An Extensible Framework for Creating Personal Web Archives of Content Behind Authentication

Mat Kelly
Director: Michele C. Weigle
Committee: Michael L. Nelson, Yaohang Li
Background

- Internet Archive crawls and preserves webpages creating web archives

- Only public sites are preserved
Problems

• A lot of content on web is not preserved
  – e.g., Social media content
• As more people document lives on social media, importance of preserving becomes greater
• Content not preserved = heritage lost
Problems:
Unsuitability of Institutional Tools

• Overhead and learning curve is steep
• Institutional tools meant for larger scale
Problems:
Complete Lack of Preservation
State of the Art in Personal Web Archiving

- Personal web archiving tools
  - Break when target sites’ hierarchy changes
  - Produce sub-optimal archives

- Some conventional web archiving practices not easily translatable to personal web archiving
Goals of Thesis

- Show social media content can be preserved
  - With output more optimal than current offerings
- Remedy the tools’ breaking problem
  - Remotely specify target sites’ hierarchies
  - Show spec is easily adaptable to tools
- Identify and consider solutions to domain-specific nuances
- Establish section commonality between social media websites
Extent of the Unpreserved
Ways to Capture Missing Content:
Supply crawler with auth credentials

- Unsuitable for institutional crawlers
- Other Personal Web Archiving problems remain
Ways to Capture Missing Content:

“Save As” Desired Pages

- Miss metadata
- Doesn’t produce interoperable output
Ways to Capture Missing Content:
Utilize Fetching Tools

- Lose look & feel
- Difficult capturing all content desired
- Frequently sub-optimal output format
Tools Utilized In Thesis:

- Archive Facebook

  - Firefox add-on
  - Creates navigable “web archives”
  - Outputs files with original file type
  - Sequential Archiving

---

8/3/2012 MS Thesis - August 2012
Tools Utilized In Thesis:
WARCreate

- Google Chrome extension
- Creates Wayback-Compatible Web ARChive (WARC) files
- Allows page manipulation prior to generating archive

localhost:8080/wayback/20120724015446/https://www.facebook.com
Integration with Other Tools

• Wayback (WARC replay system)
  – Allows WARCreate output to be re-experienced
  – Provides content for Memento

• Memento
  – Allows temporal traversal of archived pages
  – Timegate serves as relay only to local wayback instance

• XAMPP (Client-Side Server Suite)
  – Overcome Javascript inadequacies
  – Provide foundation for replay system
Institutional vs. Personal Web Archiving
Institutional vs. Personal Web Archiving
Institutional vs. Personal Web Archiving
Institutional vs. Personal Web Archiving

Crawls WWW
Institutional vs. Personal Web Archiving

Crawls WWW

outputs

Indexes

WARC

Indexing

Heritrix
Institutional vs. Personal Web Archiving

- Crawls WWW
- Indexes
- WARC
- Outputs
- Publicly viewable
- Archive replay
Institutional vs. **Personal** Web Archiving
Institutional vs. Personal Web Archiving
Institutional vs. Personal Web Archiving
Institutional vs. Personal Web Archiving
Institutional vs. Personal Web Archiving
Problems Specific to Personal Web Archiving

• **Personalization/Authentication**
  – Different users, facebook.com, different content

• **Context**
  – Different browsing tools, different site experience

• **Output Format**
  – Ad hoc approaches are often used that lose metadata, context, content, etc.
Personalization/Authentication

- Two users, same URI, vastly different content
- One user, same URI, authentication vs. no authentication, different content
  – As shown in IA’s archive of FB
Context

- Same URI+diff devices = diff content served
- Mobile vs. PC
- Firefox vs. Chrome

<!--[if lt IE 5]>
Your browser is too old and cannot render this content.
<![endif]-->
<!--[if gte IE 9]>
...features not supported by version of IE prior to 9...
<![endif]-->
Output Format
Output Format

• Saving only HTML is not enough
• Local references need manipulation
• Browser alone is insufficient replay system
• Misses HTTP headers
  • Request & Response
  • e.g., Auth
• If headers included, inputs for personalization can be viewed

```
GET / HTTP/1.1 Host: www.facebook.com User-Agent: Mozilla/5.0 (Windows NT 5.1; rv:14.0) Gecko/20100101 Firefox/14.0.1 Accept: text/html,application/xhtml+xml,application/xml;q=0.9,*/*;q=0.8 Accept-Language: en-us,en;q=0.5 Accept-Encoding: gzip, deflate Connection: keep-alive Cookie: datr=KMo6T3jicPEdEl4pY2yFnr6F; lu=TgU4dhoSBG0ZmEnThtLeyqIA; c_user=100003509861423; fr=OKMqEWNPgver2SIX.AWXf-6Ww_7iQFPPP9sFtiMPaV0; s=Aa4dL41H8UGZ-4Lf.BQGryl; xs=1%3Am7APtsN9-ev4Vg%3A0%3A1343929509; act=1343929622029%2F3%3A2; p=1; presence=EM343929627EuserFA21B03509861423A2EstateFD sb2FOET2F_5b_5dElm2Fnu11Euct2F1343929017BEmunullEt p2S2F522900E+121343929627063EutF0EdP1EnotF0C3话剧 F

HTTP/1.1 200 OK Cache-Control: private, no-cache, no-store, must-revalidate Expires: Sat, 01 Jan 2000 00:00:00 GMT P3P: CP="Facebook does not have a P3P policy. Learn why here: http://fb.me/p3p" Pragma: no-cache X-Content-Type-Options: nosniff x-frame-options: DENY X-XSS-Protection: 1; mode=block Content-Encoding: gzip Content-Type: text/html; charset=utf-8 X-FB-Debug: uMXm8343NOon00OIE142teVECApUiEqr6s7GTwnX+S= Date: Thu, 02 Aug 2012 19:26:12 GMT Transfer-Encoding: chunked Connection: keep-alive
```
Specification and OOP

• Sites’ hierarchies resemble OOP concepts (polymorphism, inheritance)
• Sites’ sections can be represented as classes
• Classes converted to XML specification
• Personal Web Archiving tools utilize this specification to become adaptive
## Commonality of “Sections” Between Social Media Websites

<table>
<thead>
<tr>
<th>Abstracted media type</th>
<th>Facebook</th>
<th>Google+</th>
<th>Twitter</th>
</tr>
</thead>
<tbody>
<tr>
<td>personal stream</td>
<td>wall</td>
<td>posts</td>
<td>my tweets</td>
</tr>
<tr>
<td>global stream</td>
<td>news feed</td>
<td>streams</td>
<td>followees’ tweets</td>
</tr>
<tr>
<td>multimedia - photos</td>
<td>photos</td>
<td>photos</td>
<td></td>
</tr>
<tr>
<td>multimedia - videos</td>
<td>videos</td>
<td>videos</td>
<td></td>
</tr>
<tr>
<td>photo collection</td>
<td>albums</td>
<td></td>
<td></td>
</tr>
<tr>
<td>posts</td>
<td>notes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>friends</td>
<td>friends</td>
<td>circles</td>
<td></td>
</tr>
</tbody>
</table>
Example: Facebook Section Objects

```php
SocialMediaWebsite facebook =
    new SocialMediaWebsite(homepage => "http://www.facebook.com")
facebook->decorate([
    new SocialMediaWebsiteSectionPersonalStream(
        name => "Wall",
        preprocessor => new SocialMediaScrollPreprocessor(
            timeBetweenFirings => 0,
            maxFirings => 0,
            conditionBeforeSubsequentFirings = null
        )
    ),
    new SocialMediaWebsiteSectionUserInfo(
        name => "Info",
    ),
    new SocialMediaWebsiteSectionMultimediaCollection(
        name => "Photos",
        preprocessor => new SocialMediaScrollPreprocessor(
            timeBetweenFirings => 0,
            maxFirings => 0,
            conditionBeforeSubsequentFirings = null
        )
    ),
    ...
```
Example: Facebook Section Objects

SocialMediaWebsite facebook =
    new SocialMediaWebsite(homepage => "http://www.facebook.com")
facebook->decorate([
    new SocialMediaWebsiteSectionPersonalStream(
        name => "Wall",
        preprocessor => new SocialMediaScrollPreprocessor(
            timeBetweenFirings => 0,
            maxFirings = 0,
            conditionBeforeSubsequentFirings = null
        )
    ),
    new SocialMediaWebsiteSectionUserInfo(
        name => "Info",
    ),
    new SocialMediaWebsiteSectionMultimediaCollection(
        name => "Photos",
        preprocessor => new SocialMediaScrollPreprocessor(
            timeBetweenFirings => 0,
            maxFirings => 0,
            conditionBeforeSubsequentFirings = null
        )
    ),
    ...
]
Example: Facebook Section Objects

```php
SocialMediaWebsite facebook =
    new SocialMediaWebsite(homepage => "http://www.facebook.com")
facebook->decorate([
    new SocialMediaWebsiteSectionPersonalStream(
        name => "Wall",
        preprocessor => new SocialMediaScrollPrepreprocessor(
            timeBetweenFirings => 0,
            maxFirings = 0,
            conditionBeforeSubsequentFirings = null
        )
    ),
    new SocialMediaWebsiteSectionUserInfo(
        name => "Info",
    ),
    new SocialMediaWebsiteSectionMultimediaCollection(
        name => "Photos",
        proprocessor => new SocialMediaScrollPreprocessor(
            timeBetweenFirings => 0,
            maxFirings => 0,
            conditionBeforeSubsequentFirings = null
        )
    ),
    ...
```
Example: Hierarchical Similarities

SocialMediaWebsite facebook =
    new SocialMediaWebsite(homepage => "http://www.facebook.com")
facebook->decorate([
    new SocialMediaWebsiteSectionPersonalStream(
        name => "Wall",
        preprocessor => new SocialMediaScrollPreprocessor(
            timeBetweenFirings => 0,
            maxFirings = 0,
            conditionBeforeSubsequentFirings = null
        )
    ),
    new SocialMediaWebsiteSectionUserInfo(
        name => "Info",
    ),
    new SocialMediaWebsiteSectionMultimediaCollection(
        name => "Photos",
        url => "http://www.facebook.com/photos",
        preprocessor => new SocialMediaScrollPreprocessor(
            timeBetweenFirings => 0,
            maxFirings => 0,
            conditionBeforeSubsequentFirings = null
        )
    ),
    ...
Spec Retrieval Process

1. Tool accesses root spec w/ URI parameter
2. Spec returns with reference to site-specific hierarchy spec
3. Tool fetches site spec
4. Updated site hierarchy returned
Concrete Usage – Tool Adaptation

• Archive Facebook
  – Map current URIs to remotely fetched URIs
  – Perform pre-processing defined in FB spec

• WARCcreate
  – Implement sequential/cohesive archiving
Evaluation 1: Tool Adaptability

1. Setup synthetic social media website
2. Define site’s remote spec
3. Change AFB to preserve synthetic site
4. Change hierarchy of synthetic site
5. Show AFB breaking
6. Change synthetic site spec
7. Show AFB functionality restored
Evaluation 1: Tool Adaptability
Step 1: Synthetic Site Creation

- Simple hierarchy for base case testing
- Requires Auth
- Utilizes CDN
- Can be manipulated
- Recursive Sections
Evaluation 1: Tool Adaptability

Step 2: Define Site Remove Spec

Social Standard Test

Just Your Typical Social Media Website

<table>
<thead>
<tr>
<th>Peer Stream</th>
<th>Mary Johnson</th>
</tr>
</thead>
<tbody>
<tr>
<td>My Stream</td>
<td></td>
</tr>
<tr>
<td>Photos</td>
<td></td>
</tr>
</tbody>
</table>

<?xml version="1.0" ?>
<socialMediaWebsite>
  <homepage>http://test.socialstandard.org</homepage>
  <sections>
    <socialMediaWebsiteSection type="SocialMediaWebsiteSectionPersonalStream">
      <name>Personal Stream</name>
      <url>http://test.socialstandard.org/personal</url>
      <preprocessor type="SocialMediaScrollPreprocessor">
        <timeBetweenFirings>0</timeBetweenFirings>
        <maxFirings>0</maxFirings>
        <conditionBeforeSubsequentFiring>?</conditionBeforeSubsequentFiring>
      </preprocessor>
    </socialMediaWebsiteSection>
    <socialMediaWebsiteSection type="SocialMediaWebsiteSectionMultimediaCollection">
      <name>Photo Albums</name>
      <url>http://test.socialstandard.org/albums</url>
      <preprocessor type="SocialMediaScrollPreprocessor">
        <timeBetweenFirings>0</timeBetweenFirings>
        <maxFirings>0</maxFirings>
        <conditionBeforeSubsequentFiring>?</conditionBeforeSubsequentFiring>
      </preprocessor>
      <children>
        <regex>&lt;div class="album.*&lt;a href="(.*)"</regex>
        <type>SocialMediaWebsiteSectionMultimediaCollection</type>
        <name>Photo Album</name>
      </children>
    </socialMediaWebsiteSection>
    <socialMediaWebsiteSection type="SocialMediaWebsiteSectionMultimediaCollection">
      <name>Photo Album</name>
      <preprocessor type="SocialMediaScrollPreprocessor">
        <timeBetweenFirings>0</timeBetweenFirings>
        <maxFirings>0</maxFirings>
        <conditionBeforeSubsequentFiring>?</conditionBeforeSubsequentFiring>
      </preprocessor>
      <children>
        <regex>&lt;div class="album.*&lt;a href="(album/[a-zA-Z0-9]+)"</regex>
      </children>
    </socialMediaWebsiteSection>
  </sections>
</socialMediaWebsite>
Evaluation 1: Tool Adaptability

Step 3: Change AFB to preserve synthetic site

- Utilize existing capture mechanisms
- Exploit guaranteed attributes (e.g., host)
- Make code general enough to be widely applicable to sections

```javascript
getCurrentSiteSpec : function(step,urlIn,hostIn){
    switch(step){
        case 0:
            var xhr = new XMLHttpRequest();
            var siteSpec = "", uriOut = "";
            $.ajax({
                url: urlIn,
                success: function(data){
                    var host = "www.facebook.com"; //hostIn n/a here
                    var parser = new DOMParser();
                    var socialMediaWebsites = $(data.childNodes[0]).children;
                    for(var i=0; i<socialMediaWebsites.length; i++){
                        var smw = socialMediaWebsites[i];
                        if($(smw).find("homepage").text().indexOf(host) != -1)
                            siteSpec = $(smw).find("specification").text();
                    }
                    getCurrentSiteSpec(1,siteSpec,host);
                }, error: function(){}, //rof
            }); //xaja
            break;
        case 1:
            $.ajax({
                url: urlIn,
                success: function(data){
                    var ls = window.content.localStorage;
                    ls.setItem("spec", (new XMLSerializer()).serializeToString(archivefbBrowserOverlay.capture(ls.getItem("spec")));
                }, error : function(){}, //rof
            }); //xaja
            break;
    }
}
```
Evaluation 1: Tool Adaptability

Step 4: Change hierarchy of synthetic site

• Simulate simply through mod_rewrite

• Previously:

  RewriteRule ^personal$ index.php?section=personal [NC]

• Updated:

  RewriteRule ^myfeed$ index.php?section=personal [NC]

• Disavow previous reference altogether to ensure 404
Evaluation 1: Tool Adaptability

Step 5: Show AFB breaking

- Run archiving procedure again, note failing of procedure or content not captured

---

Not Found

The requested URL /personal was not found on this server.

Additionally, a 404 Not Found error was encountered while trying to use an ErrorDocument to handle the request.
Evaluation 1: Tool Adaptability
Step 6: Change synthetic site spec

<xml version="1.0" ?>
<socialMediaWebsite>
  <homepage>http://test.socialstandard.org</homepage>
  <sections>
    <socialMediaWebsiteSection type="SocialMediaWebsiteSectionPersonalStream">
      <name>Personal Stream</name>
      <url>http://test.socialstandard.org/personal</url>
      <preprocessor type="SocialMediaScrollPreprocessor">
        <timeBetweenFirings>0</timeBetweenFirings>
        <maxFirings>0</maxFirings>
        <conditionBeforeSubsequentFiring>?</conditionBeforeSubsequentFiring>
      </preprocessor>
    </socialMediaWebsiteSection>
    ...
  </sections>
</socialMediaWebsite>
Evaluation 1: Tool Adaptability
Step 7: Show AFB functionality restored

• Execute archiving procedure of tool w/o modifying code
• Show that result matches step 1
Evaluation 2: Preservation of Content Behind Authentication

1. Create tool (WARCreate) to store to WARC format
2. Setup easy-to-use Replay system (local wayback)
3. Execute Tool’s Archiving Procedure
4. Verify replayability in wayback
Existing Tools’ Shortcoming: Facebook Data Dump

- Lose look & feel
- FB decides what is preserved
- Unreliable (requests not always answered)
Existing Tools’ Shortcoming:
“Save Webpage As”

- Metadata is Lost
- Archive is not Self-Contained
- Archive is not interoperable with Archive Replay Systems (e.g. wayback)
Existing Tools’ Shortcoming: warc-tools

- No archive creation facility
- Relies on incomplete WARC spec (like WARCreate)
- Only command-line access: suitable for sysadmins and power users
Existing Tools’ Shortcoming: wget & wget-warc

• No content manipulation
• Require CLI interaction
  – Issue for Ajax driven content (no JS support)
• wget-warc
  – Ext. of wget w/ WARC I/O
• No look & feel preservation
Existing Tools’ Shortcoming: Archive Facebook

• Output is not compatible w/ Wayback
• Prone to breaking when FB hierarchy changed
• Limited to Firefox web browser
• Cannot escape browser sandbox for portable archives
Existing Tools’ Shortcoming: WARCreate

- No built-in sequential archiving
- Relies on subset of WARC spec
- Limited to Chrome
Shortcoming of Spec

• Relies on accessible URIs of sites’ sections
  – If base page content does not have a URI mapping, no reference exists to direct the browser

• Not comprehensive of Social Media sites

• Likely doesn’t account for some section types
Future Work

• Expand spec website coverage
• Account for sites w/o clearly accessible URIs
• WARCreateto implement whole official WARC standard
• Other SocialMediaWebsitePreprocessor types
• Address perspective issues
  – Personalization/Auth, context, archive vs. backup
Contributions

1. Highlight Personal Web Archiving difficulties
   – ways they can be addressed

2. Provide remote spec for PWA tools to use to be more robust to sites’ hierarchy changes

3. Create tool (WARCreate)
   – allows content behind auth to be preserved to standard format

4. Leverage client-side server to exec scripts in support of personal web preservation

5. Establish section commonality between social media websites
Conclusions

- Personal web archiving has unique problems not exhibited in conventional web archiving
- Tools become more adaptive by utilizing proposed spec
- Browsers can be used as medium for preservation of personal web content
- With little work, server technologies can help to ease the task of personal web archiving
WARCreate-Related Presentations

Mat Kelly (Old Dominion University, Norfolk, VA), Michele C. Weigle (Old Dominion University, Norfolk, VA), Michael Nelson (Old Dominion University, Norfolk, VA). "WARCreate - Create Wayback-Consumable WARC Files from Any Webpage," Digital Preservation 2012, Tools Demo Session: Web Archiving; 2012 Jul 25; Washington, DC.

Mat Kelly (Old Dominion University, Norfolk, VA) and Michele C. Weigle (Old Dominion University, Norfolk, VA), "WARCreate - Create Wayback-Consumable WARC Files from Any Webpage (demo)," In Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL). Washington, DC, June 2012

For more information on:

WARCreate: http://warcreate.com
Example: Implicit Recursion

SocialMediaWebsite
- homepage : str
- sections : SocialMediaWebsiteSection []

SocialMediaWebsiteSection
- name : str
- url : str
- [preprocessor : SocialMediaWebsitePreprocessor]

SocialMediaWebsiteStream

SocialMediaWebsitePersonalStream

SocialMediaWebsitePeerStream

SocialMediaWebsiteMultimedia

SocialMediaWebsiteMultimediaCollection
- children : SocialMediaWebsiteMultimedia []

SocialMediaWebsiteMultimediaItem

SocialMediaWebsitePreprocessor
- timeBetweenFirings : int
- maxFirings : int
- condition : SocialMediaWebsitePreprocessorCondition

SocialMediaWebsitePreprocessorCondition