



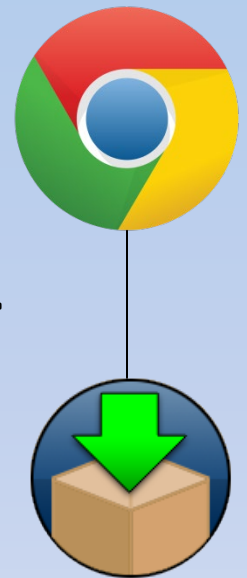
WARCreate

Create Wayback-Consumable WARC Files from Any Webpage

Mat Kelly, Michele C. Weigle, Michael L. Nelson
{mkelly,mweigle,mIn}@cs.odu.edu
Old Dominion University; Norfolk, VA

What is WARCreate?

- Google Chrome extension
- Creates WARC files
- Enables preservation by users from their browser
- First steps in bringing Institutional Archiving facilities to the PC



Target Content



- Unreachable by web crawlers
 - Behind authentication
 - Not listed in search engines (Deep Web)



- Private

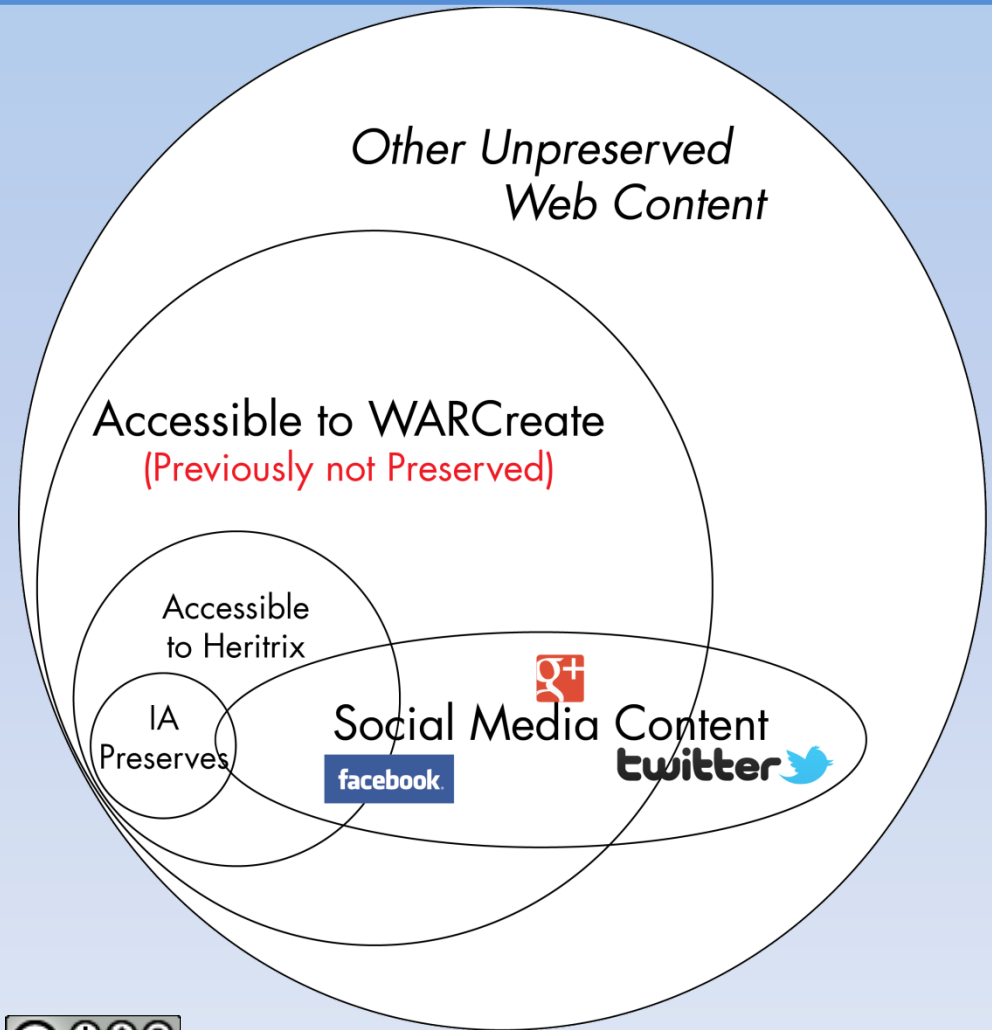


- We don't want our bank statements in Wayback
- Non-pertinent to public
 - Others have little interest in our Facebook comments



Preserving More!

- Much digital information is needlessly lost
- User chooses what they deem important
- Compatible with standard archiving tools.



WYSIWYG

Facebook-Supplied Data Dump



John Conner

John Conner Does anybody know how I get the Facebook Timeline? I do not have the big green button at the bottom of the Timeline page. :(
February 28, 2012 at 5:47 pm

John Conner Hump Day means we're half way through the week. Can't wait until the weekend!
February 22, 2012 at 1:16 pm

Aaron Januszewski John conner, didn't i see you in terminator?
February 10, 2012 at 5:06 pm

John Conner That was John Connor (note the surname) but thanks for noticing. :)
February 15, 2012 at 1:23 pm

John Conner One day closer to the weekend! Wooohoo!
February 9, 2012 at 7:16 pm

John Conner
February 9, 2012 at 2:56 pm

Downloaded by John Conner (<http://www.facebook.com/profile.php?id=100003509861423>) on March 20, 2012 at 1:32 pm

Archive created from WARCreate in Wayback



localhost:8080/wayback/2011: x

localhost:8080/wayback/20110724015446/https://www.facebook.com

Wayback Machine

1 captures
24 JUL 12 - 24 JUL 12

John Conner

What's on your mind?

Public Post

APPS

- App Center
- Photos
- Music
- Notes
- Links
- Pokes

GROUPS

- Create Group...

FRIENDS

- Close Friends
- Family
- Textile Manufacturers Limited
- Springfield High School 20+

INTERESTS

- Subscriptions
- Add Interests...

Aimee Stahler

Thank you all so much for your prayers, kindness, thoughts, well wishes, comments, gifts, calls/emails/texts, meals, flowers, visits, and love! We feel truly blessed to have so much love and support from all of you. We wish that we could thank each and every one of you personally, but we're still enjoying our time at home as a family and adjusting to having the little man around :) We just can't get enough of our sweet Grant! Love you all and praise the Lord for this amazing miracle :)

Like · Comment · July 20 at 11:29am

33 people like this.

View all 8 comments

- Kara Somers Arbutina Congratulations!
July 20 at 3:52pm · Like
- LeighAnna Yasiejko Congratulations Aimee!!! He's adorable!!! :)
July 20 at 8:08pm · Like

Write a comment...

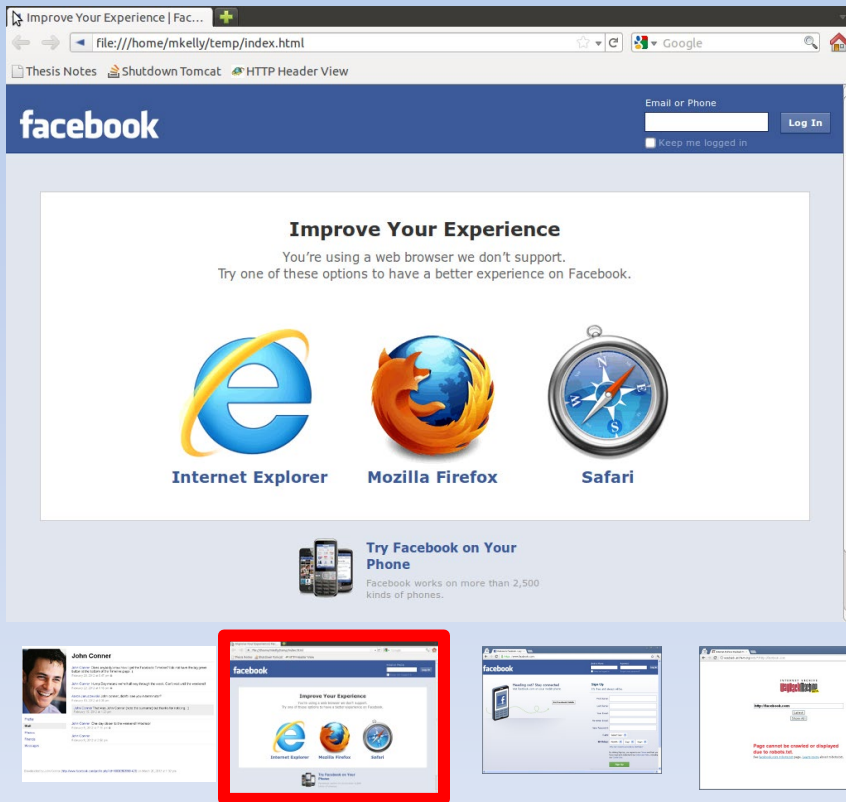
Tiffany Stinson shared ShopAtHome.com's photo.



WYSIWYG

Using Scraping Tools (e.g. wget)

Archive created from
WARCreate in Wayback



WYSIWYG

A Crawler Has No Context

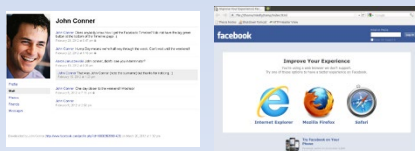
Archive created from
WARCreate in Wayback



WYSIWYG

IA/HERITRIX OBEY ROBOTS


Archive created from
WARCreate in Wayback



Goals

- Make it easy to use (GUI-based, no cmd line)
- Make it useful (fill the need)
- Demonstrate novelty of browser-instigated preservation
- Show value of WARC format for Personal Web preservation
- Bring WARC format to Personal Digital Archiving




You've gone incognito. Pages you view in this window won't appear in your browser history or search history, and they won't leave other traces, like cookies, on your computer after you close **all** open incognito windows. Any files you download or bookmarks you create will be preserved, however. 

Going incognito doesn't affect the behavior of other people, servers, or software. Be wary of:

- Websites that collect or share information about you
- Internet service providers or employers that track the pages you visit
- Malicious software that tracks your keystrokes in exchange for free smileys
- Surveillance by secret agents
- People standing behind you

[Learn more](#) about incognito browsing.

 Because Google Chrome does not control how extensions handle your personal data, all extensions have been disabled for incognito windows. You can reenable them individually in the [extensions manager](#).

Creating a WARC

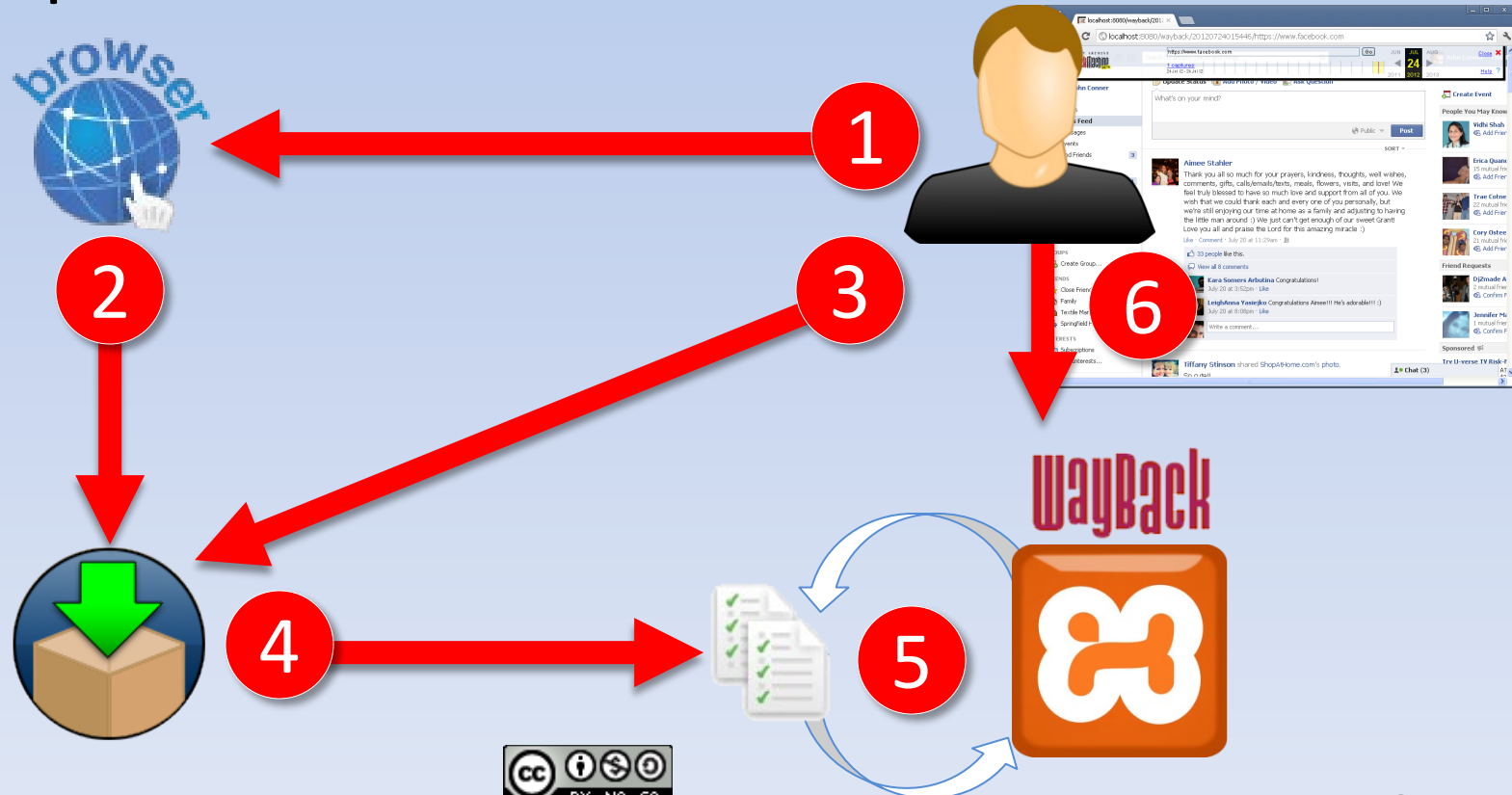
I've Made a WARC. Now what?

- What you do with the archive is up to you.
 - Install it in your local Wayback instance
- Who has their own Wayback Instance!?
 - Wayback is free & open source
- That seems like a lot of work!
 - One additional reason for users **NOT** to preserve what they would like archived



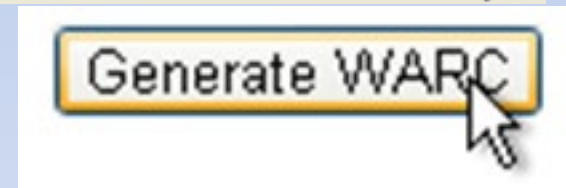
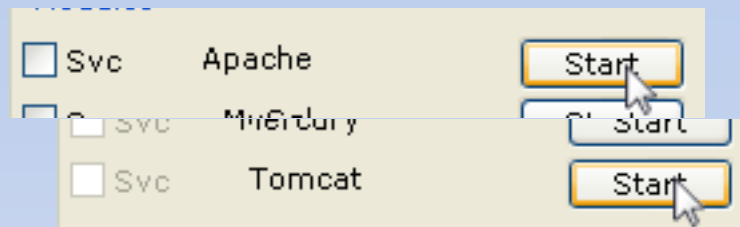
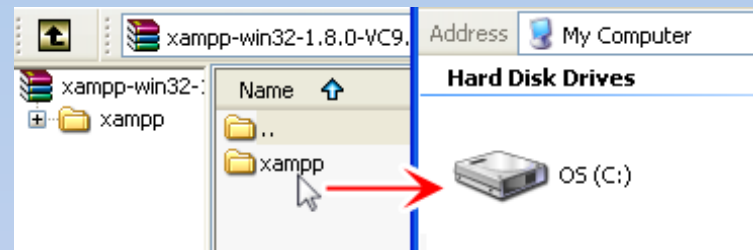
WARC Creation & Replay

1. WARC Originator (User) browses the Web Page
to determine WARC creation to local wayback



Suite Installation & Interaction

- Drag & Drop .zip to hd
- Start relevant services using GUI
- Execute WARCCreate process
- View Archive at <http://localhost/wayback>



You've gone incognito. Pages you view in this window won't appear in your browser history or search history, and they won't leave other traces, like cookies, on your computer after you close **all** open incognito windows. Any files you download or bookmarks you create will be preserved, however.



Going incognito doesn't affect the behavior of other people, servers, or software. Be wary of:

- Websites that collect or share information about you
- Internet service providers or employers that track the pages you visit
- Malicious software that tracks your keystrokes in exchange for free smileys
- Surveillance by secret agents
- People standing behind you

[Learn more](#) about incognito browsing.



Because Google Chrome does not control how extensions handle your personal data, all extensions have been disabled for incognito windows. You can reenable them individually in the [extensions manager](#).

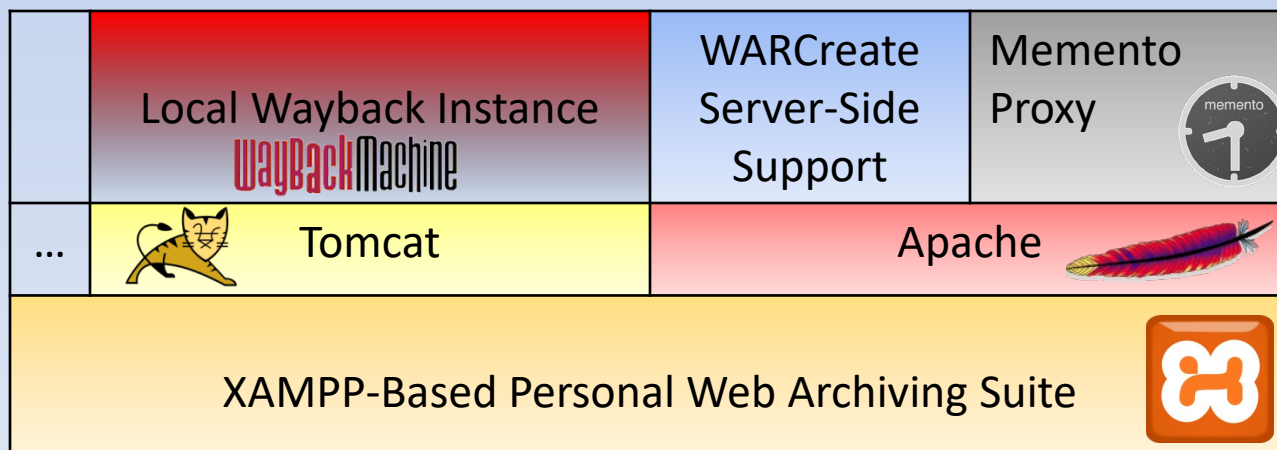
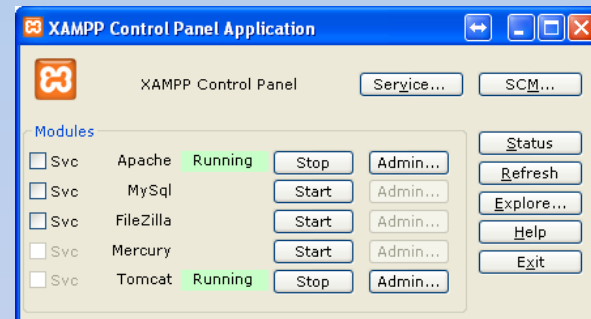
Replay of Preserved Twitter page

And My Bank Statements?

- Preserved content:
 - never leaves WARC files
 - never leaves local machine
- WARCcreate provides preliminary encoding/encryption support
- Wayback instance is hosted on your own machine – no external access by default

Why Use a Client-Side Server?

- Server scripts do what JS can't
- Can reside on your machine!
- Controls are GUI based
- Resource fetching w/o XSS issues

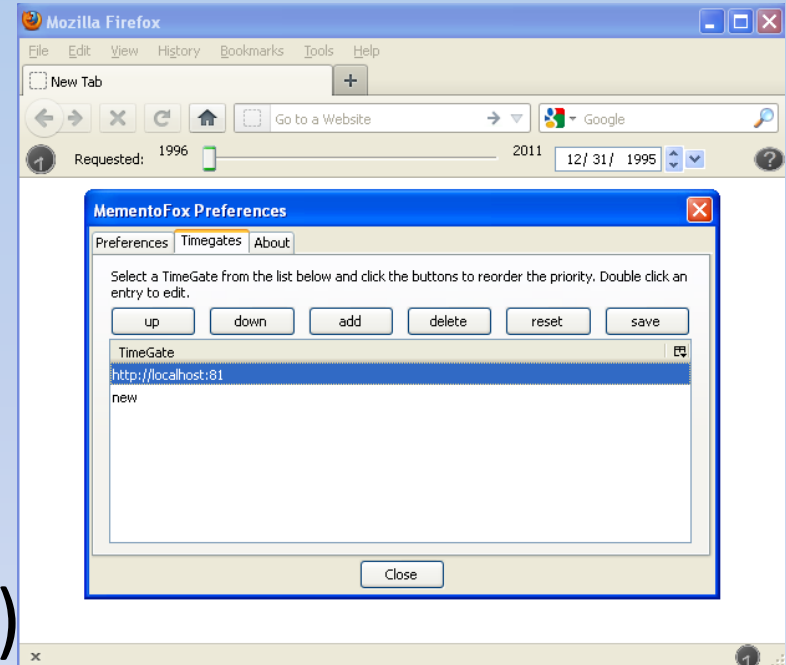


Built On
↓

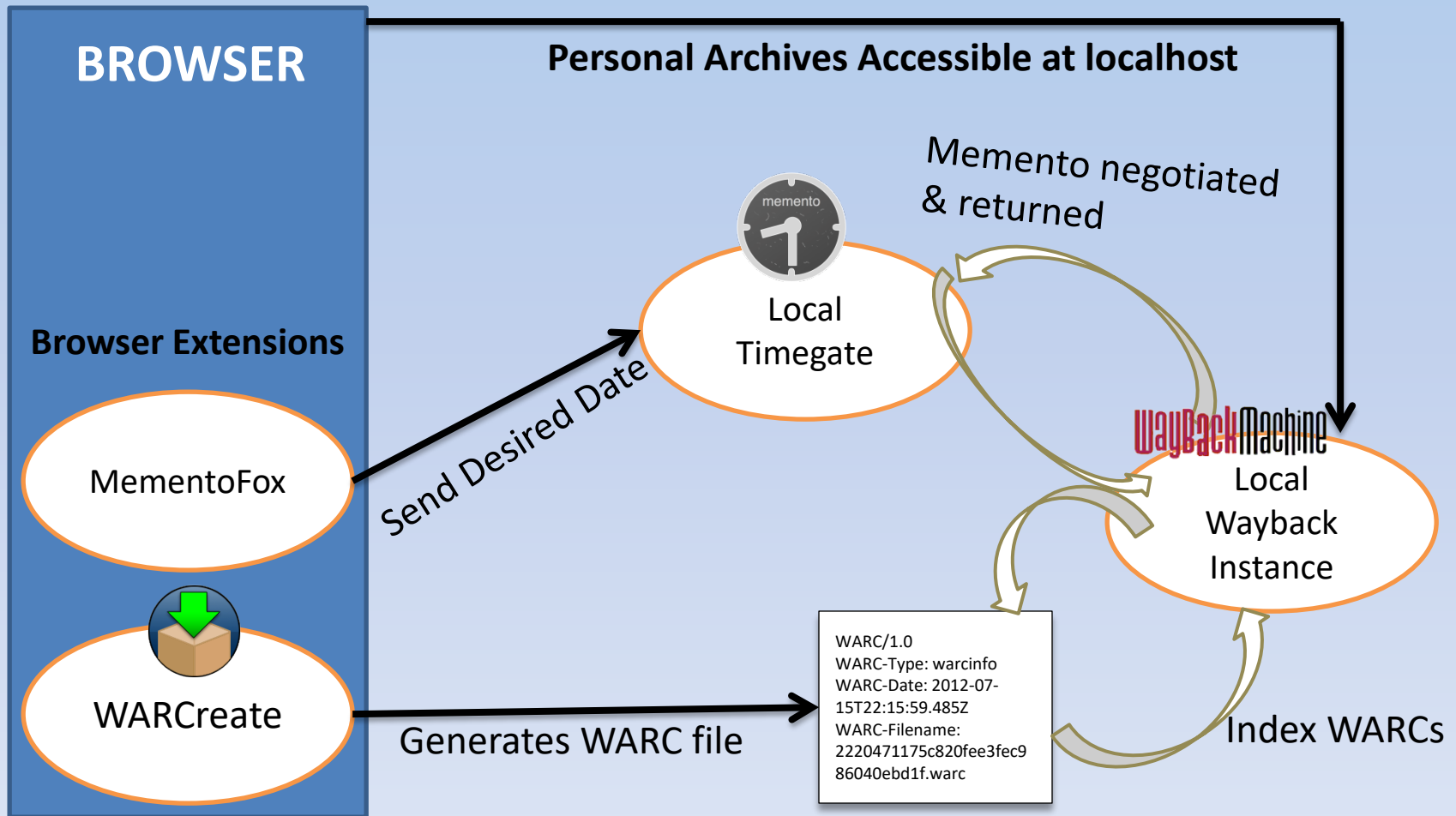


Extras: Memento Support

- Suite's includes tailored Timegate
- Memento abstraction is beyond WARC
- Point MementoFox (or other Memento tools) to localhost



How it All Relates



Contribution of Work

- Facilitate browser-based Personal Web Archiving
- Determine feasibility of fully Client-Side Preservation
- Integrate with existing tools for establishing use cases





WARCreate

Create Wayback-Consumable WARC Files from Any Webpage

<http://WARCreate.com>



Backup Slides



Future Work

- Decouple from “server”
- Refine Memento integration
- Reference full WARC spec
- Built-in WARC validation
- Built-in replay
- Compression
- Optimization (removing duplicates)
- ...many more

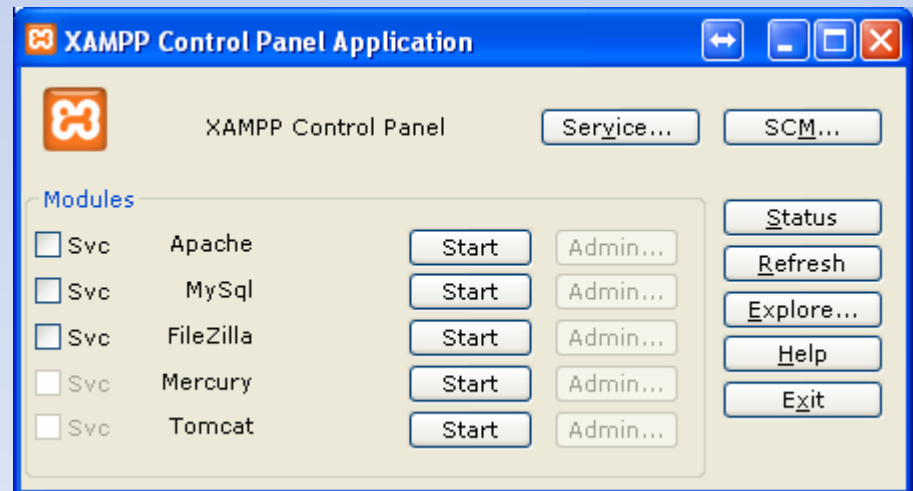


Extras: Configuration Sanity Check

- Server scripts make up for Javascript shortcomings
- The server can reside on your machine!
- Setup, Start, Stop are GUI based

- ✗ WARC Validation
- ✗ AJAX XSS Circumvention
- ✗ HTML5 Sandbox Escaping
- ✗ Memento Support
- ✗ Local Wayback Instance

In WARCcreate

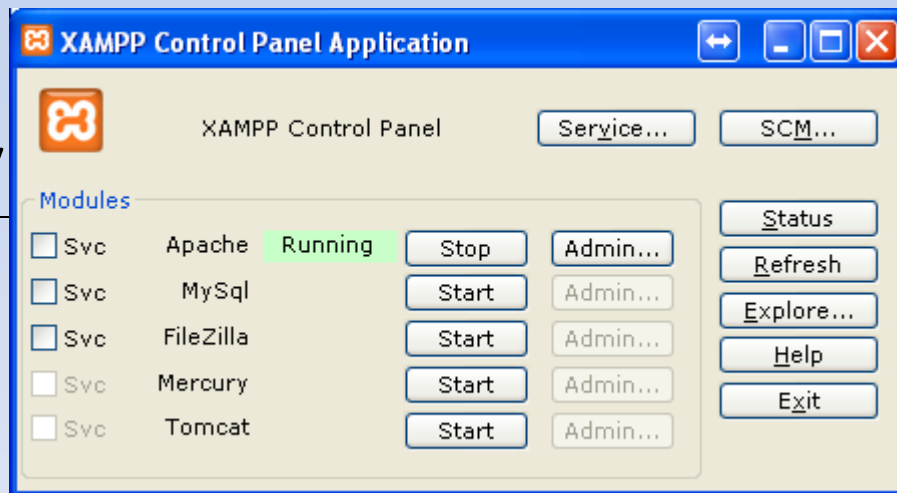


Extras: Configuration Sanity Check

- + Apache allows generated WARC's to be validated
 - + Javascript cannot write to disk, server-side scripts can
 - + Server prevents hot-linking & has security
-
- = Content better preserved using server techs

- ✓ WARC Validation
- ✓ AJAX XSS Circumvention
- ✓ HTML5 Sandbox Escaping
- ? Memento Support
- ✗ Local Wayback Instance

In WARCcreate



Extras: Configuration Sanity Check

- **Memento** requires **Wayback**
Wayback requires **Tomcat**
∴ **Memento** requires **Tomcat**
- **Memento** Timegate req's
Python+modules
(*pre-packaged + included*)

- ✓ WARC Validation
- ✓ AJAX XSS Circumvention
- ✓ HTML5 Sandbox Escaping
- ✓ Memento Support
- ✓ Local Wayback Instance

In WARCreate

