Vancouver 2010 Winter Olympics |
| 18 Dec 2009 | 2010 Vancouver Olympic Games Medals Results Schedule Sports : Vancouver 2010 Winter Olympics and Paralympics |
| 07 Feb 2010 | Vancouver Olympic Games Medals Results Sports : Vancouver 2010 Winter Olympics |
| 05 Mar 2010 | Olympic Games Medals, Results, Sports : Vancouver 2010 Winter Olympics |
| 02 Feb 2011 | Vancouver 2010 Winter Olympics | Olympic Games Photos, Videos, & News - Olympic.org |
| 11 May 2011 | vancouver 2010 Winter Olympics | Olympic Videos, Photos, News |
| 16 Dec 2011 | Jeux olympiques d'hiver de vancouver 2010 | vancouver Vidéos, Photos, Media olympique |
| 21 Dec 2012 | Vancouver 2010 Winter Olympics | Olympic Videos, Photos, News, Medals |
| 08 Jan 2013 | Vancouver 2010 Winter Olympics | Olympic Video, Medals, News |

# Unsolved Questions

**What are the anchor-text that pointed to www.vancouver2010.com through time?**

# Researchers use Page-scraping

- Researchers crawled the web archive to build their corpus. For example,
  - Weber[1] built Internet Archive Crawler (HistoryCrawl) to examine the evolution of content and hyperlink networks between websites.
  - Brügger[2] discussed the challenges of crawling the web archives as part of his study on Danish parliamenary elections.

[1]WEBER, M.S., 2012. Newspapers and the Long-Term Implications of Hyperlinking. *Journal of Computer-Mediated Communication*, 17(2), pp.187–201.
[2]BRÜGGER, N., 2012. Historical Network Analysis of the Web. *Social Science Computer Review*.

# It's More than WAT files

| WAT | ArcLink |
| --- | --- |
| Batch Process on a set of WARCs | Batch process on a set of URIs |
| For internal use | For public use |
| No-way to integerate with others WAT files in others locations | It could be aggregated with other graphs |
| No incremental update | Support incremental update |
| Access on WAT file level using Pig | Access on URI level using Web service |

# Dataset

Winter Olympics 2010 collection

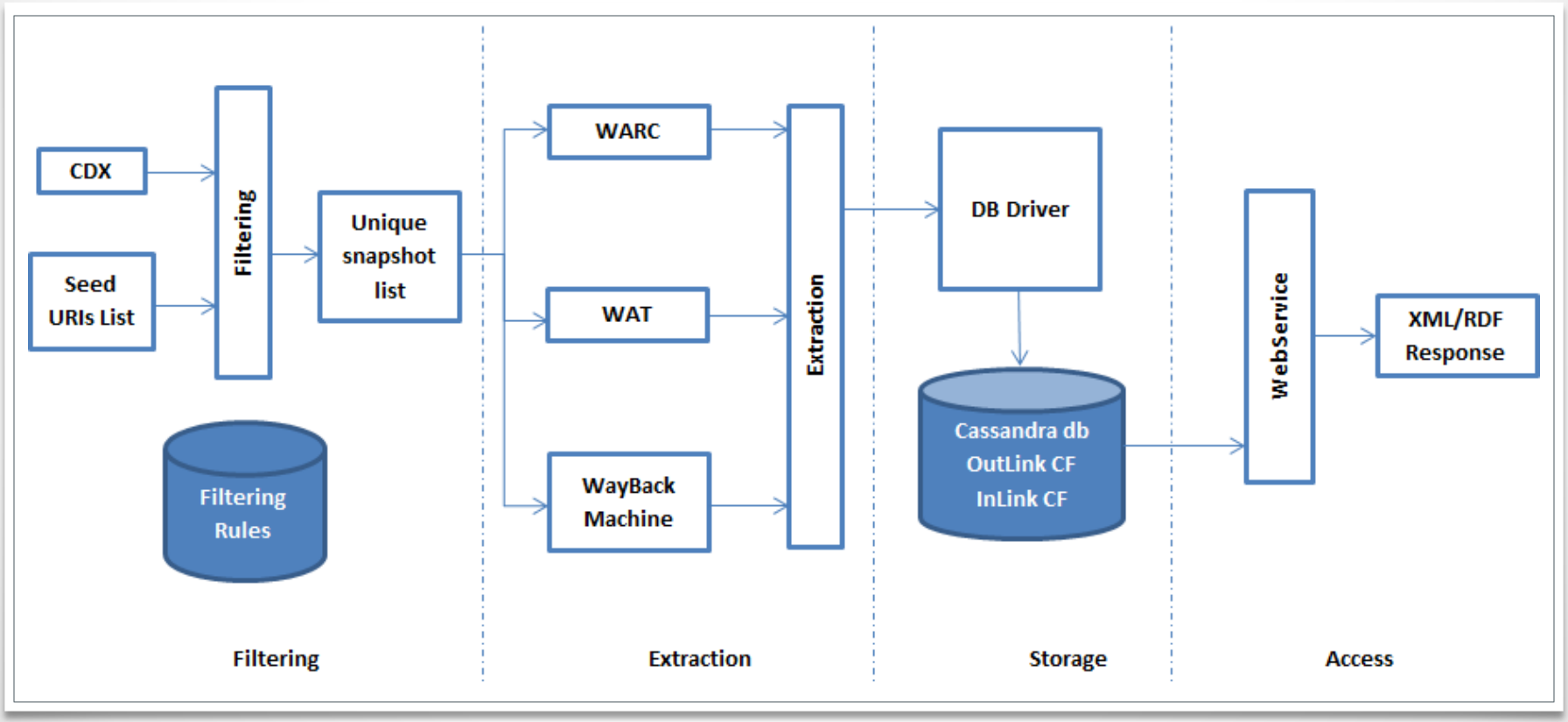
| | |
|---|---|
| Size | 700GB+ |
| From | Nov 2009 |
| To | Mar 2010 |
| #URI-R | 6.4M |
| #URI-M | 23.7M |


* http://olympics.us.archive.org/olympics2010/

# System Stages

Filtering – Extraction – Preservation - Access

# System Stages



# Filtering

- Using CDX files to filter the URI to select the mementos that will contribute to the Web Graph.
- For example,
  - Exclude non-200 HTTP status code
  - Exclude Images, style-sheets, videos, etc
  - Exclude duplicate mementos
- Technique: Using Pig Latin script on CDX files
- Results: CDX was reduced to 25% of the original size

# Extraction

- Technique: Hadoop
- Step 1: URI-ID generation
  - Canonicalized the URI into SURT format
  - Hash the canonicalized format using SimHash
  - Completely distributed
- Step 2: Define data sources
  - WARC
  - Web archive UI

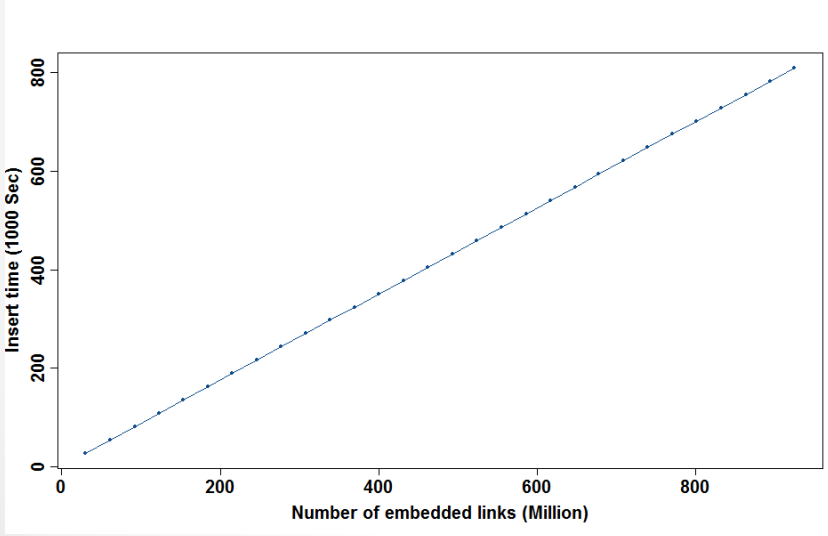
$$\left.\begin{array}{l} www.example.org/foo \\ example.org/foo \\ www1.example.org/foo \end{array}\right\} org,example)/foo$$

$$org,example)/foo \rightarrow ABCD11$$

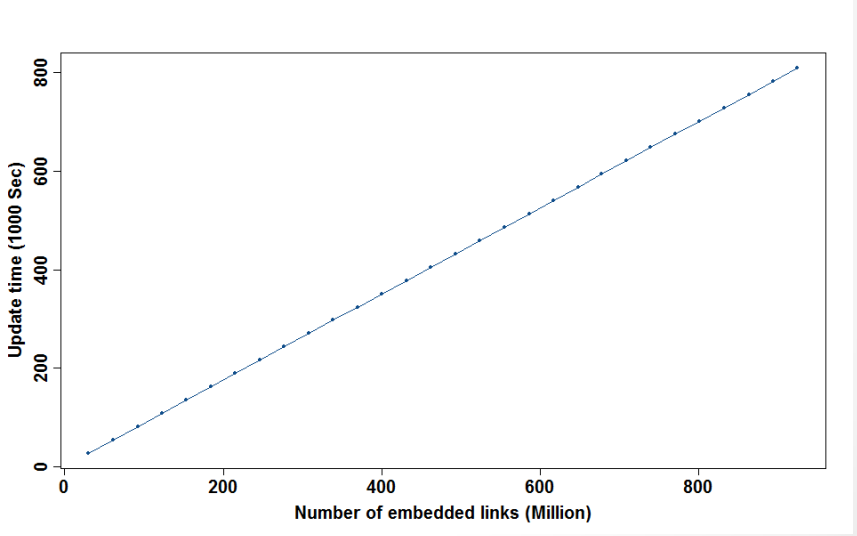
| Input | Source | Map | Reduce | Total *sec* |
|---|---|---|---|---|
| 2 Tasks | Wayback | 21,422 | 4,194 | 25,616 |
| | WARC | 13,327 | 2,770 | 16,098 (62%) |
| 5 Tasks | Wayback | 13,721 | 2,257 | 15,978 |
| | WARC | 8,304 | 1,746 | 10,051 (62%) |

# Storage

- ArcLink used Cassandra database to save the web graph




Insertion Performance




Update Performance

# Access

| | |
|---|---|
| curl command | > curl "http://localhost:8080/LinkService/linkQuery?uri=vancouver2010.com" |
| XML clause | <?xml version="1.0"?> |
| RDF clause | <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#" xmlns:twg="http://www.mementoweb.org/TemporalWebGraph/"> |
| URI | **<rdf:Description rdf:about="vancouver2010.com">** |

*To be continued…*

# Access – *Outlinks*

Outlink To

AnchorTxt

Timestamp

Outlink To

AnchorTxt

Timestamp

```
<rdf:Description rdf:about="vancouver2010.com">
…..
<twg:hasOutlinks rdf:parseType="Collection">
  <rdf:Description rdf:about="http://paralympic-games.com/news">
   <twg:type>href</twg:type> <twg:text>News</twg:text>
   <twg:timestamp>  <rdf:Bag>
    <rdf:li>20091103011307</rdf:li><rdf:li>20100130003005</rdf:li> ...
   </rdf:Bag> </twg:timestamp> </rdf:Description>

  <rdf:Description rdf:about=" http:// olympic-cross-country-skiing.com/">
   <twg:type>href</twg:type> <twg:text>Cross-Country Skiing</twg:text>
   <twg:timestamp> <rdf:Bag>
    <rdf:li>20091110011557</rdf:li> <rdf:li>20100227081100</rdf:li>  ...
   </rdf:Bag> </twg:timestamp> </rdf:Description>
.....
 </twg:hasOutlinks>
```

*To be continued…*

# Access – *Inlinks*

Inlink From

AnchorTxt

Timestamp

Inlink From

AnchorTxt

Timestamp

```
<rdf:Description rdf:about="vancouver2010.com">
…..
<twg:hasInlinks rdf:parseType="Collection">
  <rdf:Description rdf:about="http://vancouver2010.teamgb.com/gallery/gillian-cooke/">
   <twg:type>href</twg:type> <twg:text>Official Vancouver Games site</twg:text>
   <twg:timestamp> <rdf:Bag>
    <rdf:li>20100217101229</rdf:li>
   </rdf:Bag> </twg:timestamp> </rdf:Description>

    <rdf:Description rdf:about="http://swissolympic.ch/olympiablog/?tag=/verletzung">
    <twg:type>href</twg:type> <twg:text>VANOC 2010</twg:text>
   <twg:timestamp> <rdf:Bag>
    <rdf:li>20100220104902</rdf:li>
   </rdf:Bag> </twg:timestamp> </rdf:Description>
....
 </twg:hasInlinks>
</rdf:Description></rdf:RDF>
```

# Cost of Scaling Up


| | | |
|---|---|---|
| Filtering | • $Time = \frac{n}{10^6} * \frac{88}{m}$ $(sec)$<br>• $Reduction = n * 0.3$ $(mementos)$ | 58.6 hrs<br>72 * $10^9$ mementos |
| Extraction | • $Time = \frac{n}{10^6} * \frac{5.5}{m}$ $(hrs)$ | 165 days |
| Storage | • $Size = n * 10\%$ | 500 TB |

*Numbers based on Wayback Machine published statistics on Jan 2013 of 240B mementos with total size 5PB

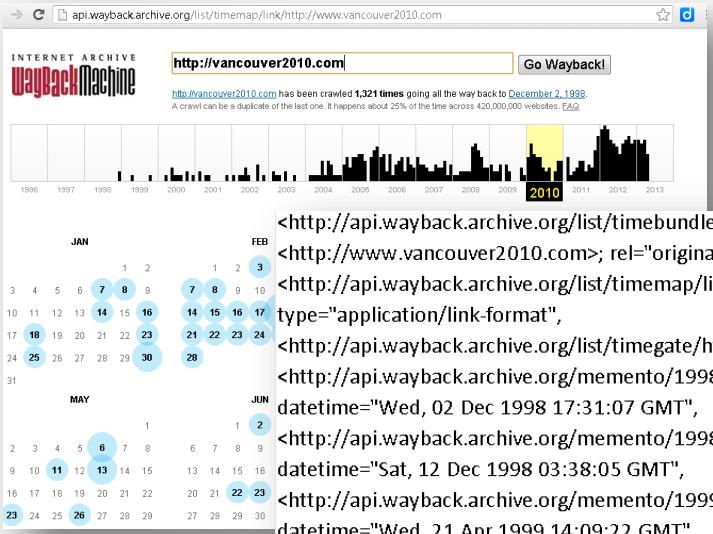
# ArcLink as Wayback Ext.

**Technical perspective**

- Both of them are URI-lookup
- Both of them are built as java web applications.

# ArcLink as Wayback Ext.

## User perspective




Wayback

Memento

ArcLink

```
<http://api.wayback.archive.org/list/timebundle/http://www.vancouver2010.com>; rel="timebundle",
<http://www.vancouver2010.com>; rel="original",
<http://api.wayback.archive.org/list/timemap/link/http://www.vancouver2010.com>; rel="timemap";
type="application/link-format",
<http://api.wayback.archive.org/list/timegate/http://www.vancouver2010.com>; rel="timegate",
<http://api.wayback.archive.org/memento/19981202173107/http://www.vancouver2010.com/>; rel="first memento";
datetime="Wed, 02 Dec 1998 17:31:07 GMT",
<http://api.wayback.archive.org/memento/19981212033805/ht
datetime="Sat, 12 Dec 1998 03:38:05 GMT",
<http://api.wayback.archive.org/memento/19990421140922/ht
datetime="Wed, 21 Apr 1999 14:09:22 GMT",
<http://api.wayback.archive.org/memento/19991012084400/ht
12 Oct 1999 08:44:00 GMT",
<http://api.wayback.archive.org/memento/20000302164857/ht
datetime="Thu, 02 Mar 2000 16:48:57 GMT",
<http://api.wayback.archive.org/memento/20000511045854/ht
datetime="Thu, 11 May 2000 04:58:54 GMT",
```

```
<?xml version="1.0"?>
- <rdf:RDF xmlns:twg="http://www.mementoweb.org/TemporalWebGraph/"
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
  - <rdf:Description rdf:about="vancouver2010.com">
    - <twg:hasOutlinks rdf:parseType="Collection">
      - <rdf:Description rdf:about="http://www.paralympic-games.com/news">
          <twg:type>href</twg:type>
          <twg:text>News</twg:text>
        - <twg:timestamp>
          - <rdf:Bag>
              <rdf:li>20091103011307</rdf:li>
              <rdf:li>20100130003005</rdf:li>
            </rdf:Bag>
          </twg:timestamp>
        </rdf:Description>
      - <rdf:Description rdf:about="http://olympic-cross-country-skiing.com/">
          <twg:type>href</twg:type>
          <twg:text>Cross-Country Skiing</twg:text>
        - <twg:timestamp>
          - <rdf:Bag>
              <rdf:li>20091110011557</rdf:li>
              <rdf:li>20100227081100</rdf:li>
            </rdf:Bag>
          </twg:timestamp>
        </rdf:Description>
```

# Applications

• • •

Let's solve our "*Unsolved Questions*"

# Time-Indexed Inlinks Information

| Date | Anchor Text |
|---|---|
| 04-Nov-09 | vancouver2010.com |
| 11-Nov-09 | vancouver2010.com |
| 18-Nov-09 | vancouver2010.com |
| 16-Jan-10 | Vancouver 2010 Olympic Games |
| 16-Jan-10 | Vancouver 2010 Olympic Games |
| 23-Jan-10 | vancouver2010.com |
| 23-Jan-10 | 2010 Vancouver Olympic Games Medals Results Schedule Sports |
| 30-Jan-10 | 2010 Vancouver Olympic Games Medals Results Schedule Sports |
| 30-Jan-10 | vancouver2010.com |
| 30-Jan-10 | Vancouver 2010 Olympic Games |
| 13-Feb-10 | Vancouver 2010 Olympic Winter Games |
| 15-Feb-10 | Vancouver 2010 Olympic Games |
| 18-Feb-10 | Official Vancouver Games site |
| 19-Feb-10 | vancouver2010.com |
| 20-Feb-10 | Official Vancouver Games site |
| 21-Feb-10 | VANOC 2010 |

# Temporal Page Rank

| | Nov-2009 | Dec-2009 | Jan-2010 |
|---|---|---|---|
| 1 | vancouver2010.com/code | - | topsport.com/sportch/liveticker/ |
| 2 | vancouver2010.com/en/langpolicy | - | vancouver2010.com/code |
| 3 | vancouver2010.com/forgotpassword | - | canadacode.vancouver2010.com/user/register |
| 4 | vancouver2010.com/store | - | canadacode.vancouver2010.com |
| 5 | vancouver2010.com/store/index.html | - | canadacode.vancouver2010.com/explore |
| 6 | vancouver2010.com/ | - | canadacode.vancouver2010.com/user/login?destination=node/add/image |
| 7 | canadacode.vancouver2010.com | - | canadacode.vancouver2010.com/pulse |
| 8 | canadacode.vancouver2010.com/nfb-onf | - | canadacode.vancouver2010.com/challenge |
| 9 | canadacode.vancouver2010.com/contact | - | i-credible.nl |
| 10 | canadacode.vancouver2010.com/resources | - | vpzschaatsteam.nl |

| | Feb-2010 | Mar-2010 | Collection ( Nov-09 to Mar-10 ) |
|---|---|---|---|
| 1 | monlibe.liberation.fr | monlibe.liberation.fr | monlibe.liberation.fr |
| 2 | topsport.com/sportch/liveticker/ | laprovence.com/la-provence-le-faq-de-la-moderation | vancouver2010.com/code |
| 3 | lefigaro.fr | get.adobe.com/flashplayer | lefigaro.fr |
| 4 | laprovence.com/la-provence-le-faq-de-la-moderation | vancouver2010.teamgb.com /teamgb/team-behind-team-gb/filenotfound.aspx | laprovence.com/la-provence-le-faq-de-la-moderation |
| 5 | lefigaro.fr/sport | ledauphine.com | lefigaro.fr/sport |
| 6 | get.adobe.com/flashplayer | lefigaro.fr/economie | get.adobe.com/flashplayer |
| 7 | lefigaro.fr/meteo | lefigaro.fr/sport | lefigaro.fr/meteo |
| 8 | lefigaro.fr/le-talk | lefigaro.fr/actualites-a-la-une | lefigaro.fr/le-talk |
| 9 | dosb.de/de/vancouver-2010/vancouver-ticker/detail/printer.html | lemonde.fr/cgv | topsport.com/sportch/liveticker/ |
| 10 | ledauphine.com | ffs.fr/index.php | vancouver2010.com/en/langpolicy |

# How to get it

- Open source code on Google-Code
  - https://code.google.com/p/arcsys/
- Try it *Soon*

aalsum@cs.odu.edu
@aalsum