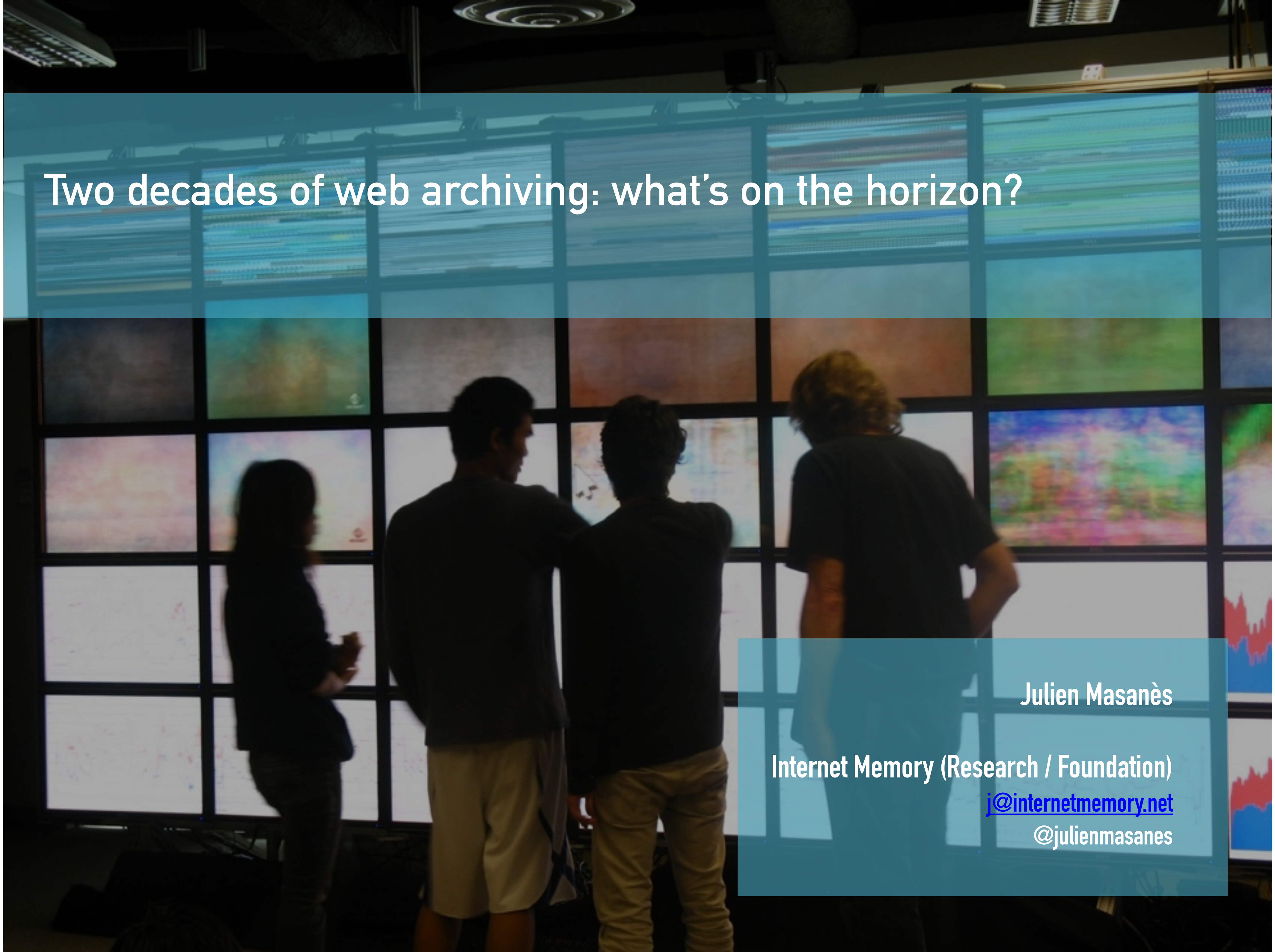
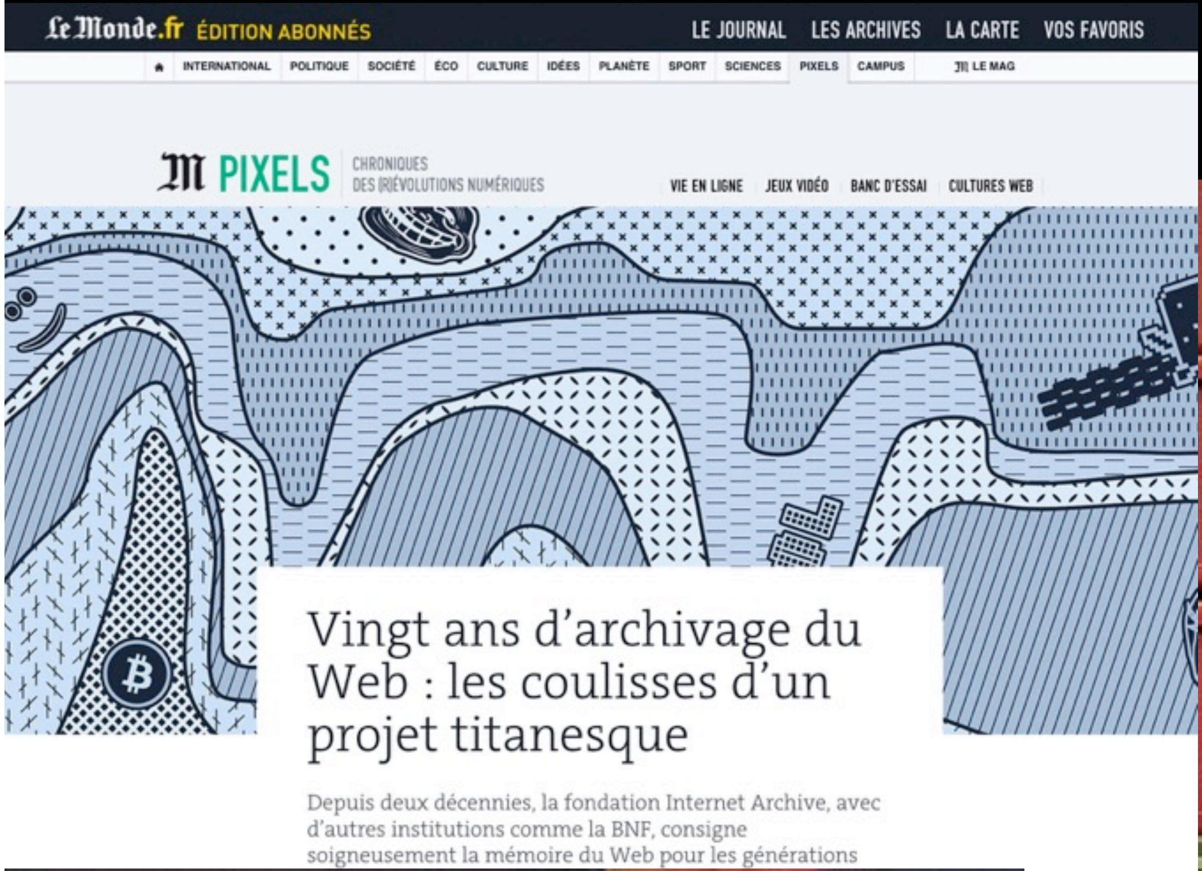
Two decades of web archiving: what's on the horizon?



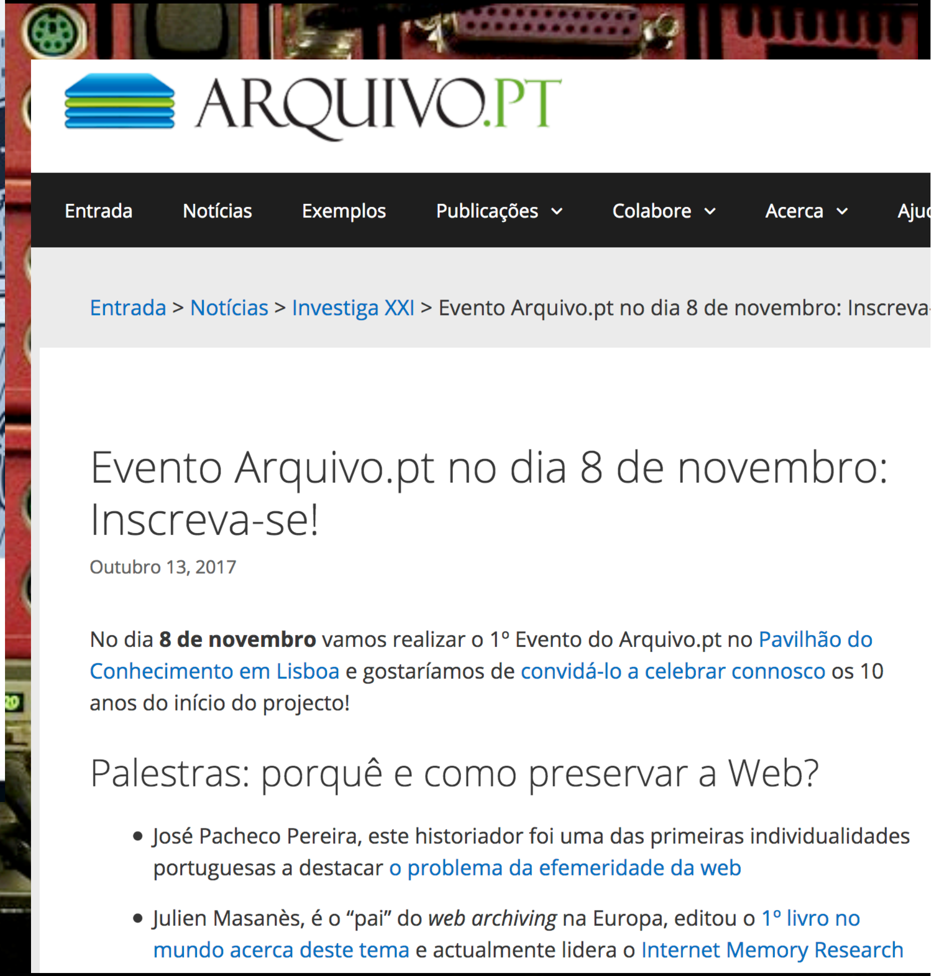
Julien Masanès

Internet Memory (Research / Foundation)
j@internetmemory.net
@julienmasanes



Vingt ans d'archivage du Web : les coulisses d'un projet titanesque

Depuis deux décennies, la fondation Internet Archive, avec d'autres institutions comme la BNF, consigne soigneusement la mémoire du Web pour les générations



Entrada > Notícias > Investiga XXI > Evento Arquivo.pt no dia 8 de novembro: Inscreva

Evento Arquivo.pt no dia 8 de novembro: Inscreva-se!

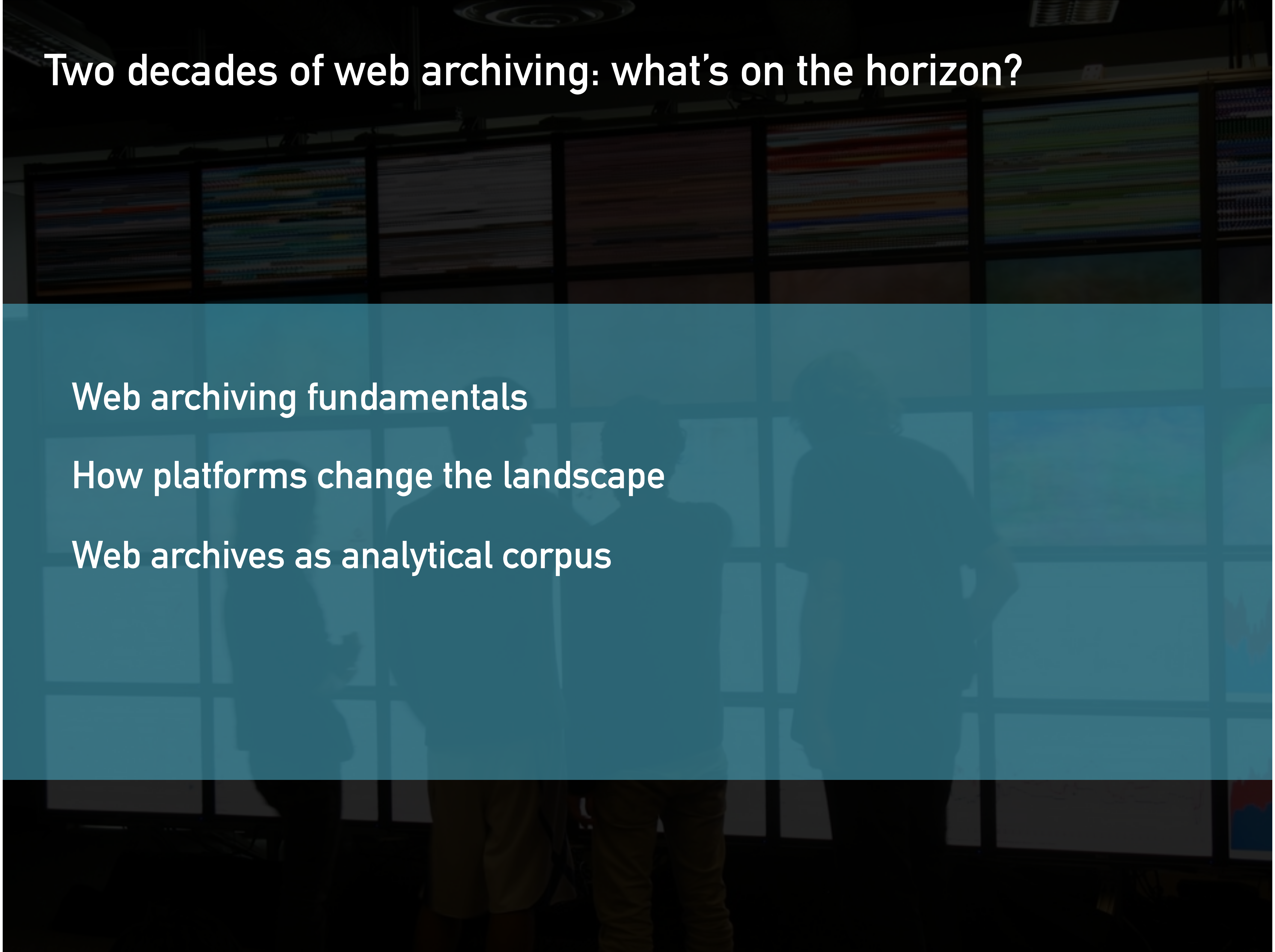
Outubro 13, 2017

No dia **8 de novembro** vamos realizar o 1º Evento do Arquivo.pt no Pavilhão do Conhecimento em Lisboa e gostaríamos de convidá-lo a celebrar connosco os 10 anos do início do projecto!

Palestras: porquê e como preservar a Web?

- José Pacheco Pereira, este historiador foi uma das primeiras individualidades portuguesas a destacar o problema da efemeridade da web
- Julien Masanès, é o "pai" do *web archiving* na Europa, editou o 1º livro no mundo acerca deste tema e actualmente lidera o Internet Memory Research

Two decades of web archiving: what's on the horizon?

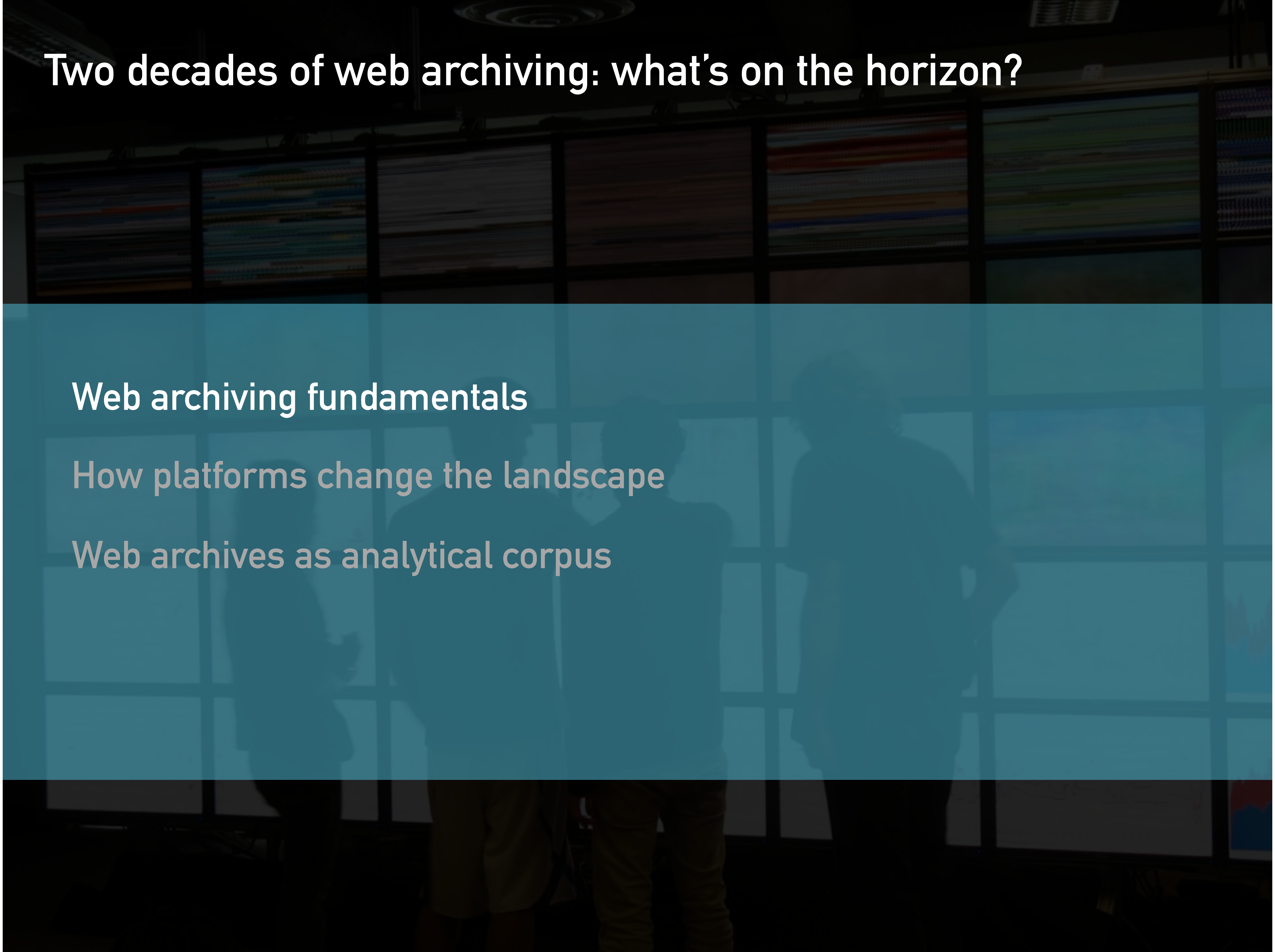


Web archiving fundamentals

How platforms change the landscape

Web archives as analytical corpus

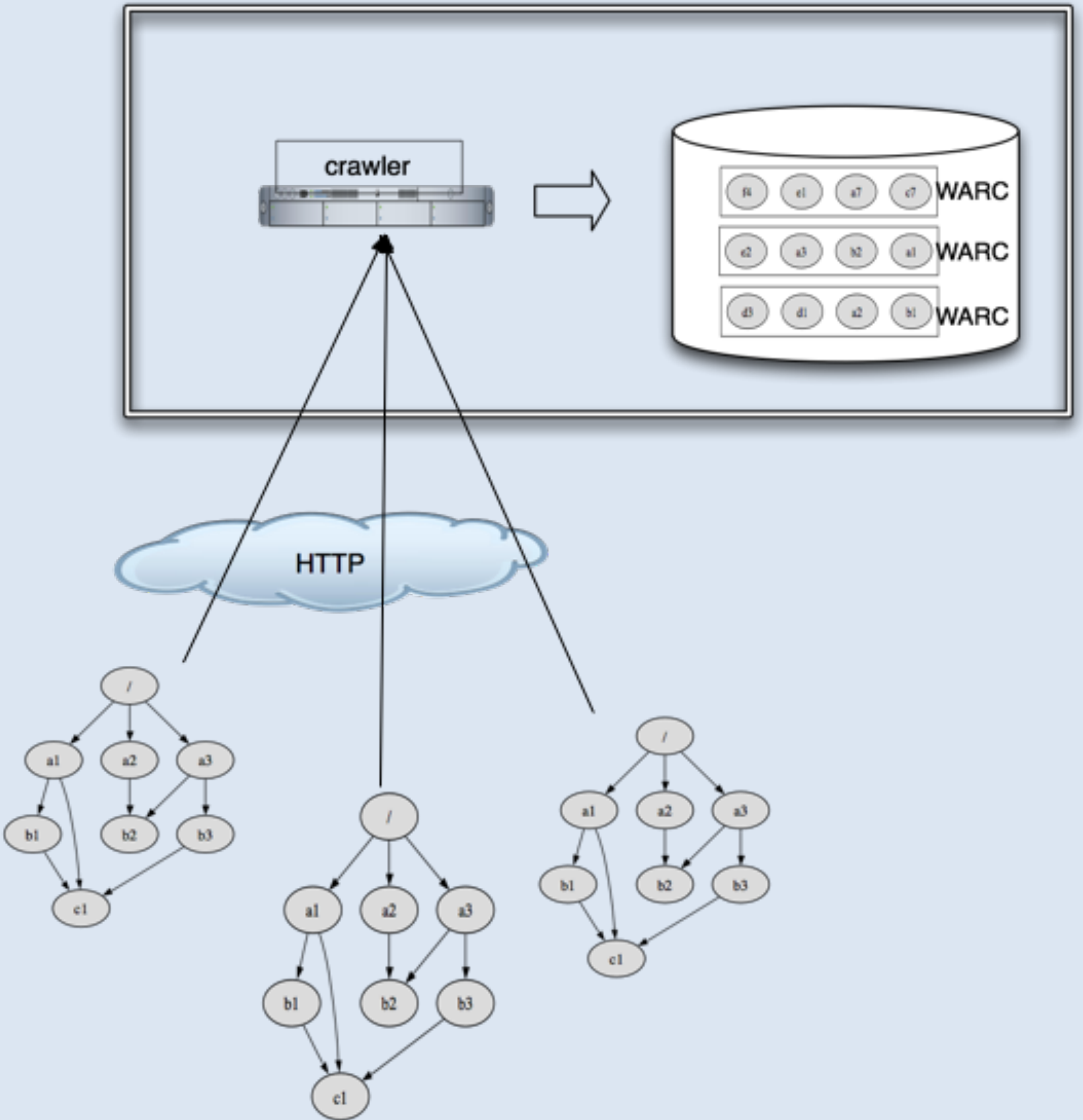
Two decades of web archiving: what's on the horizon?



Web archiving fundamentals

How platforms change the landscape

Web archives as analytical corpus


crawler
WARC
f4
e1
a7
c7
WARC
c2
a3
b2
a1
WARC
d3
d1
a2
b1
HTTP
/
a1
a2
a3
b1
b2
b3
c1
/
a1
a2
a3
b1
b2
b3
c1
/
a1
a2
a3
b1
b2
b3
c1

Shaping the archive

1. Explicit limits

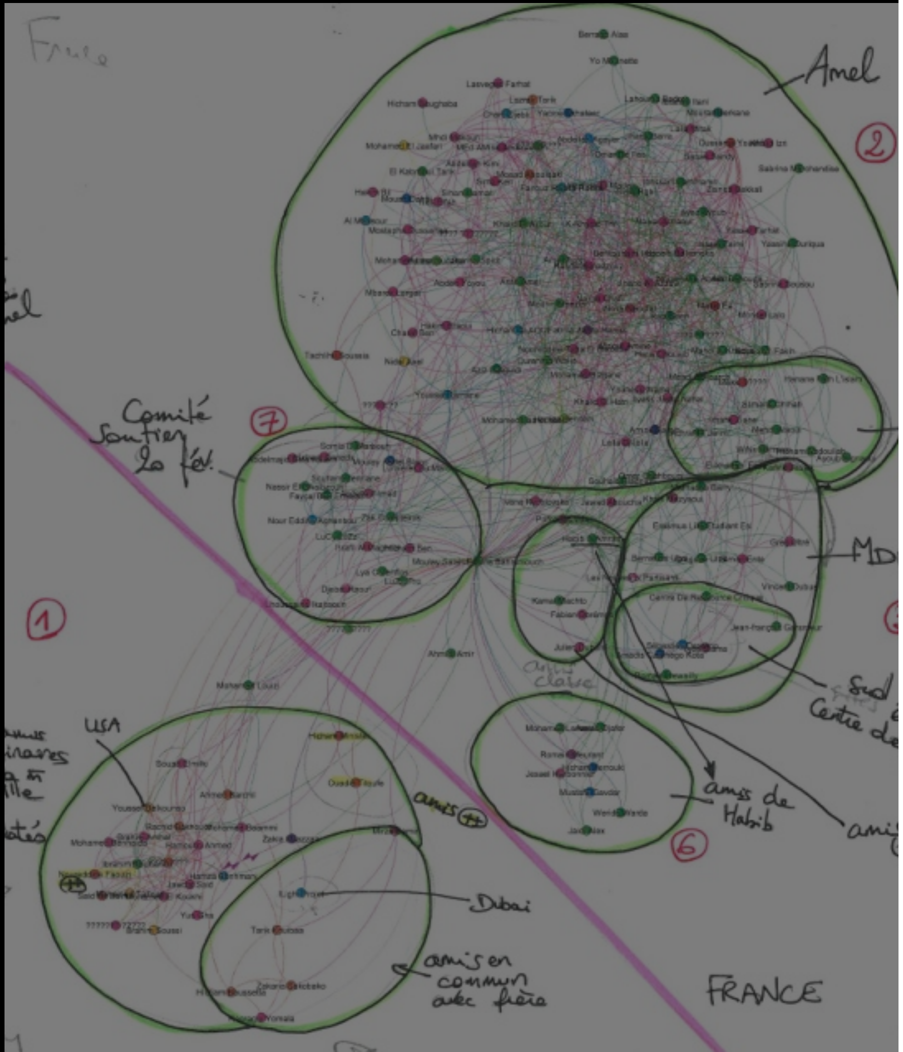
   robots.txt

2. Implicit limits (no DoS)

   Politeness

3. Technical limits (arm race)

   Client-side code execution



Cardinality of preserved object

Different for different institutions (museums, archives, libraries)

Cardinality of the incunabula

20 M (books) / 30 000 (editions) = 650

Consequence of high cardinality for preservation: redundancy and time

What about the Web?

Virtually infinite number of copies

High dependancy on one single server

More like a museum than like libraries

Active publishing system

- Web Information Systems
- Controlled by the producer
- Continuous publishing

Preserving: actively pursuing the exact opposite

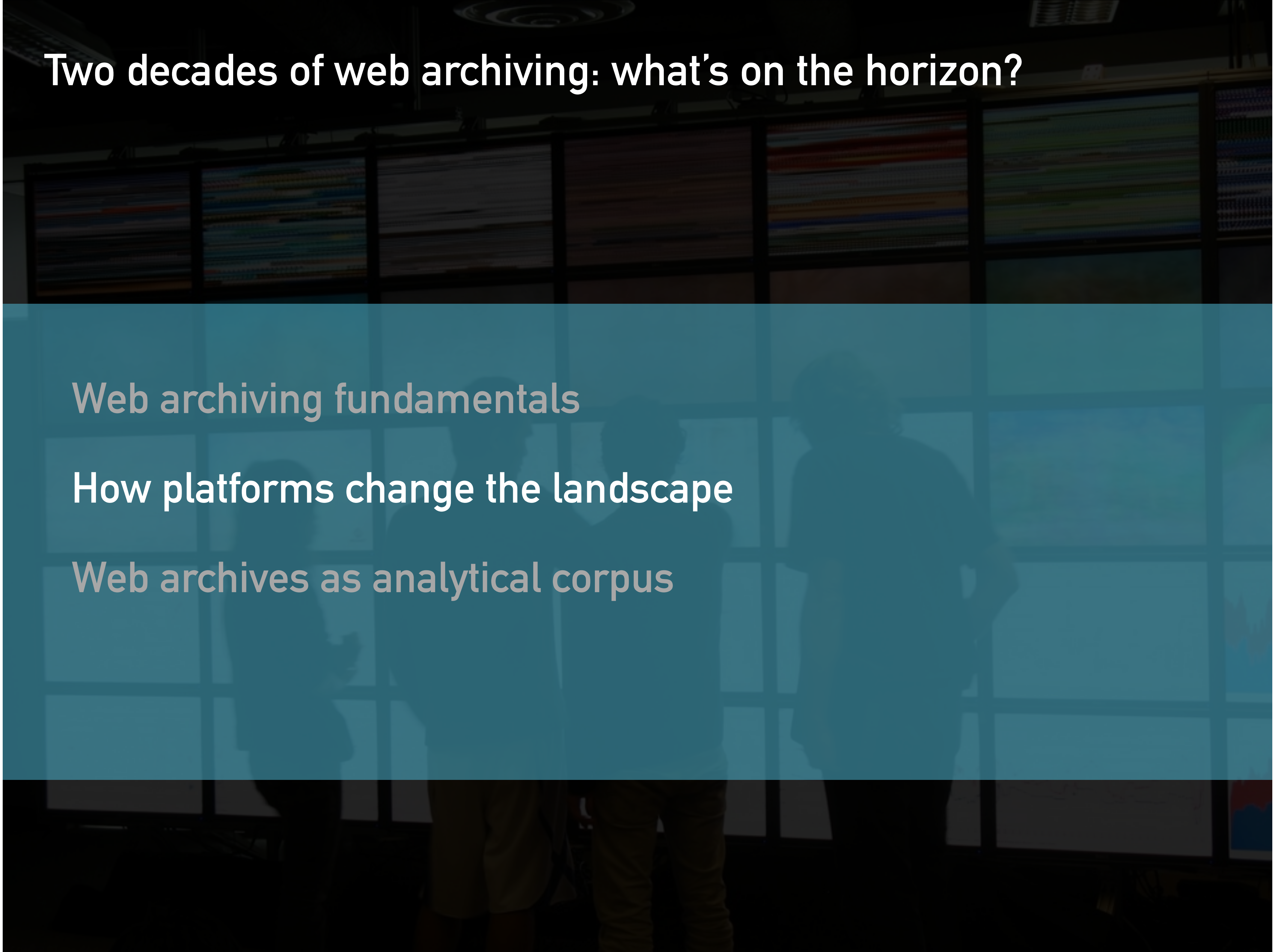
Internet Memory?

- Automated sampling of a virtually infinite information space

- Search and analysis specially leveraging the temporal dimension of the archive

- Has to be part of the internet

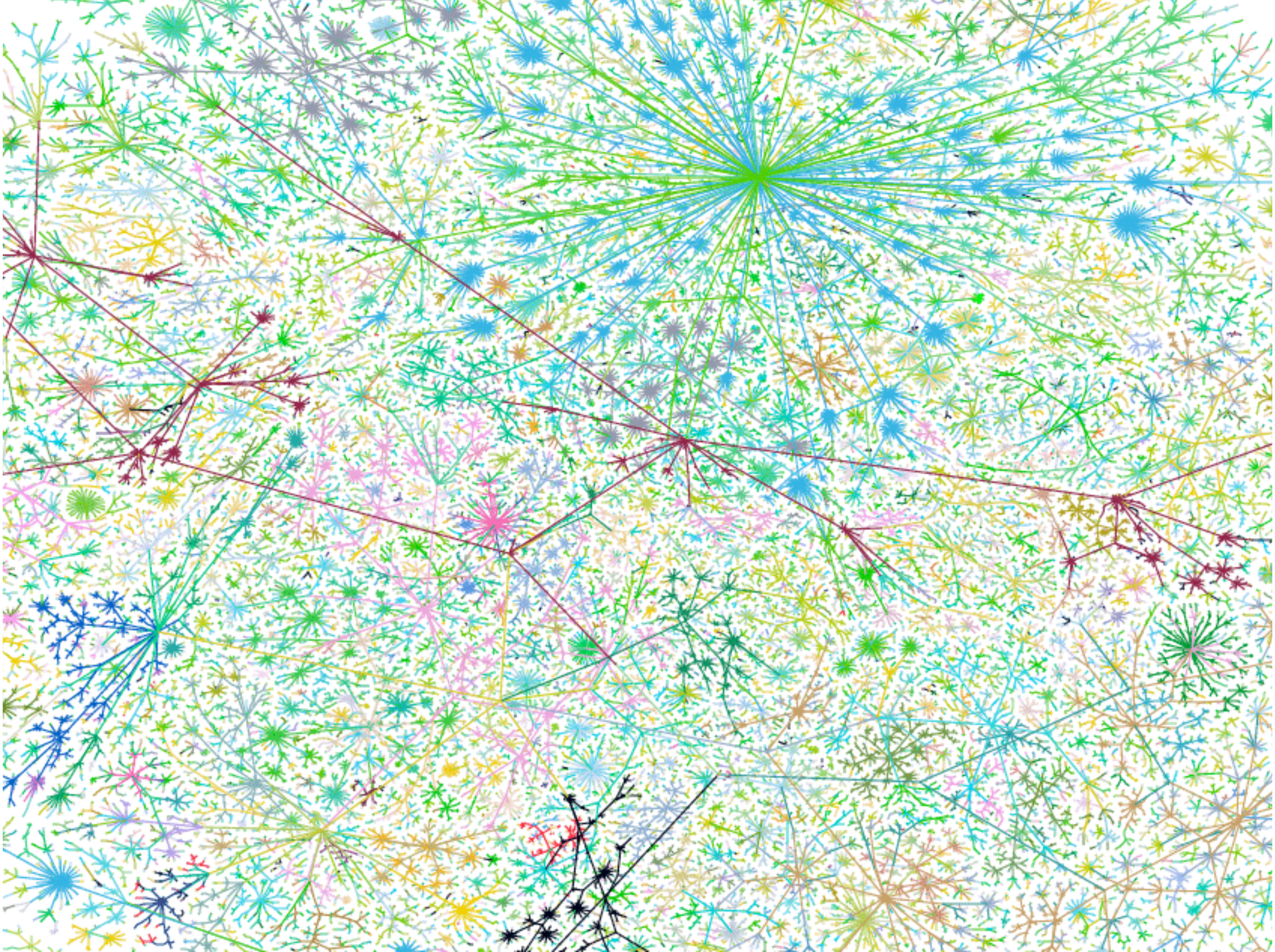
Two decades of web archiving: what's on the horizon?



Web archiving fundamentals

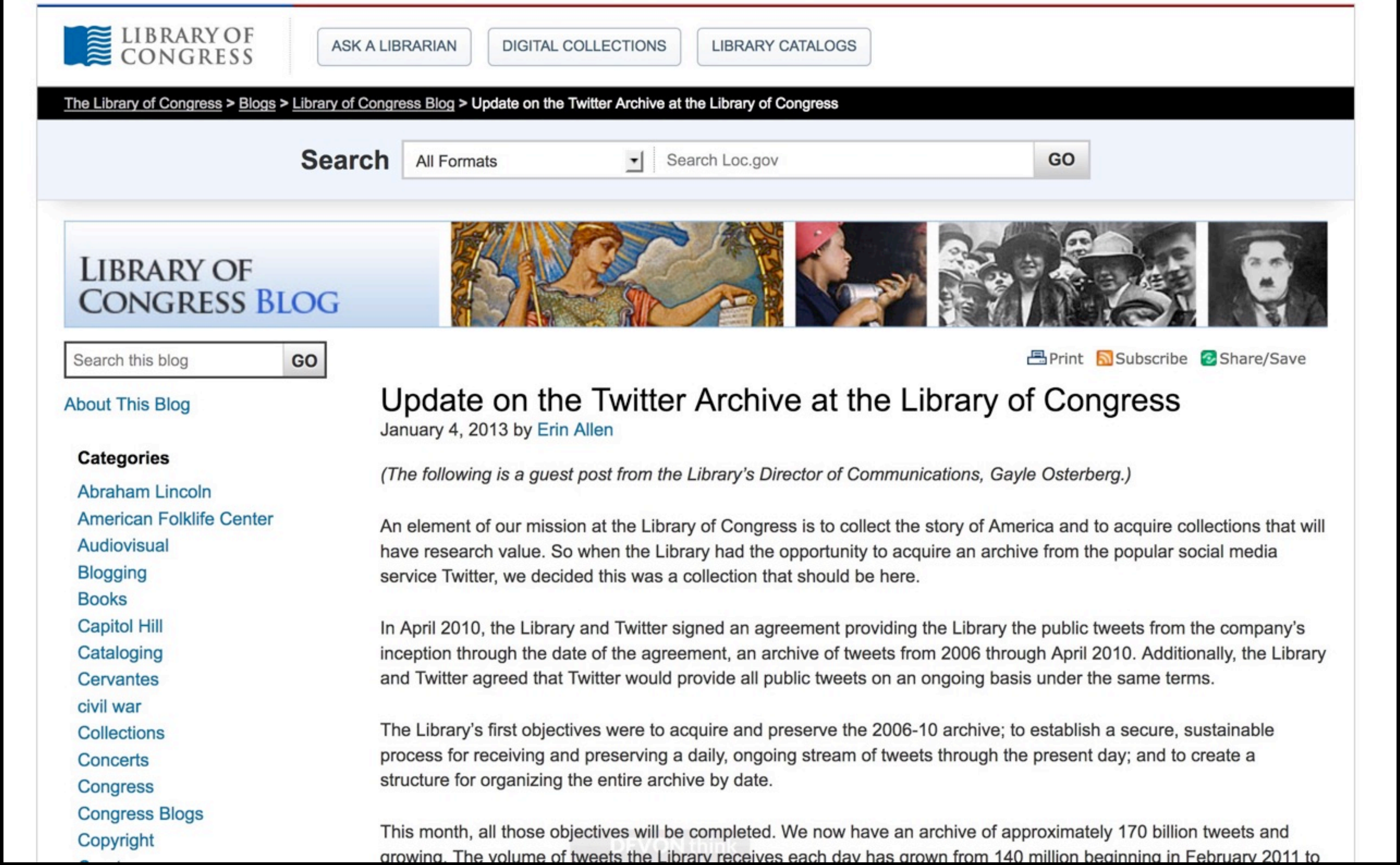
How platforms change the landscape

Web archives as analytical corpus



You
Tube
X

A partir du 28 avril
VAGUE N°5
TOUJOURS PLUS DE POUVOIR D'ACHAT
BAISSE de PRIX
SUR PLUS DE 100 PRODUITS
identifiés en rayon.
+15% GRATUIT
HUGGIES
Okay
presse



Update on the Twitter Archive at the Library of Congress

January 4, 2013 by Erin Allen

(The following is a guest post from the Library's Director of Communications, Gayle Osterberg.)

An element of our mission at the Library of Congress is to collect the story of America and to acquire collections that will have research value. So when the Library had the opportunity to acquire an archive from the popular social media service Twitter, we decided this was a collection that should be here.

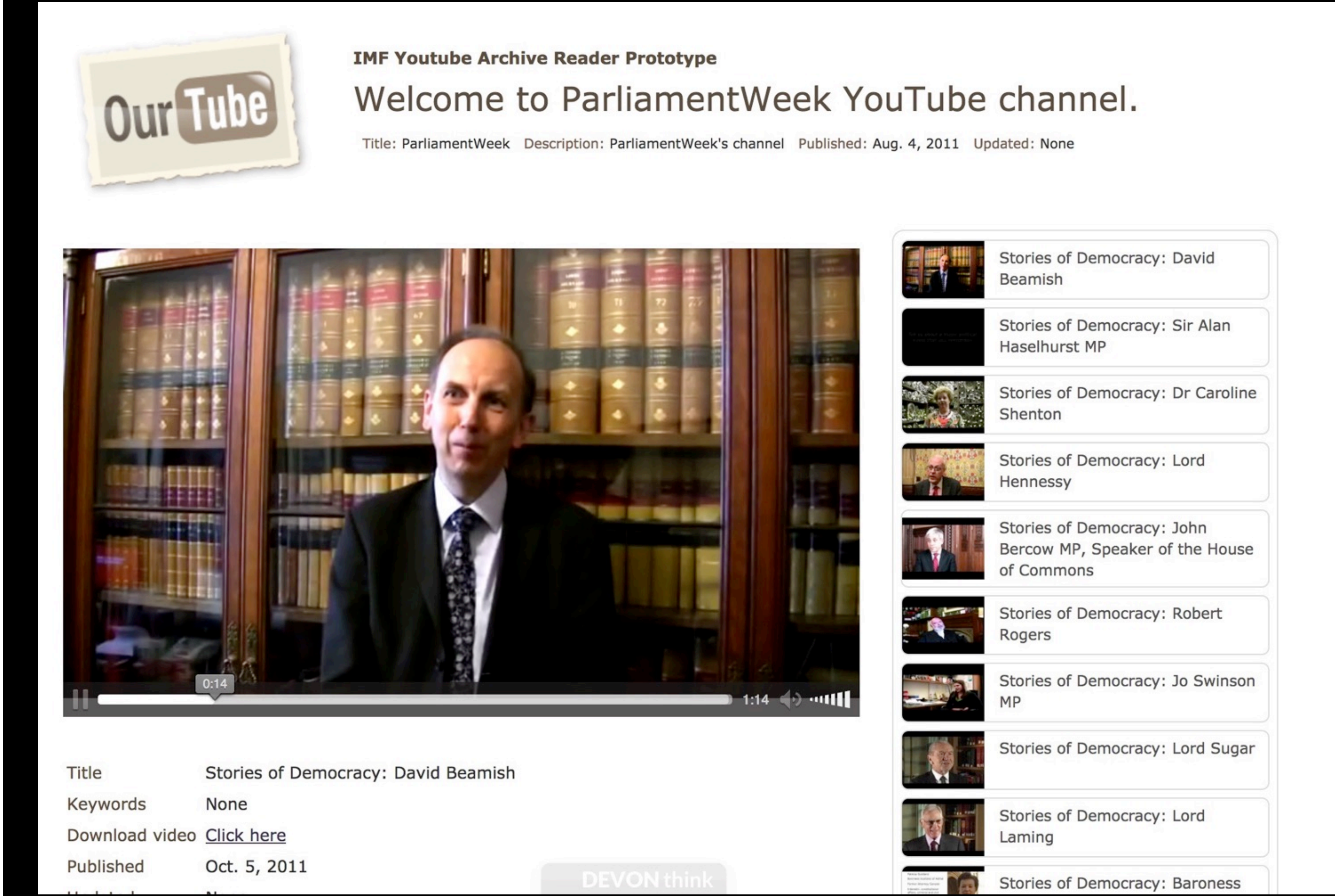
In April 2010, the Library and Twitter signed an agreement providing the Library the public tweets from the company's inception through the date of the agreement, an archive of tweets from 2006 through April 2010. Additionally, the Library and Twitter agreed that Twitter would provide all public tweets on an ongoing basis under the same terms.

The Library's first objectives were to acquire and preserve the 2006-10 archive; to establish a secure, sustainable process for receiving and preserving a daily, ongoing stream of tweets through the present day; and to create a structure for organizing the entire archive by date.

This month, all those objectives will be completed. We now have an archive of approximately 170 billion tweets and growing. The volume of tweets the Library receives each day has grown from 140 million beginning in February 2011 to

Preserving the platforms?

- Who decides?
  - A handful of private company can now decide what will be preserved?
- Legal limits
  - They have technical capacity enable archiving, Updating regulation?
- Technical limits
  - We need more machines!
  - API access, how we deal with loss of presentation context?



Our Tube

IMF Youtube Archive Reader Prototype

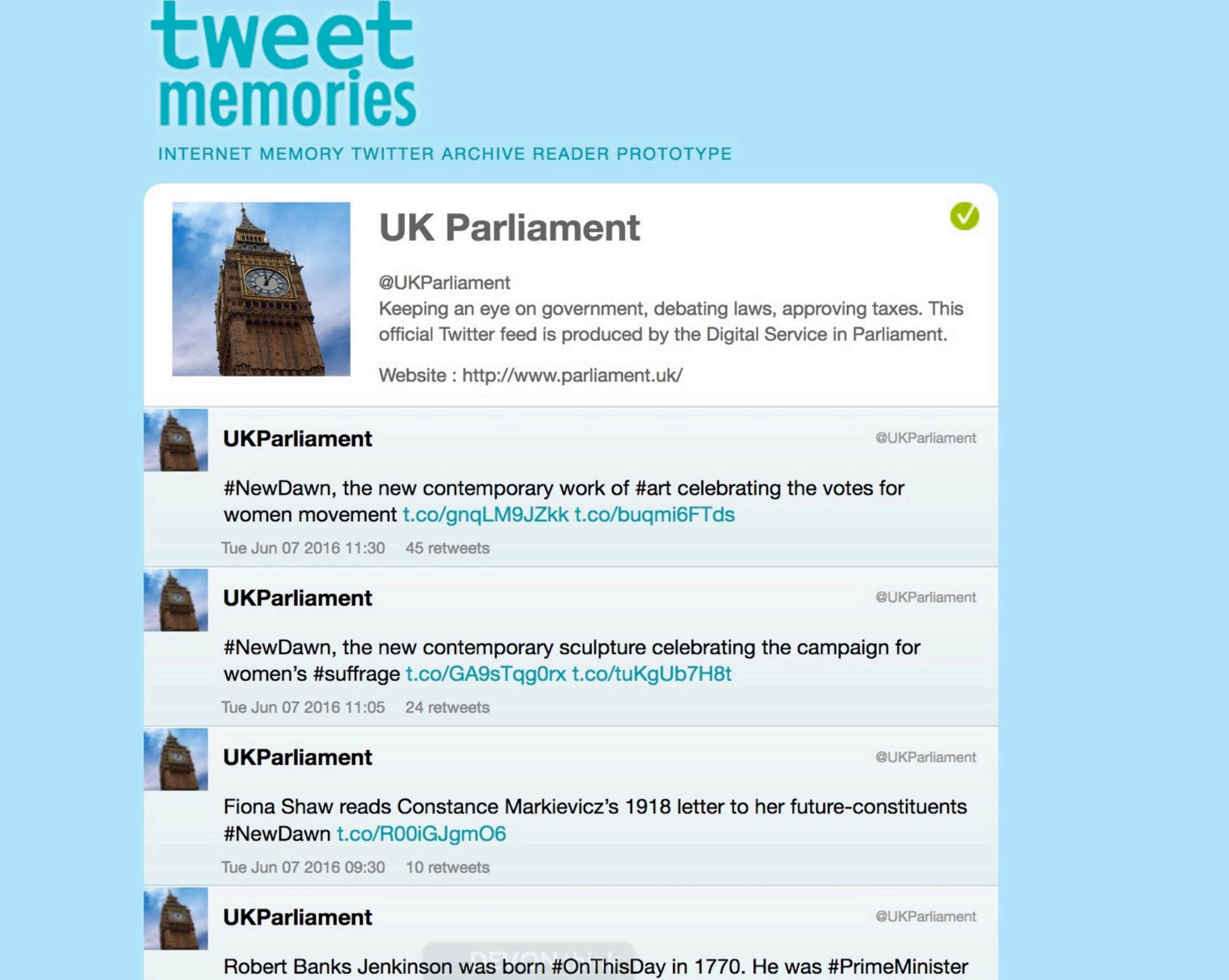
Welcome to ParliamentWeek YouTube channel.

Title: ParliamentWeek Description: ParliamentWeek's channel Published: Aug. 4, 2011 Updated: None

Title	Stories of Democracy: David Beamish
Keywords	None
Download video	Click here
Published	Oct. 5, 2011

- Stories of Democracy: David Beamish
- Stories of Democracy: Sir Alan Haselhurst MP
- Stories of Democracy: Dr Caroline Shenton
- Stories of Democracy: Lord Hennessy
- Stories of Democracy: John Bercow MP, Speaker of the House of Commons
- Stories of Democracy: Robert Rogers
- Stories of Democracy: Jo Swinson MP
- Stories of Democracy: Lord Sugar
- Stories of Democracy: Lord Laming
- Stories of Democracy: Baroness

http://webarchive.parliament.uk/20160607164449/http://www.youtube.com/user/ParliamentWeek



tweet
memories

INTERNET MEMORY TWITTER ARCHIVE READER PROTOTYPE

UK Parliament

@UKParliament
Keeping an eye on government, debating laws, approving taxes. This official Twitter feed is produced by the Digital Service in Parliament.

Website : http://www.parliament.uk/

UKParliament @UKParliament

#NewDawn, the new contemporary work of #art celebrating the votes for women movement t.co/gnqLM9JZkk t.co/buqmi6FTds

Tue Jun 07 2016 11:30 45 retweets

UKParliament @UKParliament

#NewDawn, the new contemporary sculpture celebrating the campaign for women's #suffrage t.co/GA9sTqg0rx t.co/tuKgUb7H8t

Tue Jun 07 2016 11:05 24 retweets

UKParliament @UKParliament

Fiona Shaw reads Constance Markievicz's 1918 letter to her future-constituents #NewDawn t.co/R00iGJgmO6

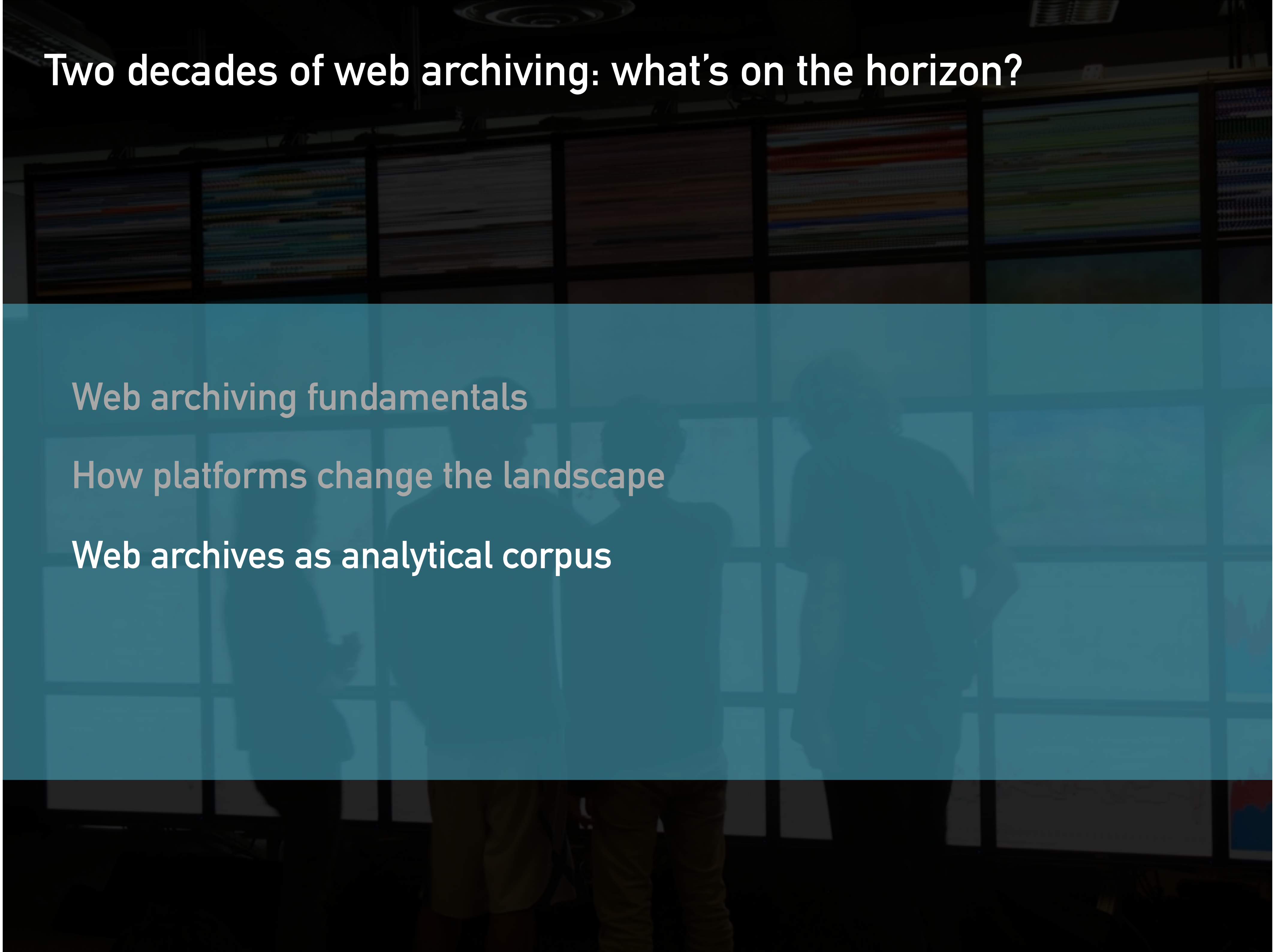
Tue Jun 07 2016 09:30 10 retweets

UKParliament @UKParliament

Robert Banks Jenkinson was born #OnThisDay in 1770. He was #PrimeMinister

http://webarchive.parliament.uk/*/http://twitter.com/ukparliament

Two decades of web archiving: what's on the horizon?



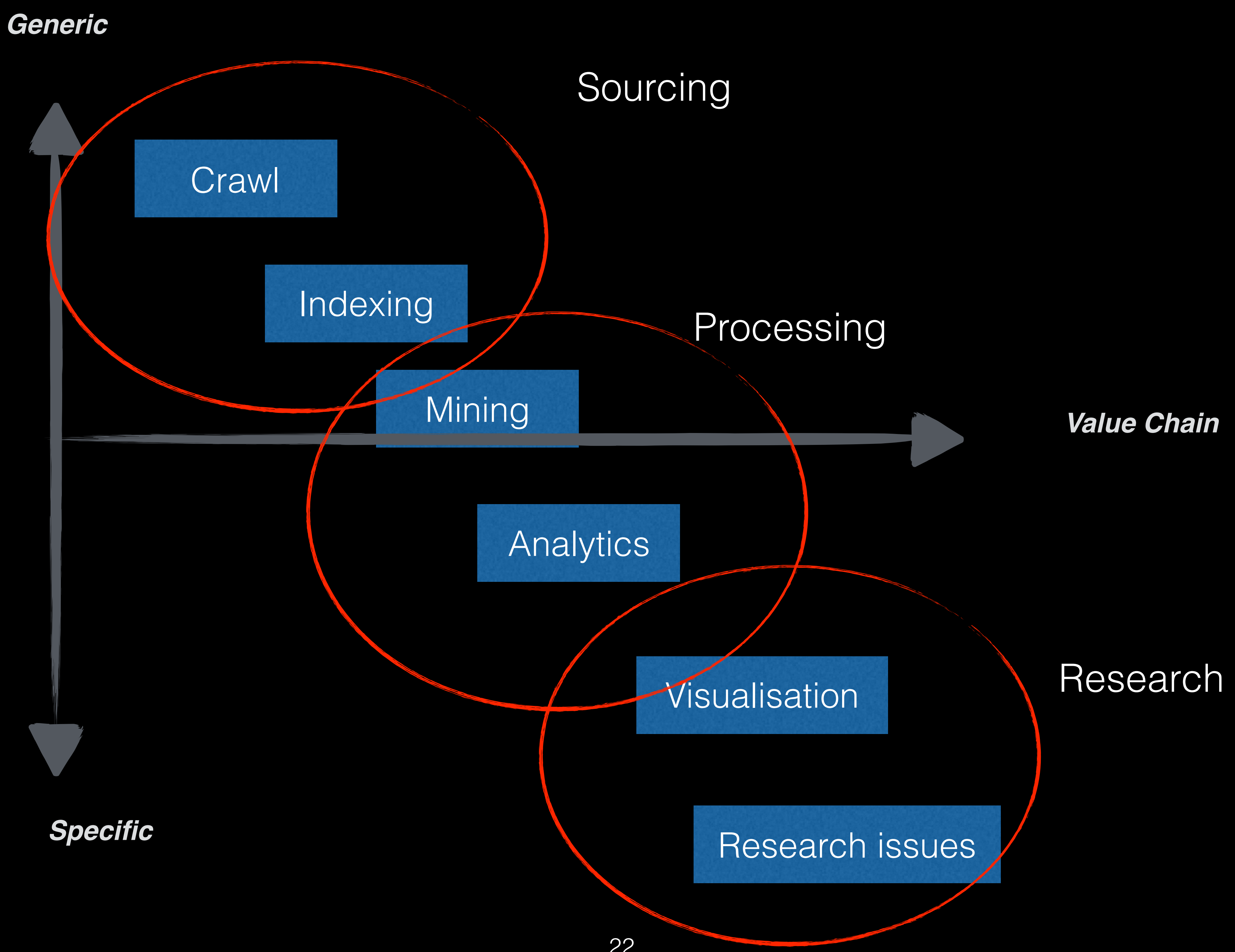
Web archiving fundamentals

How platforms change the landscape

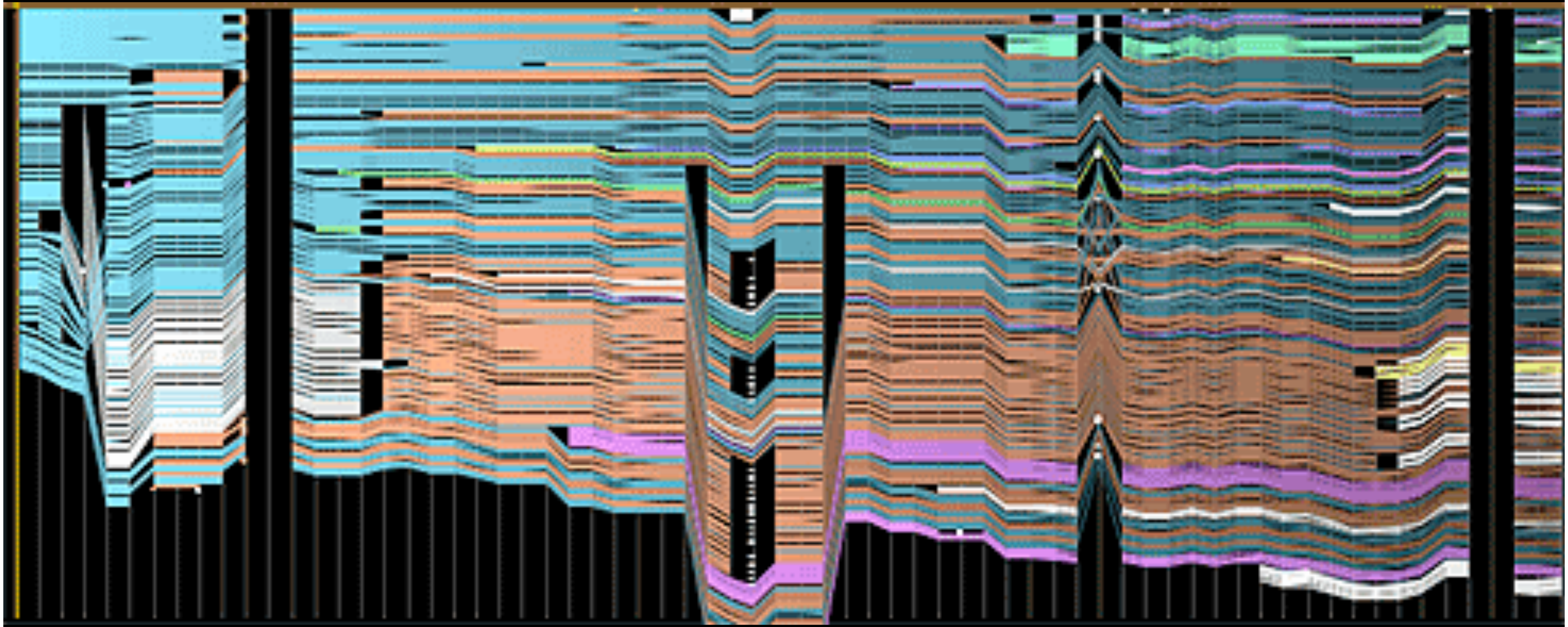
Web archives as analytical corpus

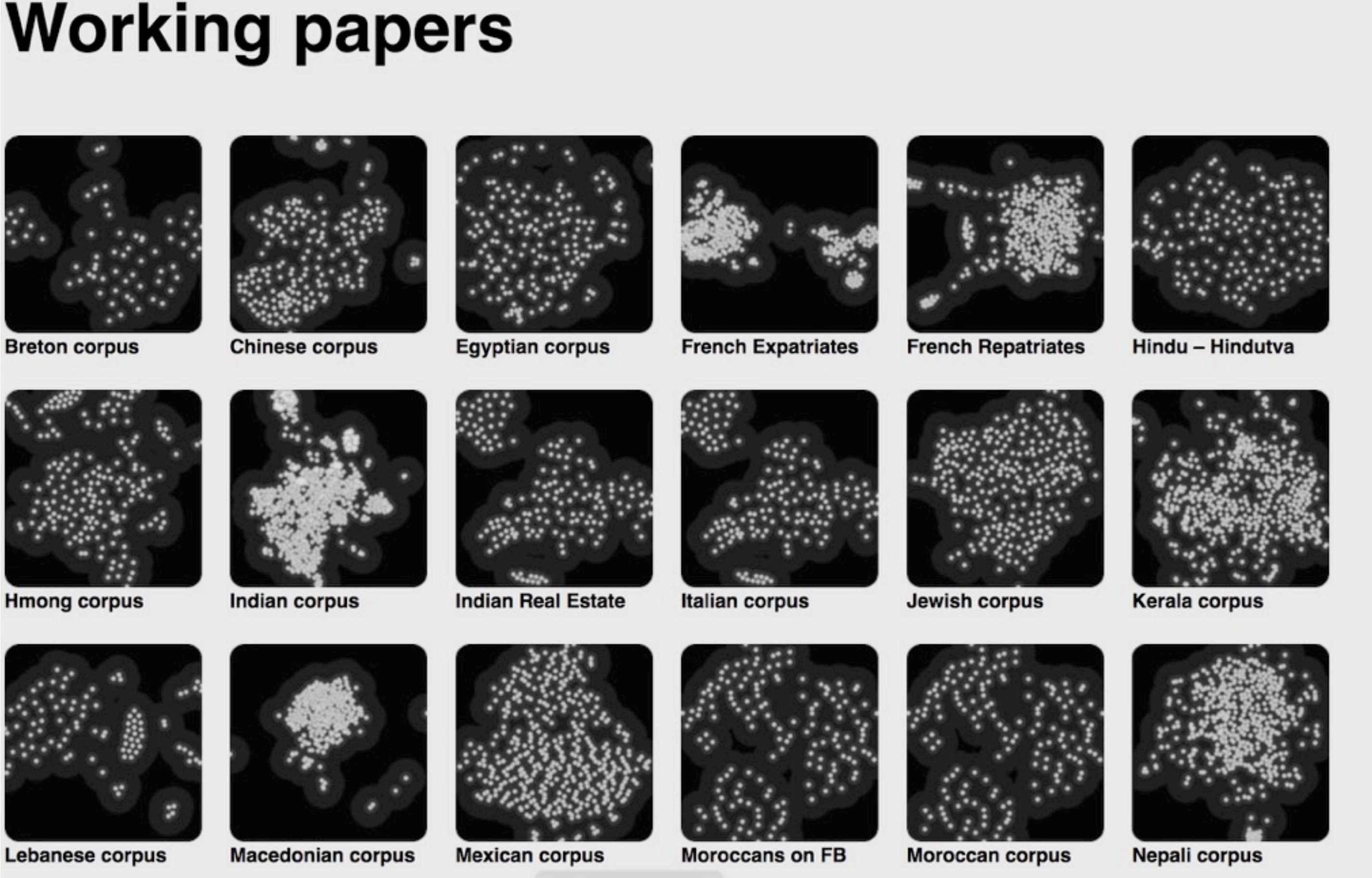
Macroanalysis

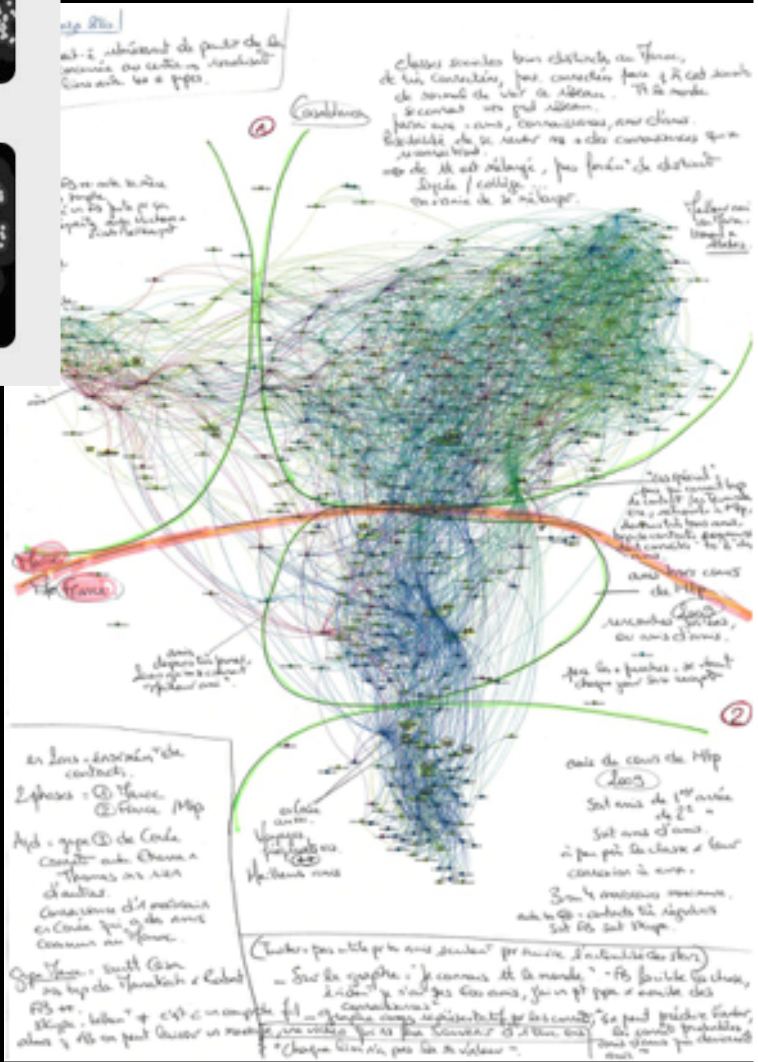
- Focus is no longer single document/page
- Extraction of data
- Indexing
- Analytics: statistical distribution , correlation, graph analyses



Generic
Sourcing
Crawl
Indexing
Processing
Mining
Value Chain
Analytics
Research
Visualisation
Research issues
Specific



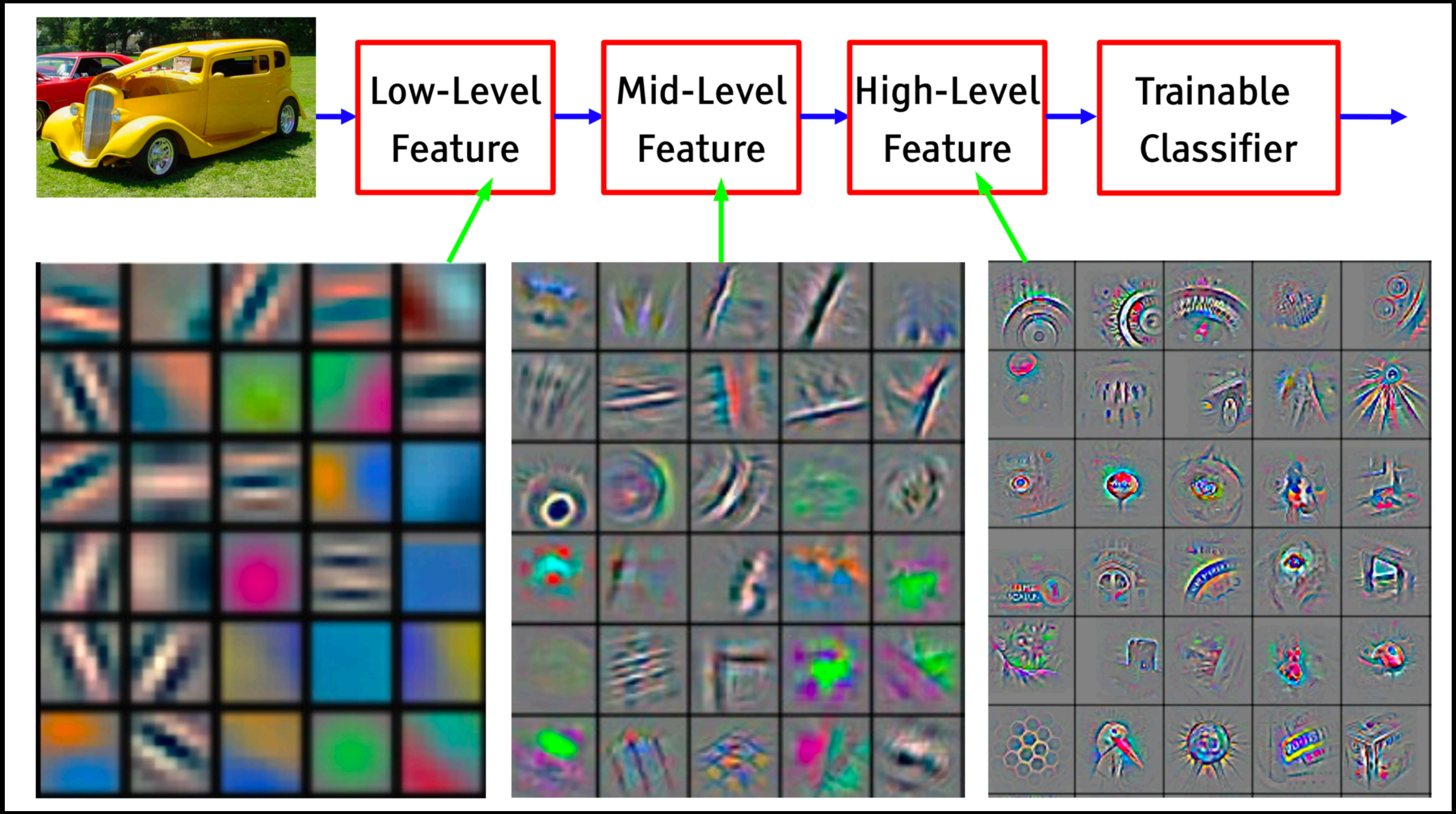




http://www.e-diasporas.fr/

Occupying the social Web: Moroccan students on the move. A method resulting from fieldwork
Sabrina Marchandise, 2012.

New possibilities to large scale analysis of web archives: Deep learning



Feature visualization of convolutional net trained on ImageNet from [Zeiler & Fergus 2013]

Deep learning to detect objects « luggages »

Also to help building the archive!

ARQUIVO.PT

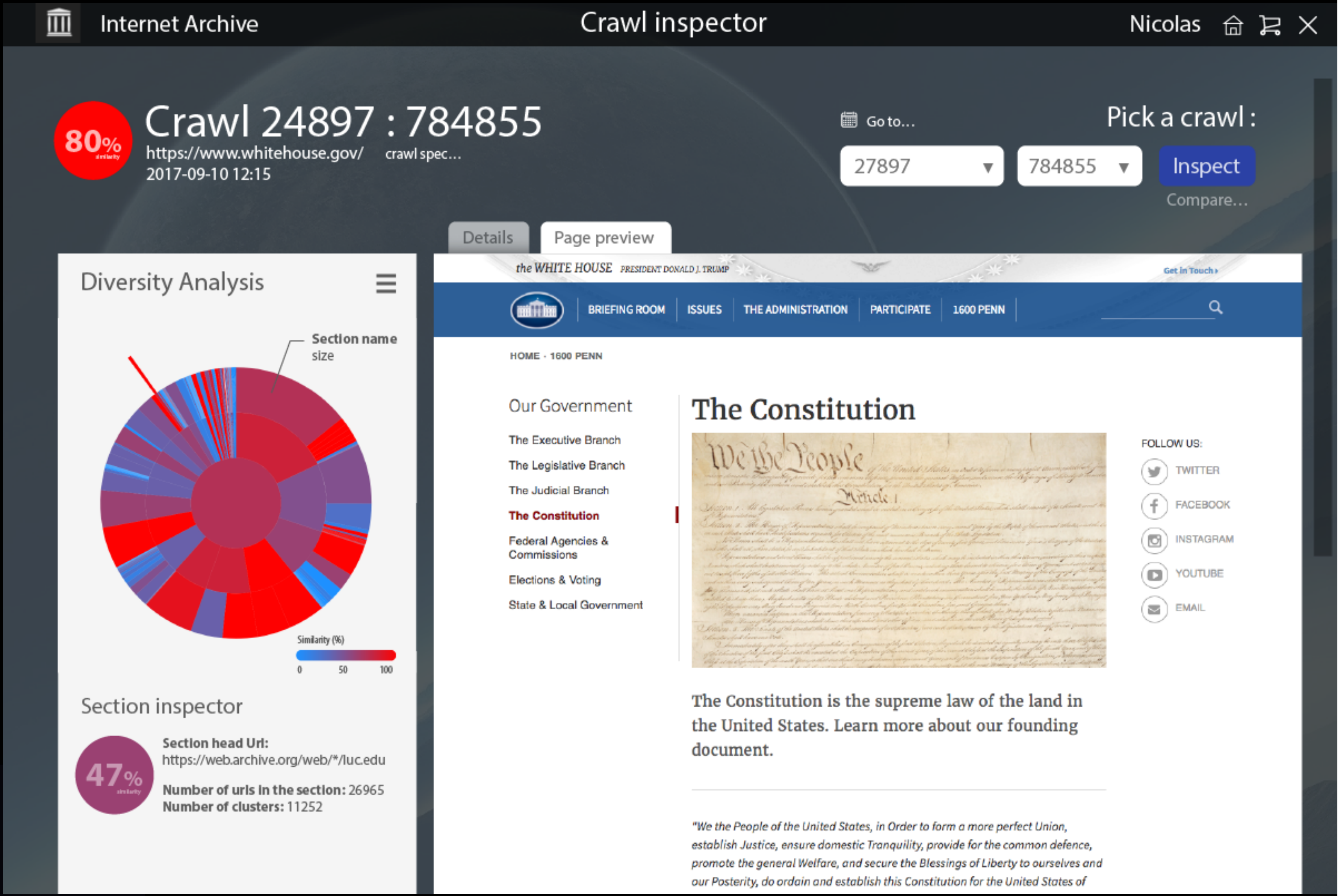
Novas formas de viajar ao passado

Search pages from the p

Millions of contents archived since 1996

Meet the service

Better crawl



Internet Archive

Crawl inspector

Nicolas

80% similarity

Crawl 24897 : 784855

https://www.whitehouse.gov/ crawl spec...

2017-09-10 12:15

Go to...

Pick a crawl :

27897 ▾ 784855 ▾ Inspect

Compare...

Details | Page preview

Diversity Analysis

Section inspector

47% similarity

Section head Url: https://web.archive.org/web/*/luc.edu

Number of urls in the section: 26965
Number of clusters: 11252

the WHITE HOUSE PRESIDENT DONALD J. TRUMP

Get in Touch ›

BRIEFING ROOM | ISSUES | THE ADMINISTRATION | PARTICIPATE | 1600 PENN

HOME · 1600 PENN

Our Government

- The Executive Branch
- The Legislative Branch
- The Judicial Branch
- The Constitution
- Federal Agencies & Commissions
- Elections & Voting
- State & Local Government

The Constitution

FOLLOW US:

- TWITTER
- FACEBOOK
- INSTAGRAM
- YOUTUBE
- EMAIL

The Constitution is the supreme law of the land in the United States. Learn more about our founding document.

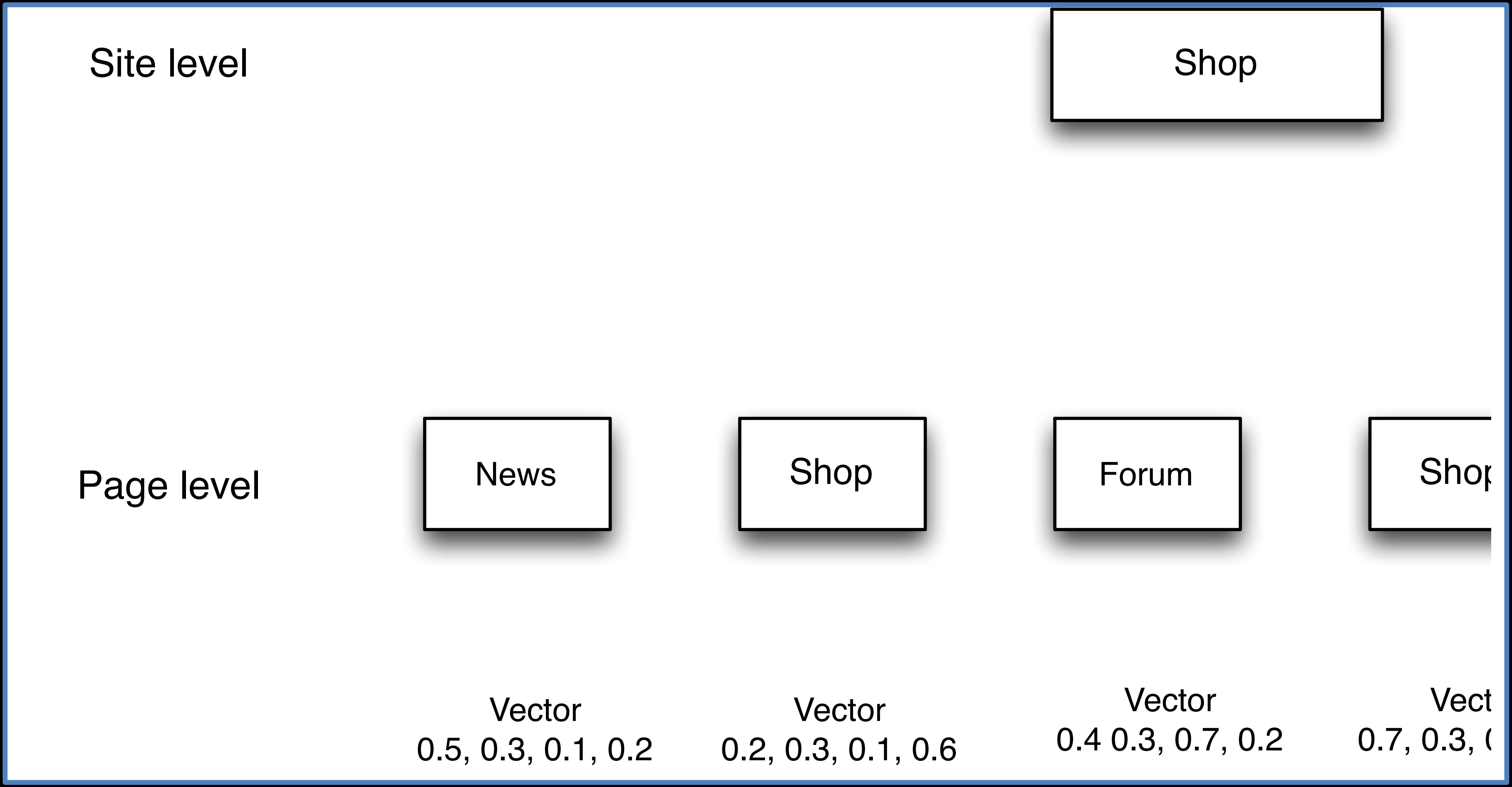
"We the People of the United States, in Order to form a more perfect Union, establish Justice, ensure domestic Tranquility, provide for the common defence, promote the general Welfare, and secure the Blessings of Liberty to ourselves and our Posterity, do ordain and establish this Constitution for the United States of

IMR INTERNET MEMORY RESEARCH

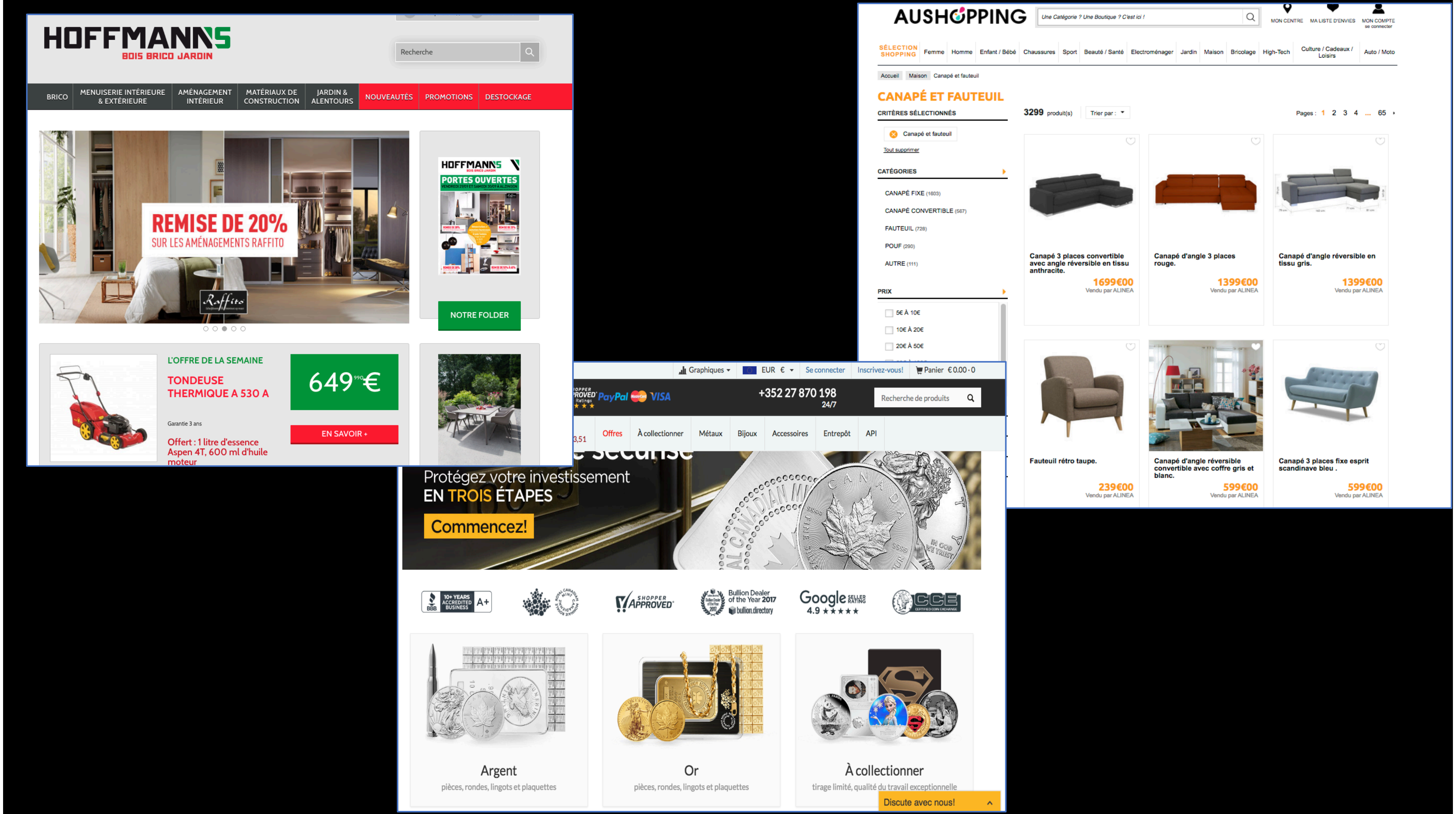
Genre analysis

- Create model, language independent for publication genres:
  - News/e-commerce/Blogs/Forum/Porn/Company/
  - Detect webspam
- Evaluate how this can be use for large scale crawling priorization
  - Change revisit frequency depending on genre
  - Change crawling budget depending on genre
  - Blacklist spam

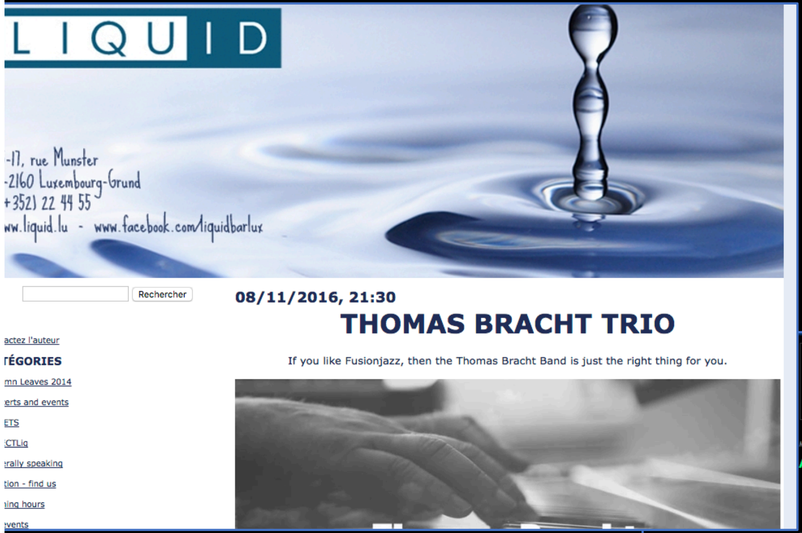
Learning to detect genres



e-commerce sites



Blog sites



L I Q U I D

-17, rue Munster
-2160 Luxembourg-Grund
+352) 22 44 55
ww.liquid.lu - www.facebook.com/liquidbarlux

Rechercher

actez l'auteur

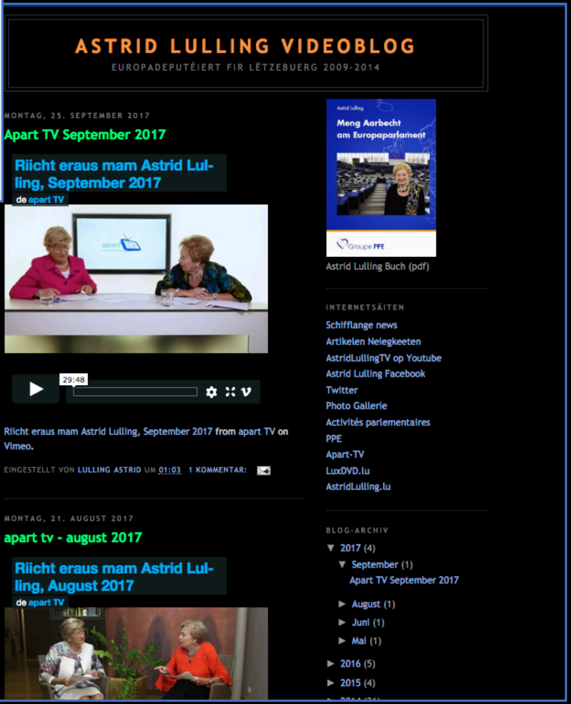
TÉGORIES

mn Leaves 2014
erts and events
ETS
CTLiq
rally speaking
tion - find us
ing hours
vents

08/11/2016, 21:30

THOMAS BRACHT TRIO

If you like Fusionjazz, then the Thomas Bracht Band is just the right thing for you.



ASTRID LULLING VIDEOBLOG

EUROPADEPUTÉIERT FIR LËTZEBUERG 2009-2014

MONTAG, 25. SEPTEMBER 2017

Apart TV September 2017

Riicht eraus mam Astrid Lulling, September 2017

de apart TV

29:48

Riicht eraus mam Astrid Lulling, September 2017 from apart TV on Vimeo.

EINGESTELLT VON LULLING ASTRID UM 01:03 1 KOMMENTAR:

MONTAG, 21. AUGUST 2017

apart tv - august 2017

Riicht eraus mam Astrid Lulling, August 2017

de apart TV

Astrid Lulling
Meng Aarbecht am Europaparlament
Groupe PPE

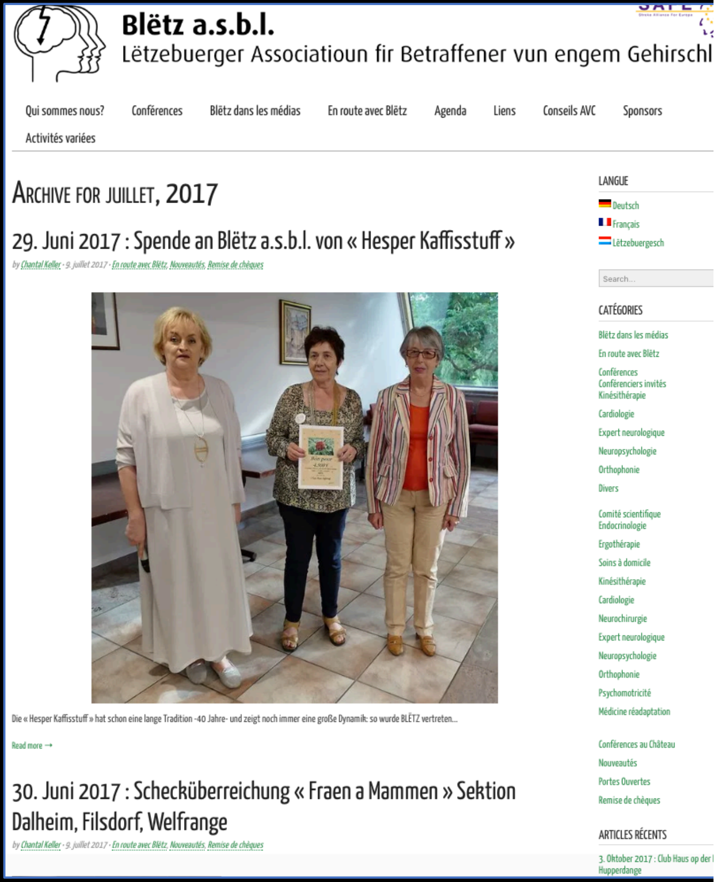
Astrid Lulling Buch (pdf)

INTERNETSÄITEN

Schifflange news
Artikelen Neiegkeeten
AstridLullingTV op Youtube
Astrid Lulling Facebook
Twitter
Photo Gallerie
Activités parlementaires
PPE
Apart-TV
LuxDVD.lu
AstridLulling.lu

BLOG-ARCHIV

- ▼ 2017 (4)
  - ▼ September (1)
    - Apart TV September 2017
  - ► August (1)
  - ► Juni (1)
  - ► Mai (1)
- ► 2016 (5)
- ► 2015 (4)



Blëtz a.s.b.l.

Lëtzebuerger Associatioun fir Betraffener vun engem Gehirschl

Qui sommes nous? Conférences Blëtz dans les médias En route avec Blëtz Agenda Liens Conseils AVC Sponsors Activités variées

ARCHIVE FOR JUILLET, 2017

29. Juni 2017 : Spende an Blëtz a.s.b.l. von « Hesper Kaffisstuff »

by Chantal Keller · 9. juillet 2017 · En route avec Blëtz, Nouveautés, Remise de chèques

Die « Hesper Kaffisstuff » hat schon eine lange Tradition -40 Jahre- und zeigt noch immer eine große Dynamik: so wurde BLËTZ vertreten...

Read more →

30. Juni 2017 : Schecküberreichung « Fraen a Mammen » Sektion Dalheim, Filsdorf, Welfrange

by Chantal Keller · 9. juillet 2017 · En route avec Blëtz, Nouveautés, Remise de chèques

LANGUE

Deutsch
Français
Lëtzebuergesch

Search...

CATÉGORIES

Blëtz dans les médias
En route avec Blëtz
Conférences
Conférenciers invités
Kinésithérapie
Cardiologie
Expert neurologique
Neuropsychologie
Orthophonie
Divers
Comité scientifique
Endocrinologie
Ergothérapie
Soins à domicile
Kinésithérapie
Cardiologie
Neurochirurgie
Expert neurologique
Neuropsychologie
Orthophonie
Psychomotricité
Médicine réadaptation
Conférences au Château
Nouveautés
Portes Ouvertes
Remise de chèques

ARTICLES RÉCENTS

3. Oktober 2017 : Club Haus op der
Hupperdange

Obligado !

Julien Masanès

Internet Memory (Research / Foundation)
j@internetmemory.net
@julienmasanes