Archive What I See Now

Personal Web Archiving with WARCs

Michele C. Weigle, Michael L. Nelson, **Mat Kelly**, and John Berlin
Web Science and Digital Libraries (WS-DL) Research Group
Old Dominion University
ws-dl.cs.odu.edu • @WebSciDL

@machawk1

IIPC Web Archiving Conference 2017
June 15, 2017
London, UK

http://bit.ly/iipcWAC2017

NATIONAL ENDOWMENT FOR THE Humanities
HD-51670-13 • HK-50181-14

Web Archiving Tools for Web Users

Standard Web archiving tools are difficult for non IT experts.

“Save Page As” is not suitable for archiving purposes.

Pages are behind authentication.

Pages change quickly, but current state needs archiving.

ARCHIVE

WHAT I SEE

NOW

http://bit.ly/iipcWAC2017

NATIONAL ENDOWMENT FOR THE Humanities
HD-51670-13 • HK-50181-14

Why?

- Allow non-technical users to locally create+replay own archives
- Preserve the previously unpreserved

more archives → more better

CREATION

+

ACCESS

of personal and private web archives

IIPC Web Archiving Conference 2017
June 15, 2017
London, UK

http://bit.ly/iipcWAC2017

@machawk1

NATIONAL ENDOWMENT FOR THE
Humanities

Goals: Advance Development of 3 Tools

WARCreate

Create a WARC from what you see in your browser

Web Archiving Integration Layer (WAIL)

Replay the WARC using software of your desktop

Your captures never leaves your machine

Mink

See how your captures temporally integrate with institutions'

Submit new URIs to Web archives (was to-WAIL in scope?)

IIPC Web Archiving Conference 2017
June 15, 2017
London, UK

http://bit.ly/iipcWAC2017

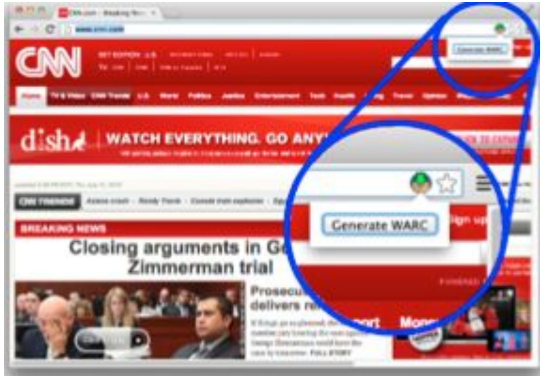
@machawk1

NATIONAL ENDOWMENT FOR THE Humanities

WARCreate

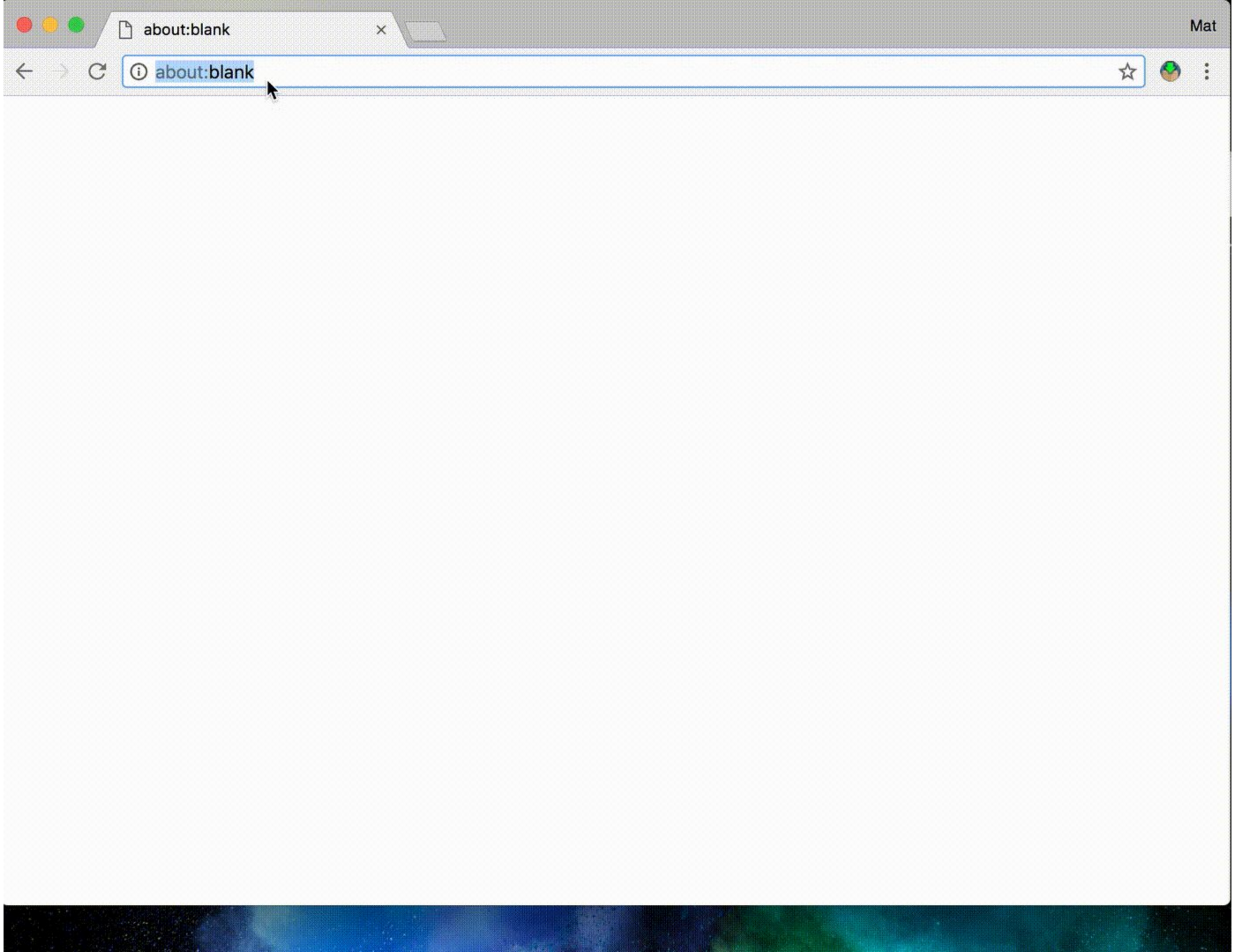
WARCreate

- Google Chrome browser extension
- Save WARC files from your browser
- No credentials pass through 3rd party
- Heavily leverages Chrome webRequest API
- Built in ‘12, APIs and libraries have evolved!

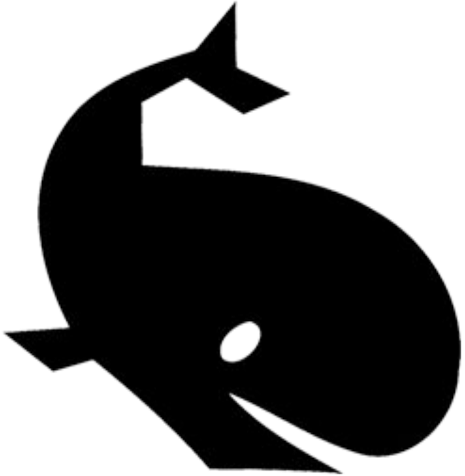


WARCreate - Recent Advancements

- Three New Modes for Browser-Based Preservation
  - Record Mode - retain buffer as you browse
  - Countdown Mode - preserve reloading page on an interval
  - Event Mode - preserve page when it's automatically reloaded
- Save to local Web archive (e.g., WAIL)



about:blank
Mat
about:blank



Web Archiving Integration Layer (WAIL)

Web Archiving Integration Layer (WAIL)

- Stand-alone desktop application
- Collection-based Web Archiving
- Includes Heritrix for crawling, OpenWayback for Replay
- Python scripts compiled to OS-native binaries (.app, .exe)

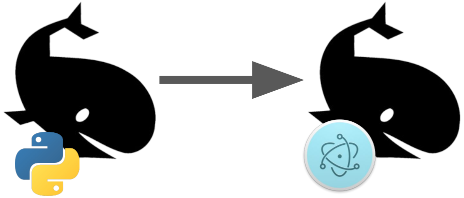
- What to do with WARCs?

- See: [How WAIL came about, "Lipstick or Ham"]()

NATIONAL ENDOWMENT FOR THE
Humanities
HD-51670-13 • HK-50181-14

WAIL - Recent Advancements

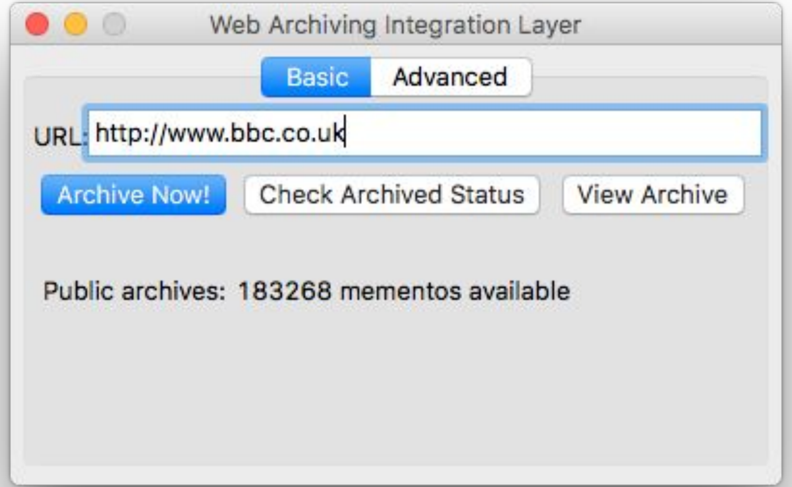
- New User Interface
- Ported from Python to Electron
  - Now using Web technologies to archive the Web
- Single archive to collection-based archiving
- OpenWayback to pywb
- Twitter integration



WAIL-Electron Feature Walk-through

NATIONAL ENDOWMENT FOR THE Humanities
HD-51670-13 • HK-50181-14

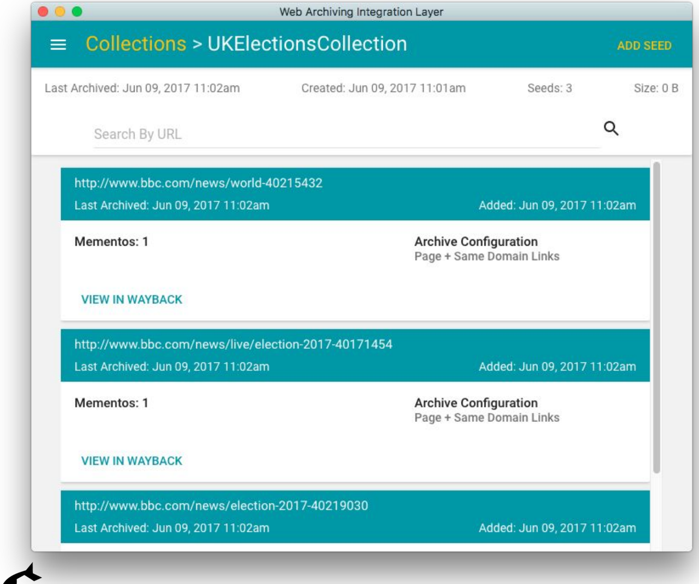
WAIL - New User Interface



URL: http://www.bbc.co.uk

Public archives: 183268 mementos available

Original one-click interface



Web Archiving Integration Layer
Collections > UKElectionsCollection
ADD SEED
Search By URL
VIEW IN WAYBACK
VIEW IN WAYBACK

Last Archived: Jun 09, 2017 11:02am Created: Jun 09, 2017 11:01am Seeds: 3 Size: 0 B

http://www.bbc.com/news/world-40215432
Last Archived: Jun 09, 2017 11:02am Added: Jun 09, 2017 11:02am
Mementos: 1
Archive Configuration
Page + Same Domain Links

http://www.bbc.com/news/live/election-2017-40171454
Last Archived: Jun 09, 2017 11:02am Added: Jun 09, 2017 11:02am
Mementos: 1
Archive Configuration
Page + Same Domain Links

http://www.bbc.com/news/election-2017-40219030
Last Archived: Jun 09, 2017 11:02am Added: Jun 09, 2017 11:02am

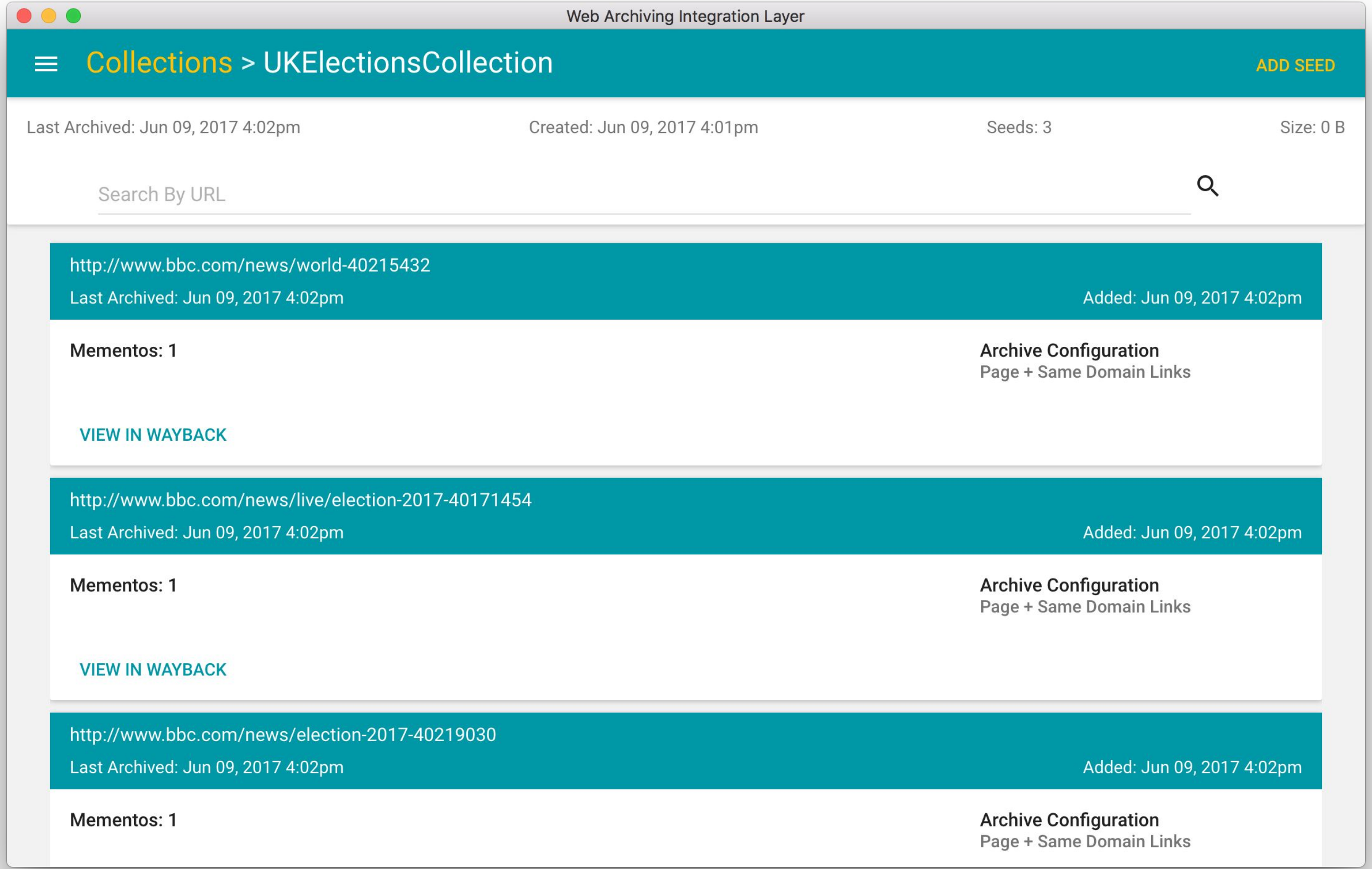
New collection-based interface

IIPC Web Archiving Conference 2017
June 15, 2017
London, UK

http://bit.ly/iipcWAC2017

@machawk1

NATIONAL ENDOWMENT FOR THE Humanities
HD-51670-13 • HK-50181-14



Web Archiving Integration Layer

Collections > UKElectionsCollection

Last Archived: Jun 09, 2017 4:02pm
Created: Jun 09, 2017 4:01pm
Seeds: 3
Size: 0 B

http://www.bbc.com/news/world-40215432
Last Archived: Jun 09, 2017 4:02pm
Added: Jun 09, 2017 4:02pm

Mementos: 1

Archive Configuration
Page + Same Domain Links

http://www.bbc.com/news/live/election-2017-40171454
Last Archived: Jun 09, 2017 4:02pm
Added: Jun 09, 2017 4:02pm

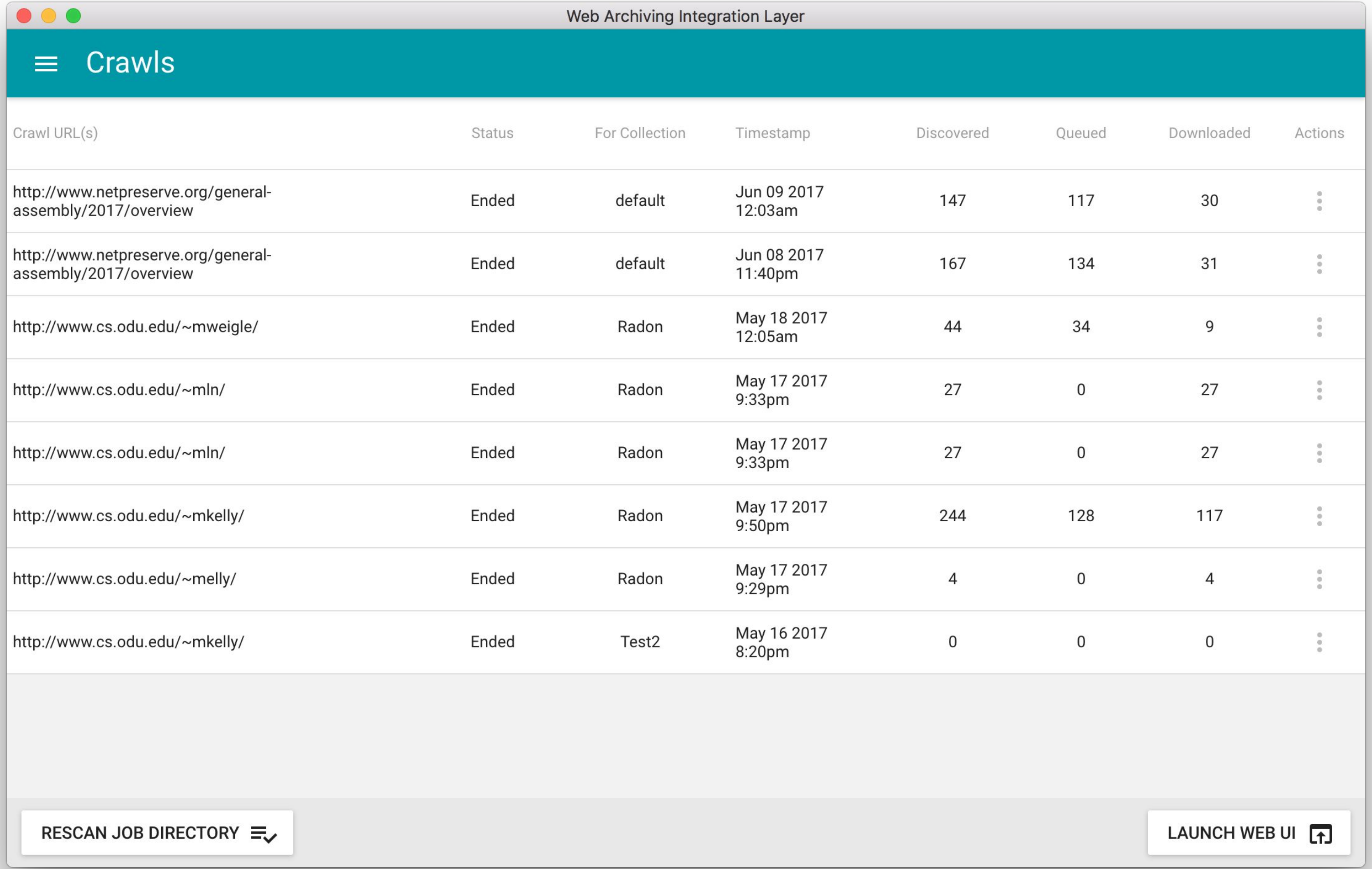
Mementos: 1

Archive Configuration
Page + Same Domain Links

http://www.bbc.com/news/election-2017-40219030
Last Archived: Jun 09, 2017 4:02pm
Added: Jun 09, 2017 4:02pm

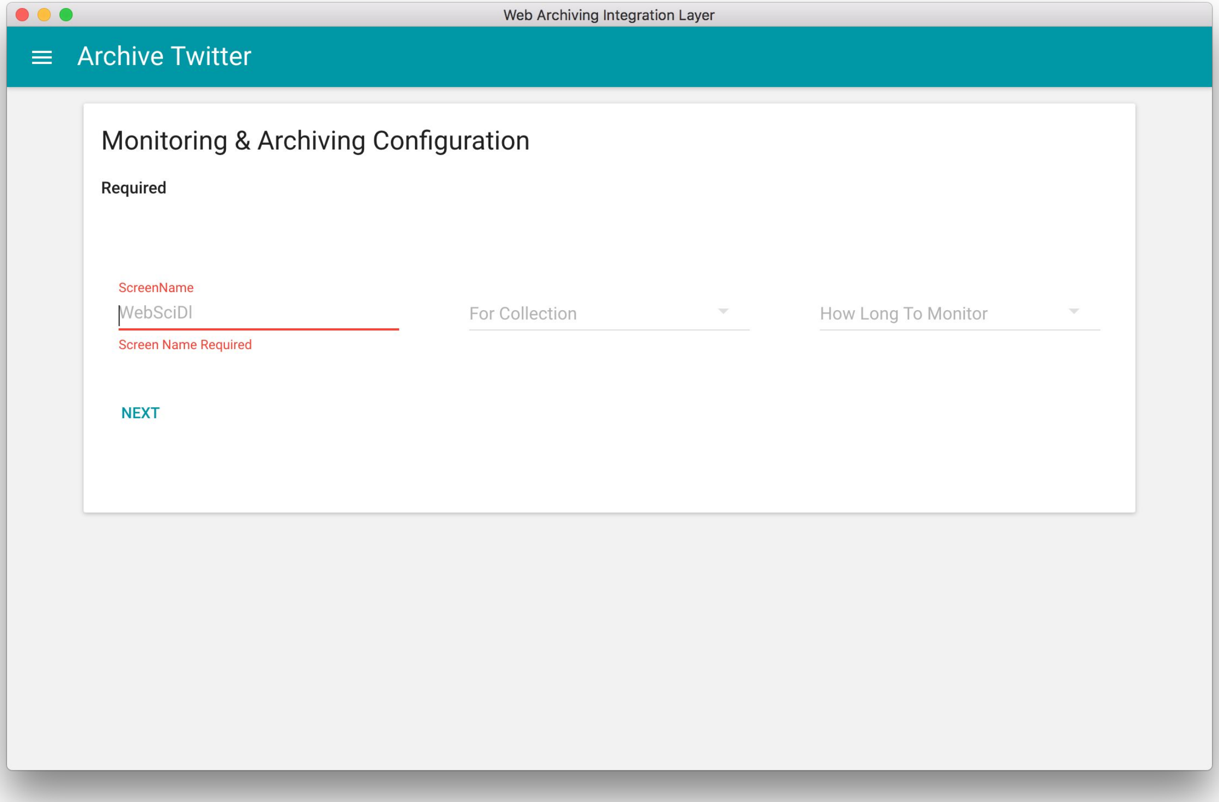
Mementos: 1

Archive Configuration
Page + Same Domain Links



Crawls

Crawl URL(s)	Status	For Collection	Timestamp	Discovered	Queued	Downloaded	Actions
http://www.netpreserve.org/general-assembly/2017/overview	Ended	default	Jun 09 2017 12:03am	147	117	30	⋮
http://www.netpreserve.org/general-assembly/2017/overview	Ended	default	Jun 08 2017 11:40pm	167	134	31	⋮
http://www.cs.odu.edu/~mweigle/	Ended	Radon	May 18 2017 12:05am	44	34	9	⋮
http://www.cs.odu.edu/~mln/	Ended	Radon	May 17 2017 9:33pm	27	0	27	⋮
http://www.cs.odu.edu/~mln/	Ended	Radon	May 17 2017 9:33pm	27	0	27	⋮
http://www.cs.odu.edu/~mkelly/	Ended	Radon	May 17 2017 9:50pm	244	128	117	⋮
http://www.cs.odu.edu/~melly/	Ended	Radon	May 17 2017 9:29pm	4	0	4	⋮
http://www.cs.odu.edu/~mkelly/	Ended	Test2	May 16 2017 8:20pm	0	0	0	⋮



Web Archiving Integration Layer

Archive Twitter

Monitoring & Archiving Configuration

Required

ScreenName: WebSciDl

Screen Name Required

For Collection

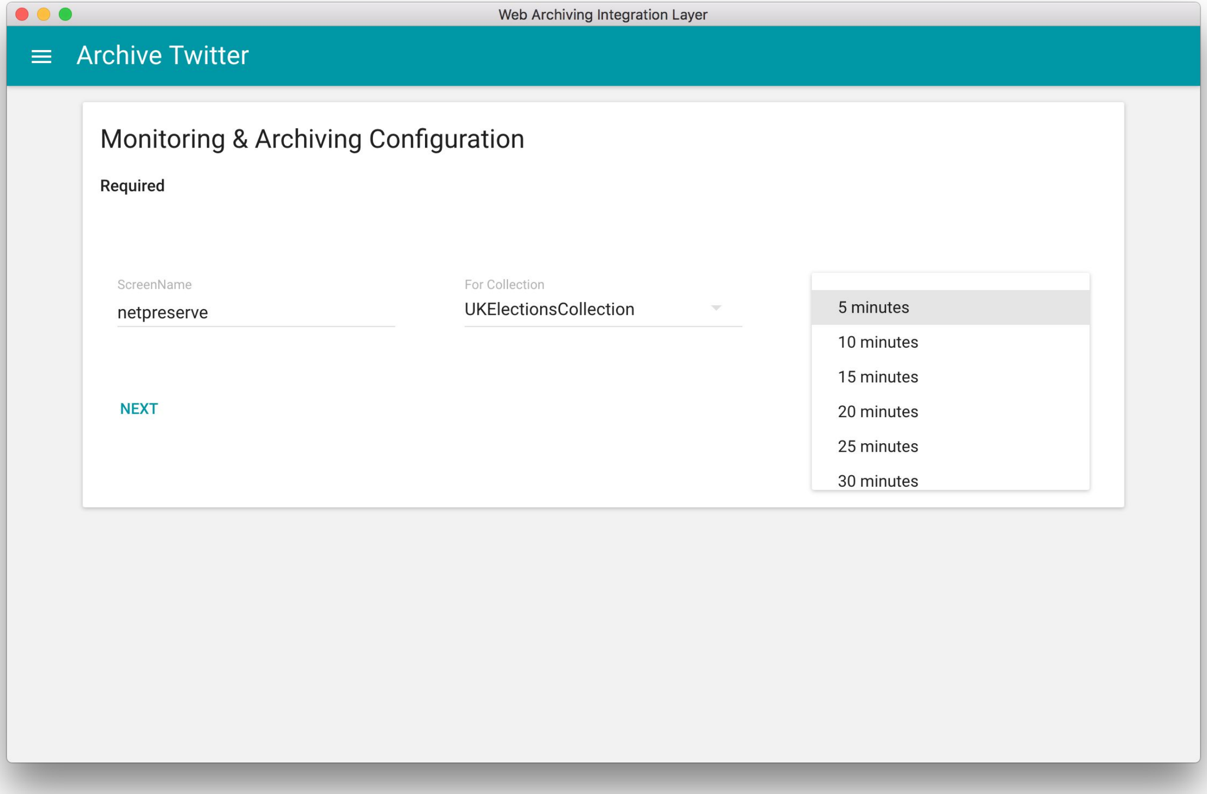
How Long To Monitor

IIPC Web Archiving Conference 2017
June 15, 2017
London, UK

http://bit.ly/iipcWAC2017

@machawk1

NATIONAL ENDOWMENT FOR THE Humanities
HD-51670-13 • HK-50181-14



Web Archiving Integration Layer

Archive Twitter

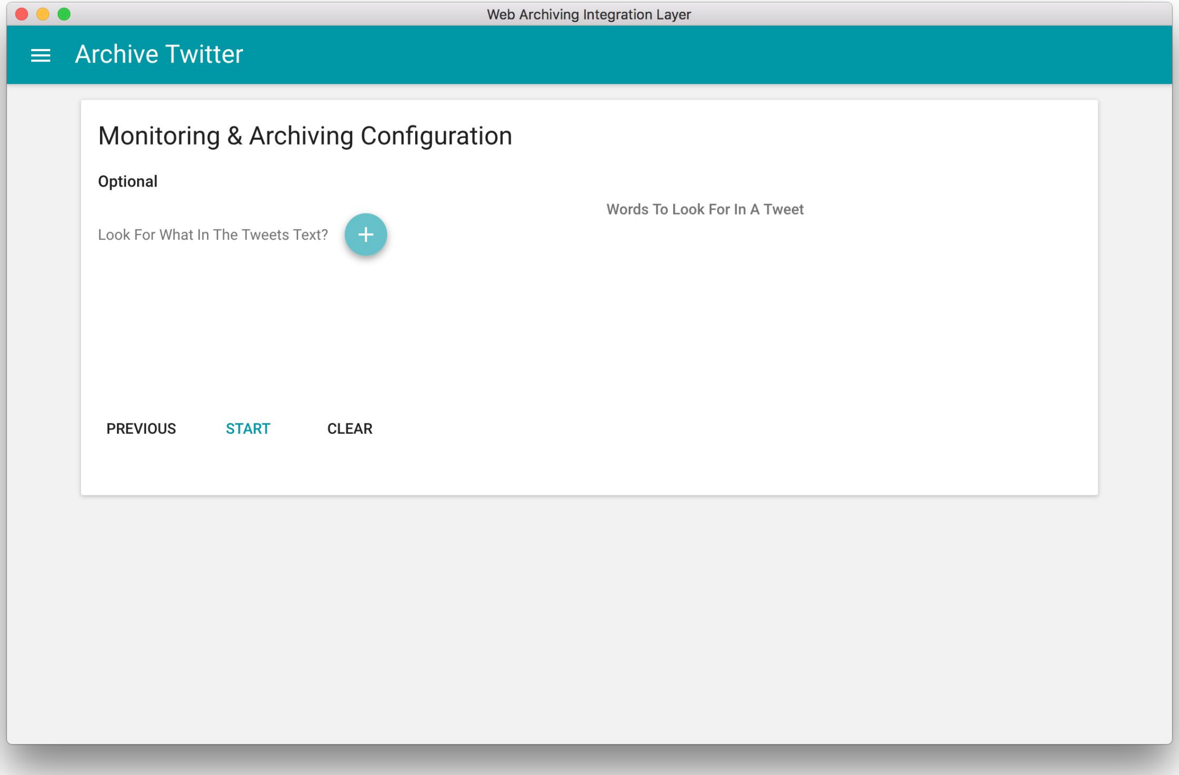
Monitoring & Archiving Configuration

Required

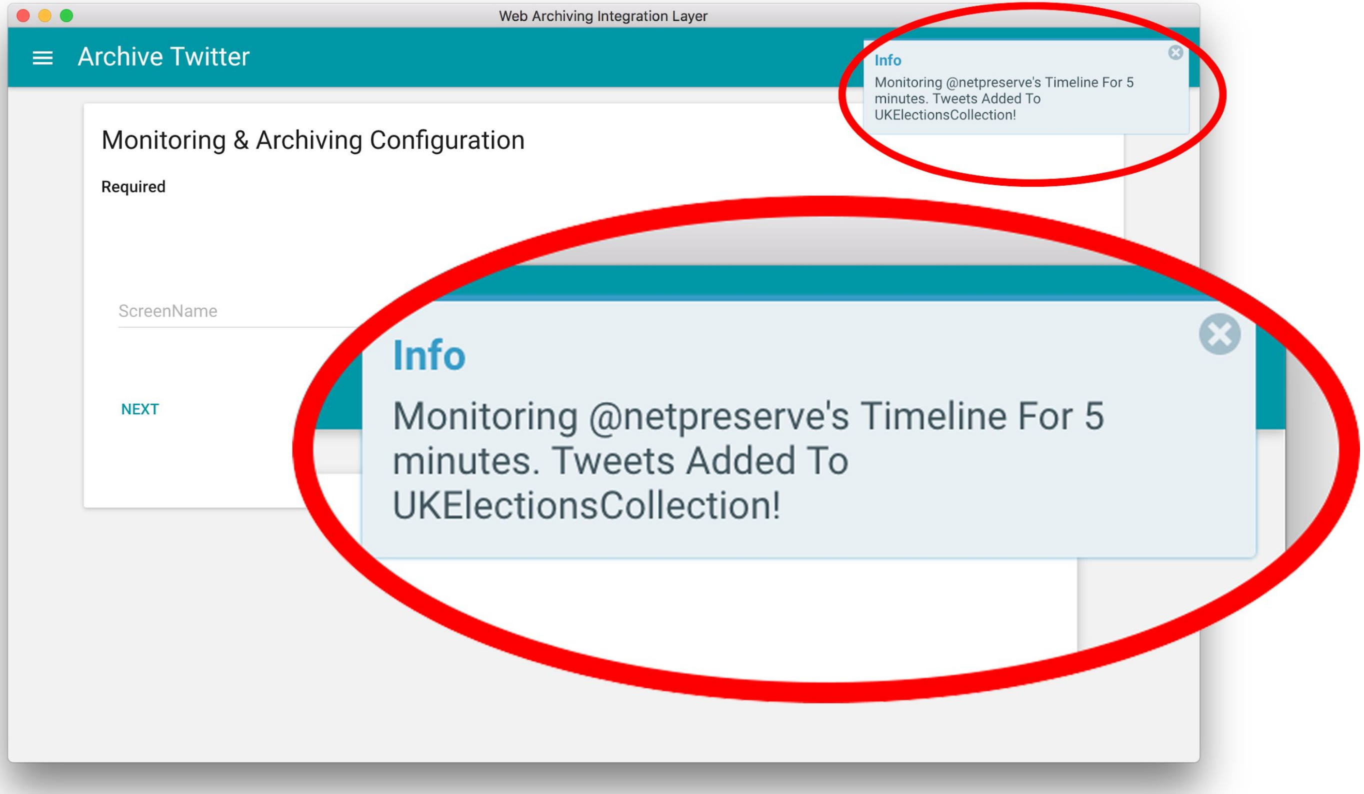
ScreenName: netpreserve

For Collection: UKElectionsCollection

NATIONAL ENDOWMENT FOR THE Humanities
HD-51670-13 • HK-50181-14


Web Archiving Integration Layer
Archive Twitter
Monitoring & Archiving Configuration
Optional
Look For What In The Tweets Text?
+
Words To Look For In A Tweet
PREVIOUS
START
CLEAR

NATIONAL ENDOWMENT FOR THE Humanities
HD-51670-13 • HK-50181-14



Monitoring & Archiving Configuration

Info

Monitoring @netpreserve's Timeline For 5 minutes. Tweets Added To UKElectionsCollection!

Info

Monitoring @netpreserve's Timeline For 5 minutes. Tweets Added To UKElectionsCollection!

memento
MINK
INTEGRATING the WEBS

Mink

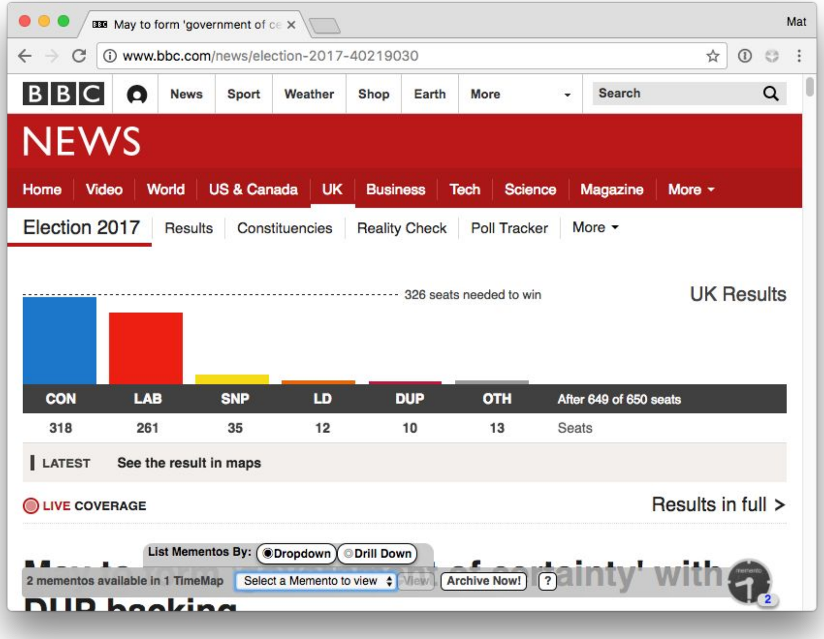
- Google Chrome browser extension
- Indicates archival capture count as you browse
- Quickly submit URI to multiple archives from UI
- From Mink(owski Space)

NATIONAL ENDOWMENT FOR THE Humanities
HD-51670-13 • HK-50181-14

Mink - Recent Advancements

- Enhance interface
  - Add number of archived pages to icon at bottom of page
  - Allow users to set preferences on how to view large set of mementos
- Communication with user-specified (or local) archive in additional to aggregated institutional archives' results

Mink - Previous Interface



CON	LAB	SNP	LD	DUP	OTH	After 649 of 650 seats
318	261	35	12	10	13	Seats

- ➢ Interface affected by page CSS
- ➢ Obtrusive on the viewport by default
- ➢ Haphazard, inconsistent animations

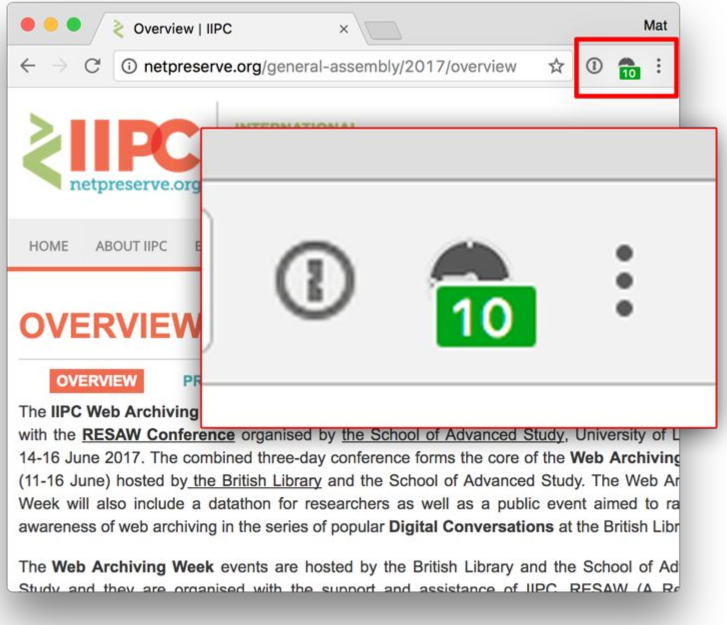
IIPC Web Archiving Conference 2017
June 15, 2017
London, UK

http://bit.ly/iipcWAC2017

@machawk1

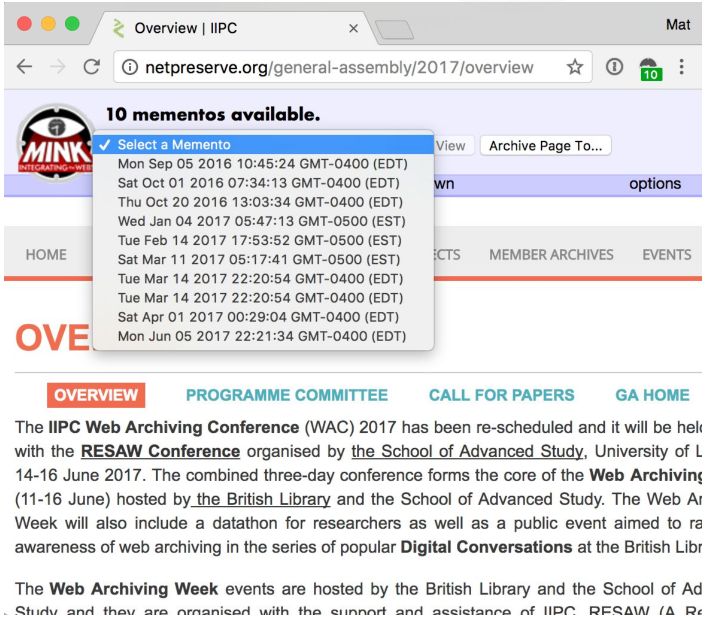
NATIONAL ENDOWMENT FOR THE
Humanities
HD-51670-13 • HK-50181-14

Mink - User Interface Revamp



The IIPC Web Archiving with the RESAW Conference organised by the School of Advanced Study, University of L 14-16 June 2017. The combined three-day conference forms the core of the Web Archiving (11-16 June) hosted by the British Library and the School of Advanced Study. The Web Ar Week will also include a datathon for researchers as well as a public event aimed to ra awareness of web archiving in the series of popular **Digital Conversations** at the British Libr

The **Web Archiving Week** events are hosted by the British Library and the School of Ad Study and they are organised with the support and assistance of IIPC, RESAW (A Re



10 mementos available.

- ✓ Select a Memento
- Mon Sep 05 2016 10:45:24 GMT-0400 (EDT)
- Sat Oct 01 2016 07:34:13 GMT-0400 (EDT)
- Thu Oct 20 2016 13:03:34 GMT-0400 (EDT)
- Wed Jan 04 2017 05:47:13 GMT-0500 (EST)
- Tue Feb 14 2017 17:53:52 GMT-0500 (EST)
- Sat Mar 11 2017 05:17:41 GMT-0500 (EST)
- Tue Mar 14 2017 22:20:54 GMT-0400 (EDT)
- Tue Mar 14 2017 22:20:54 GMT-0400 (EDT)
- Sat Apr 01 2017 00:29:04 GMT-0400 (EDT)
- Mon Jun 05 2017 22:21:34 GMT-0400 (EDT)

The **IIPC Web Archiving Conference** (WAC) 2017 has been re-scheduled and it will be hel with the **RESAW Conference** organised by the School of Advanced Study, University of L 14-16 June 2017. The combined three-day conference forms the core of the **Web Archiving** (11-16 June) hosted by the British Library and the School of Advanced Study. The Web Ar Week will also include a datathon for researchers as well as a public event aimed to ra awareness of web archiving in the series of popular **Digital Conversations** at the British Libr

The **Web Archiving Week** events are hosted by the British Library and the School of Ad Study and they are organised with the support and assistance of IIPC, RESAW (A Re

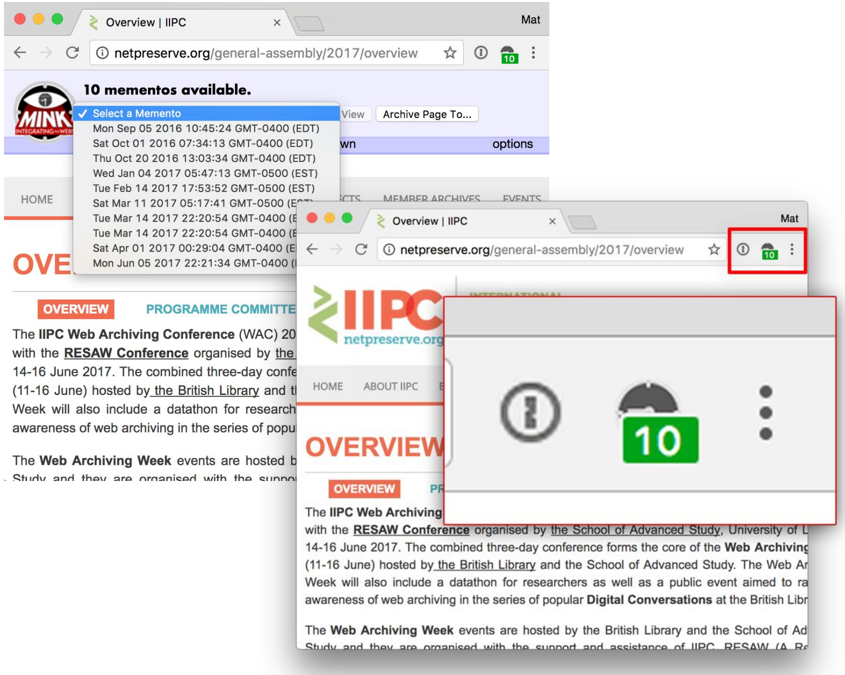
IIPC Web Archiving Conference 2017
June 15, 2017
London, UK

http://bit.ly/iipcWAC2017

@machawk1

NATIONAL ENDOWMENT FOR THE Humanities
HD-51670-13 • HK-50181-14

Mink - User Interface Revamp


Overview | IIPC
Mat
netpreserve.org/general-assembly/2017/overview
10
10 mementos available.
✓ Select a Memento
Mon Sep 05 2016 10:45:24 GMT-0400 (EDT)
Sat Oct 01 2016 07:34:13 GMT-0400 (EDT)
Thu Oct 20 2016 13:03:34 GMT-0400 (EDT)
Wed Jan 04 2017 05:47:13 GMT-0500 (EST)
Tue Feb 14 2017 17:53:52 GMT-0500 (EST)
Sat Mar 11 2017 05:17:41 GMT-0500 (E
Tue Mar 14 2017 22:20:54 GMT-0400 (E
Tue Mar 14 2017 22:20:54 GMT-0400 (E
Sat Apr 01 2017 00:29:04 GMT-0400 (E
Mon Jun 05 2017 22:21:34 GMT-0400 (
View
Archive Page To...
wn
options
HOME
ABOUT IIPC
CTS
MEMBER ARCHIVES
EVENTS
IIPC
netpreserve.org
OVERVIEW
OVERVIEW
PROGRAMME COMMITTE
10

The IIPC Web Archiving Conference (WAC) 20… with the RESAW Conference organised by the School of Advanced Study, University of L… 14-16 June 2017. The combined three-day conference forms the core of the Web Archiving… (11-16 June) hosted by the British Library and the School of Advanced Study. The Web Ar… Week will also include a datathon for researchers as well as a public event aimed to ra… awareness of web archiving in the series of popular Digital Conversations at the British Libr…

The Web Archiving Week events are hosted by the British Library and the School of Ad… Study and they are organised with the support and assistance of IIPC, RESAW (A Re…

- ➢ Interface-on-demand
- ➢ Shadow DOM, no CSS intrusion
- ➢ More consistent, intuitive Miller columns for many captures

IIPC Web Archiving Conference 2017
June 15, 2017
London, UK

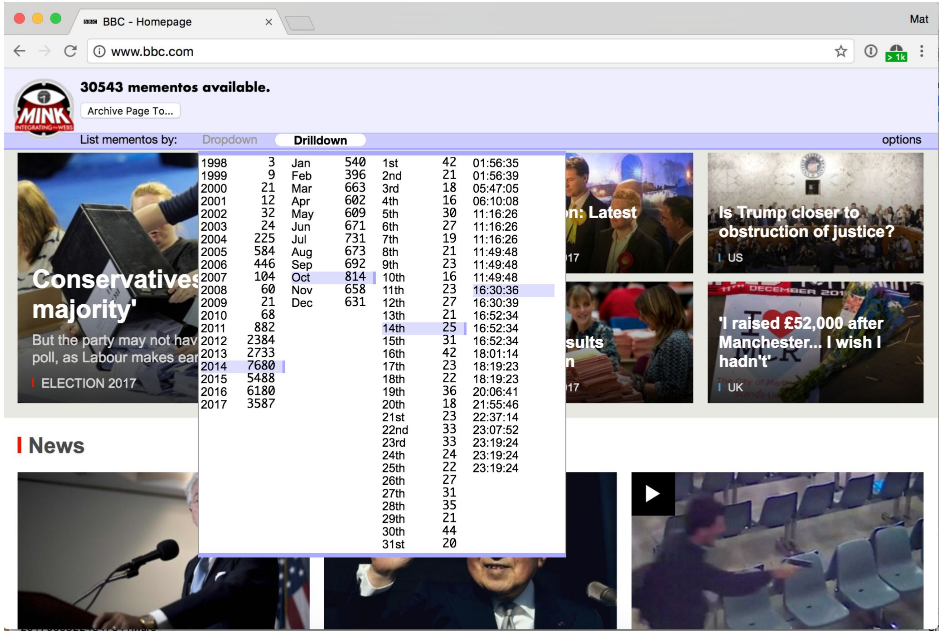
http://bit.ly/iipcWAC2017

@machawk1

NATIONAL ENDOWMENT FOR THE Humanities
HD-51670-13 • HK-50181-14

Mink - User Interface Revamp

MINK
INTEGRATING the WEBS

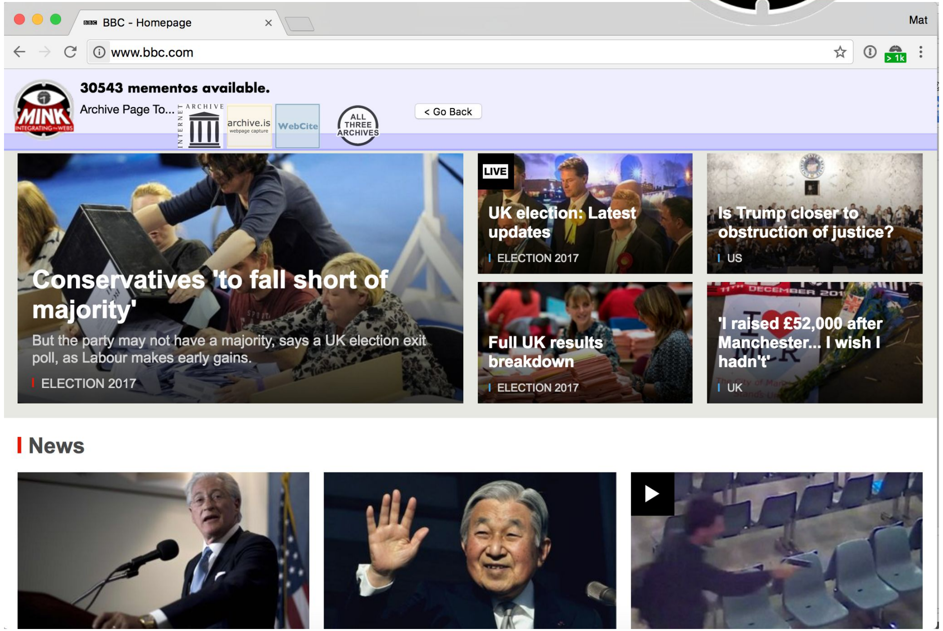


30543 mementos available.

Archive Page To...

List mementos by: Dropdown **Drilldown** options

1998	3	Jan	540	1st	42	01:56:35
1999	9	Feb	396	2nd	21	01:56:39
2000	21	Mar	663	3rd	18	05:47:05
2001	12	Apr	602	4th	16	06:10:08
2002	32	May	609	5th	30	11:16:26
2003	24	Jun	671	6th	27	11:16:26
2004	225	Jul	731	7th	19	11:16:26
2005	584	Aug	673	8th	21	11:49:48
2006	446	Sep	692	9th	23	11:49:48
2007	104	Oct	814	10th	16	11:49:48
2008	60	Nov	658	11th	23	16:30:36
2009	21	Dec	631	12th	27	16:30:39
2010	68			13th	21	16:52:34
2011	882			14th	25	16:52:34
2012	2384			15th	31	16:52:34
2013	2733			16th	42	18:01:14
2014	7680			17th	23	18:19:23
2015	5488			18th	22	18:19:23
2016	6180			19th	36	20:06:41
2017	3587			20th	18	21:55:46
				21st	23	22:37:14
				22nd	33	23:07:52
				23rd	33	23:19:24
				24th	24	23:19:24
				25th	22	23:19:24
				26th	27	
				27th	31	
				28th	35	
				29th	21	
				30th	44	
				31st	20	



30543 mementos available.

Archive Page To...

< Go Back

Conservatives 'to fall short of majority'

But the party may not have a majority, says a UK election exit poll, as Labour makes early gains.

ELECTION 2017

UK election: Latest updates

ELECTION 2017

Is Trump closer to obstruction of justice?

US

Full UK results breakdown

ELECTION 2017

'I raised £52,000 after Manchester... I wish I hadn't'

UK

News

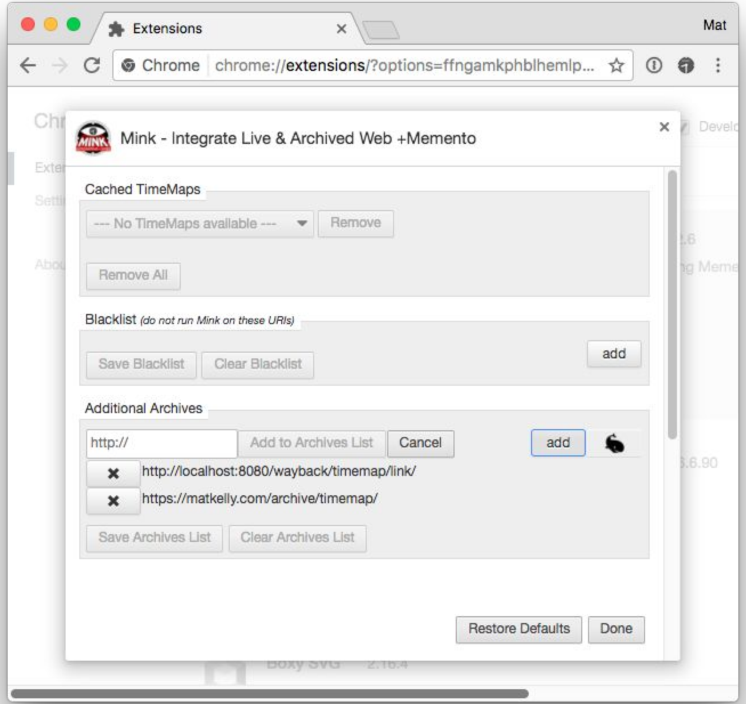
IIPC Web Archiving Conference 2017
June 15, 2017
London, UK

http://bit.ly/iipcWAC2017

@machawk1

NATIONAL ENDOWMENT FOR THE
Humanities
HD-51670-13 • HK-50181-14

Mink - Communication with Local Archives



Mink - Integrate Live & Archived Web +Memento

Cached TimeMaps

--- No TimeMaps available --- ▾ Remove

Remove All

Blacklist *(do not run Mink on these URIs)*

Save Blacklist | Clear Blacklist | add

Additional Archives

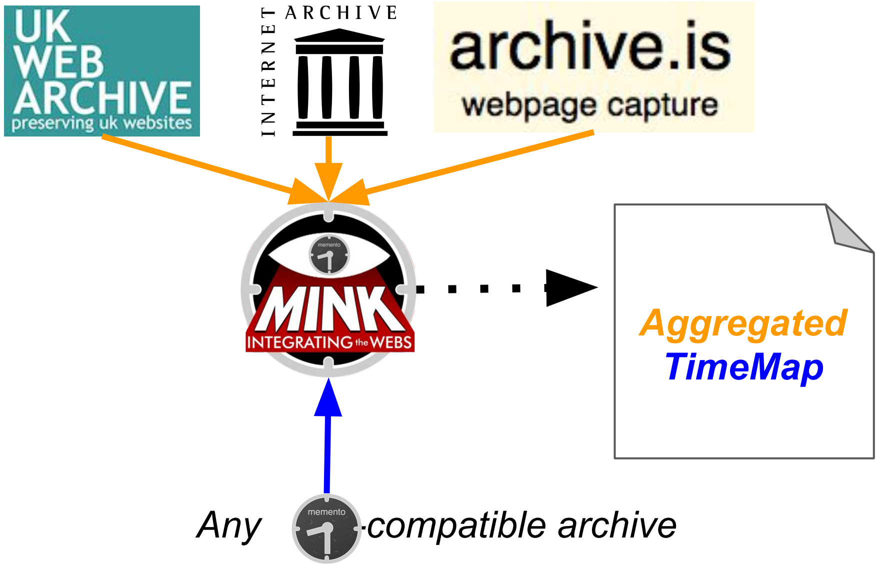
http:// | Add to Archives List | Cancel | add

✖ http://localhost:8080/wayback/timemap/link/

✖ https://matkelly.com/archive/timemap/

Save Archives List | Clear Archives List

Restore Defaults | Done



Aggregated TimeMap

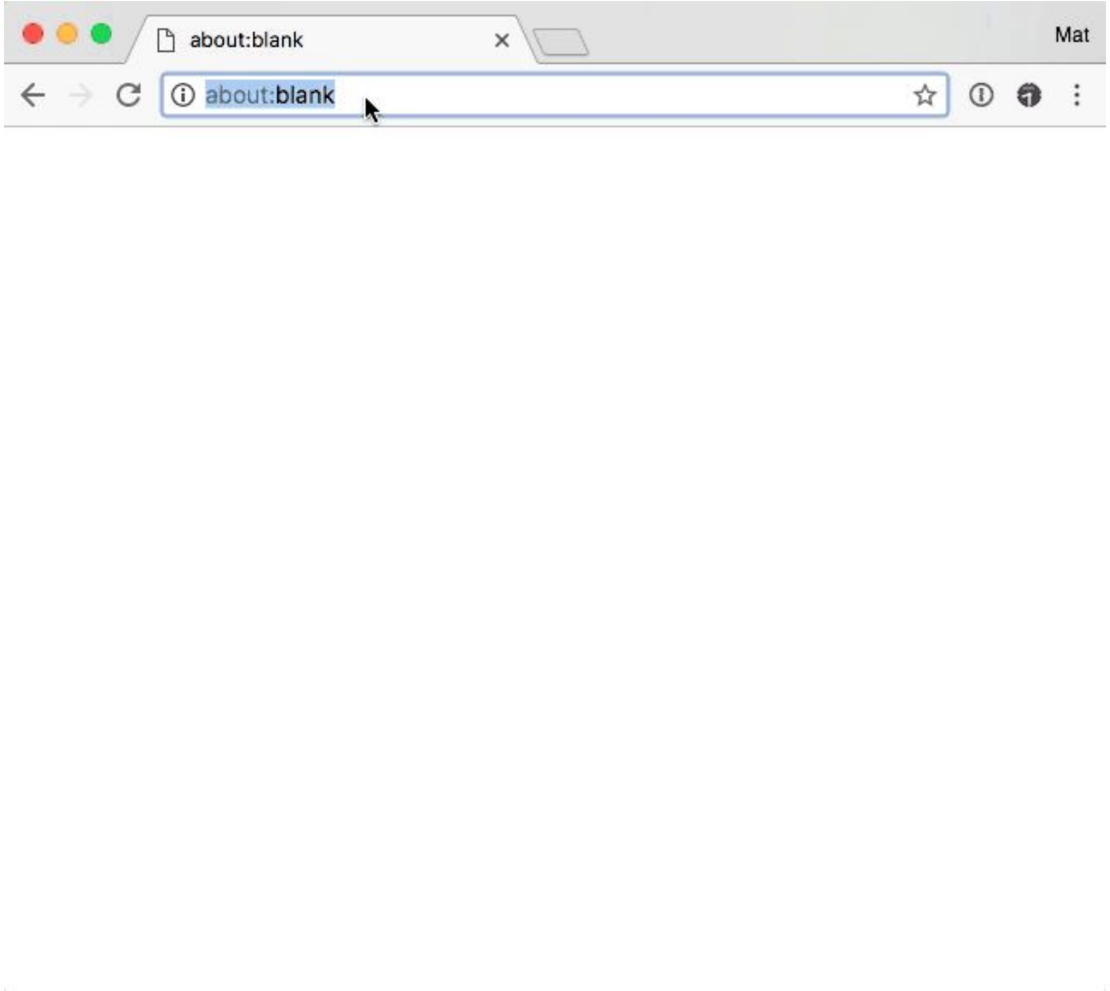
Any memento *compatible archive*

IIPC Web Archiving Conference 2017
June 15, 2017
London, UK

http://bit.ly/iipcWAC2017

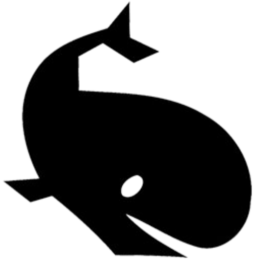
@machawk1

NATIONAL ENDOWMENT FOR THE Humanities
HD-51670-13 • HK-50181-14



Mink usage GIF, also available at:
https://youtu.be/bGjxofpTgv4

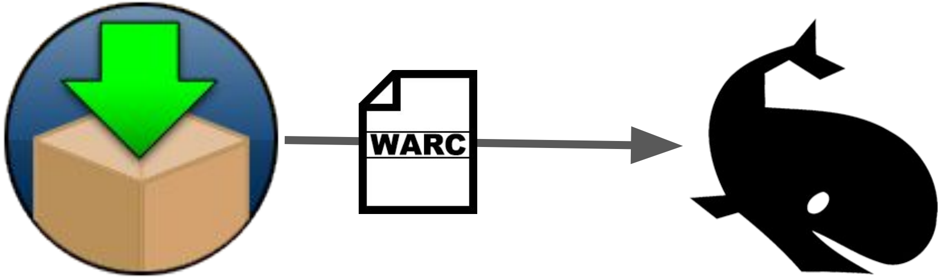
Tools’ Integration



NATIONAL ENDOWMENT FOR THE
Humanities
HD-51670-13 • HK-50181-14

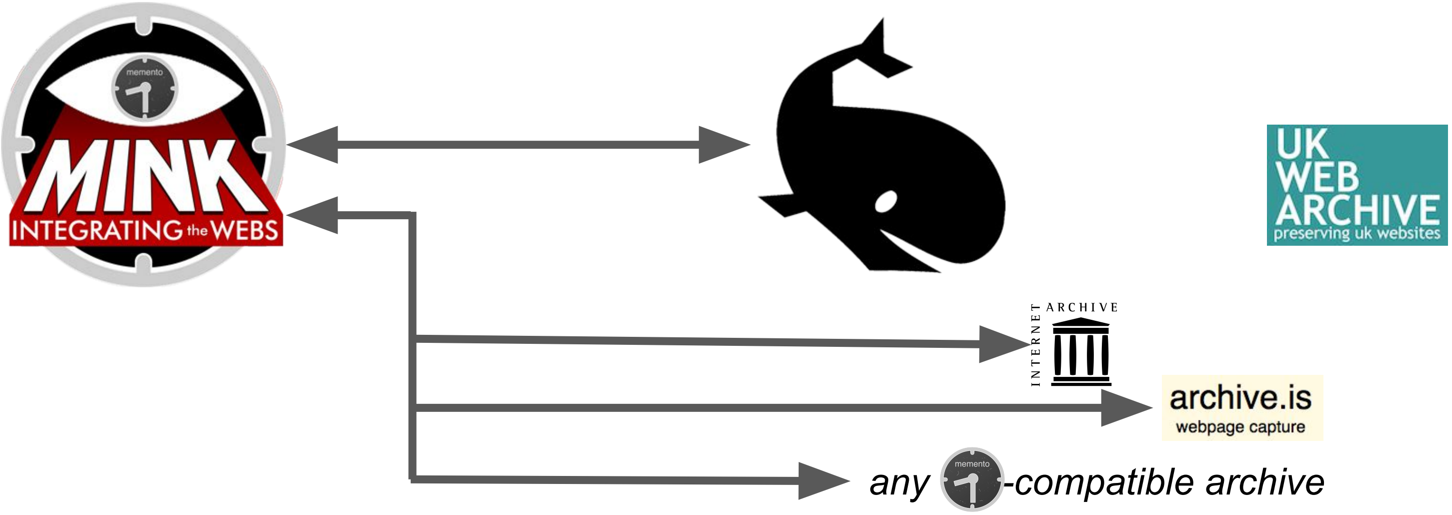
Tools' Integration: WARCreate→WAIL

- Save WARC directly to local archive (by reference [easier]
    - By-value integration feasibility being investigated a la WASAPI
- Automatically indexed and replayable



NATIONAL ENDOWMENT FOR THE Humanities
HD-51670-13 • HK-50181-14

Tools' Integration: Mink→WAIL



any -compatible archive

IIPC Web Archiving Conference 2017
June 15, 2017
London, UK

http://bit.ly/iipcWAC2017

@machawk1

Some Future Work

- Decouple Mink from external Memento aggregator
  - Client-side customizable set of archives instead
- WARC replay using browser extensions/apps
- Further integration with other archiving tools in WAIL
  - Re-add Memgator Memento aggregator (removed from Electron version)
- Firefox version of tools
  - XUL→WebExtensions
  - Decouple from Chrome APIs
- Integration with InterPlanetary Wayback (speaking about later today)

NATIONAL ENDOWMENT FOR THE
Humanities
HD-51670-13 • HK-50181-14

Acknowledgements

- NEH Grant #s HD-51670-13 • HK-50181-14
- Dr. Liza Potts and WIDE Research Center at Michigan State University
- ODU SEES Travel Grant

NATIONAL ENDOWMENT FOR THE
Humanities
HD-51670-13 • HK-50181-14

Archive What I See Now

Personal Web Archiving with WARCs

Michele C. Weigle, Michael L. Nelson, **Mat Kelly**, and John Berlin
Web Science and Digital Libraries (WS-DL) Research Group
Old Dominion University
ws-dl.cs.odu.edu

@machawk1

IIPC Web Archiving Conference 2017
June 15, 2017
London, UK

http://bit.ly/iipcWAC2017

NATIONAL ENDOWMENT FOR THE Humanities
HD-51670-13 • HK-50181-14