BSB Bayerische StaatsBibliothek
Information in erster Linie

Twittervane

Projektbericht

MDZ Münchener DigitalisierungsZentrum Digitale Bibliothek

Grundzüge des Projekts

- finanziert von IIPC (International Internet Preservation Consortium)
- Projekt
  - Prototyp-Entwicklung von Twittervane
  - Weiterentwicklung und Evaluierung von Twittervane
- durchgeführt durch die BL
- Projektstatus: abgeschlossen
- Prototyp
- Open-Source, verfügbar auf github: https://github.com/ukwa/twittervane

Ausgangslage

- Aktuelle Vorgehensweise bei Auswahl von Websites:
  - weitgehend manuell durch einige wenige Experten
  - zeitaufwendig und teuer
  - Man kann kaum auf aktuelle Ereignisse reagieren
  - Auswahl ist subjektiv

Lösung?

- Social Media nutzen, um relevante Websites zu aktuellen Ereignissen zu selektieren

Kurzbeschreibung Twittervane

- Nutzt das Wissen der Menge (Crowd), um Websites für die Langzeitarchivierung zu sammeln
- Datenbasis: Twitter
- Sammeln und extrahieren von URLs, die in Twitternachrichten verbreitet werden, nicht von Twitternachrichten selbst
- Selektionskriterium: Popularität von Websites
- Fördert idealerweise Websites zu Tage, die einem sonst durch die Lappen gegangen wären
- Keine rückwirkenden Suchanfragen möglich

TwitterVane

Crowd sourcing for Web Archiving

Welcome

Selection of web archive is a manual process that relies upon people to select quality sites and nominate or submit them to web archives for inclusion. Past (national and/or international) efforts to automate selection have either failed to convince staff that automation was identifying quality resources, or have focused on providing an interface for selectors to submit sites rather than carry out the selection per se. Selection has thus remained, for most institutions, an element of the workflow wholly dependent on contributions from externals. The number of selectors providing sites is typically small and their contributions are inevitably subjective. The resulting collections, whilst immensely valuable, are therefore mostly representative of the expertly selected sites and do not fully represent the sites frequently used in a more social setting.

Crowd sourcing is an opportunity to develop a new approach to this problem. It taps into the growth of social networks to outsource tasks typically performed by an employee or contractor, to an undefined, large group of people or community (a "crowd") (Wikipedia, 2011). It is a particularly attractive option in the current economic climate, where we are all being asked to 'do more with less'. A number of cultural heritage institutions and/or projects have already begun to leverage the power of the crowd for digitised collections, including the National Library of Australia (through Trove), the Transcribe Bentham project at ULCC, and the National Library of Finland (through the DigitalKoot program). No such projects have yet been launched for web archives.

This project will develop an automated approach to selection for web archiving based on the principles of crowd sourcing. It has been awarded partial funding from the International Internet Preservation Consortium and supports the forthcoming web archiving strategy by increasing the number of selections and automating the selection process.

Links

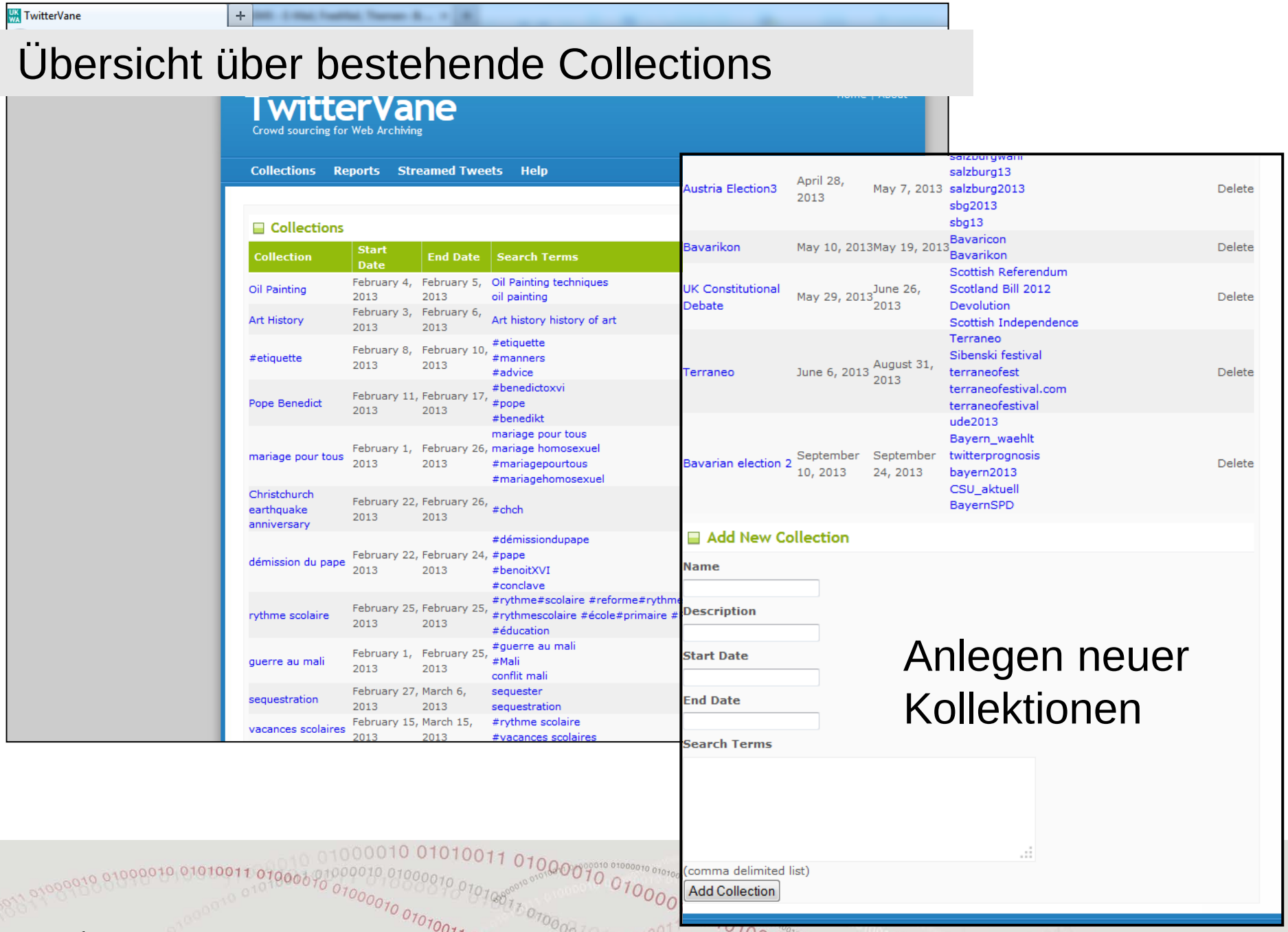
Twitter API

UK Web Archive

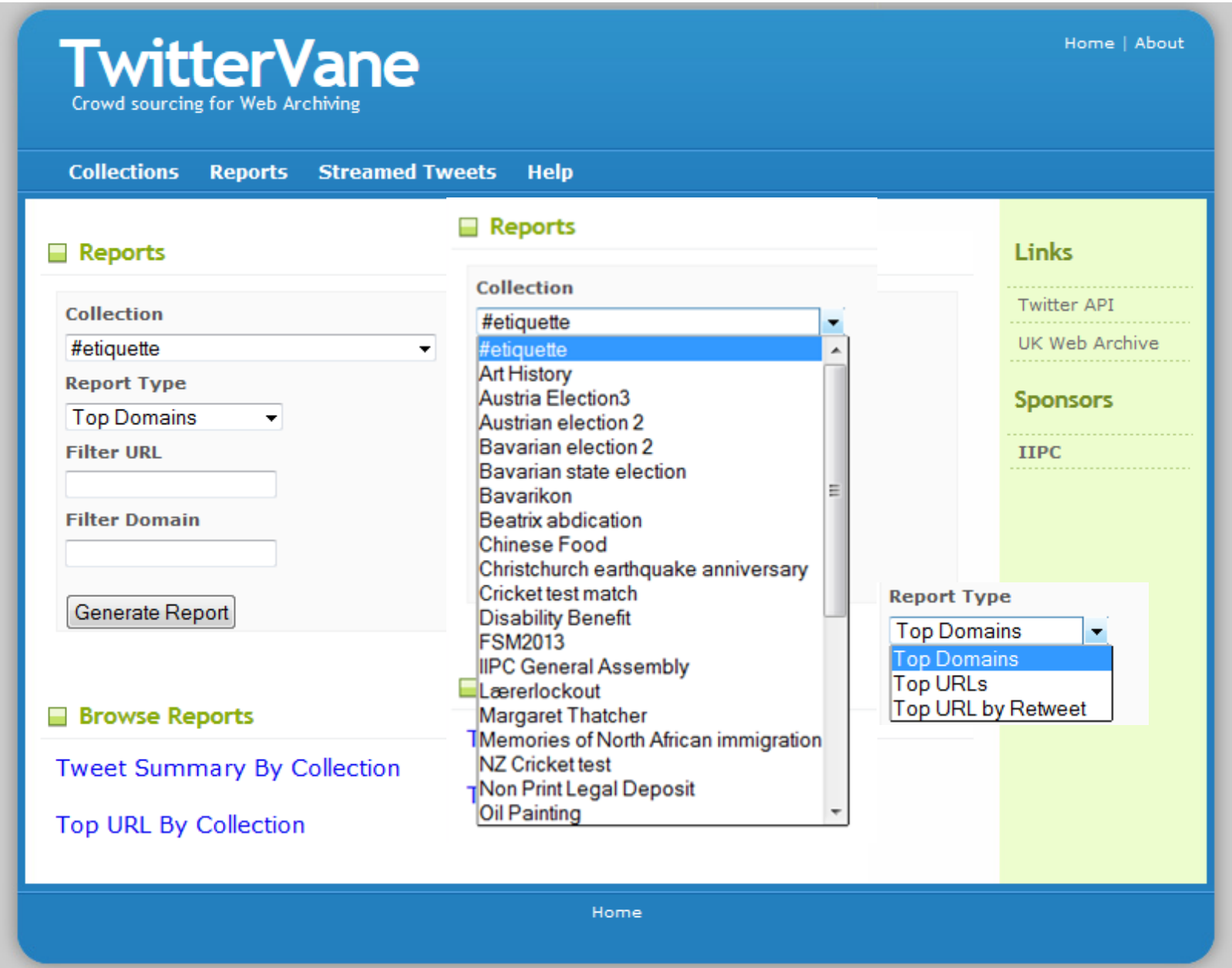
Sponsors

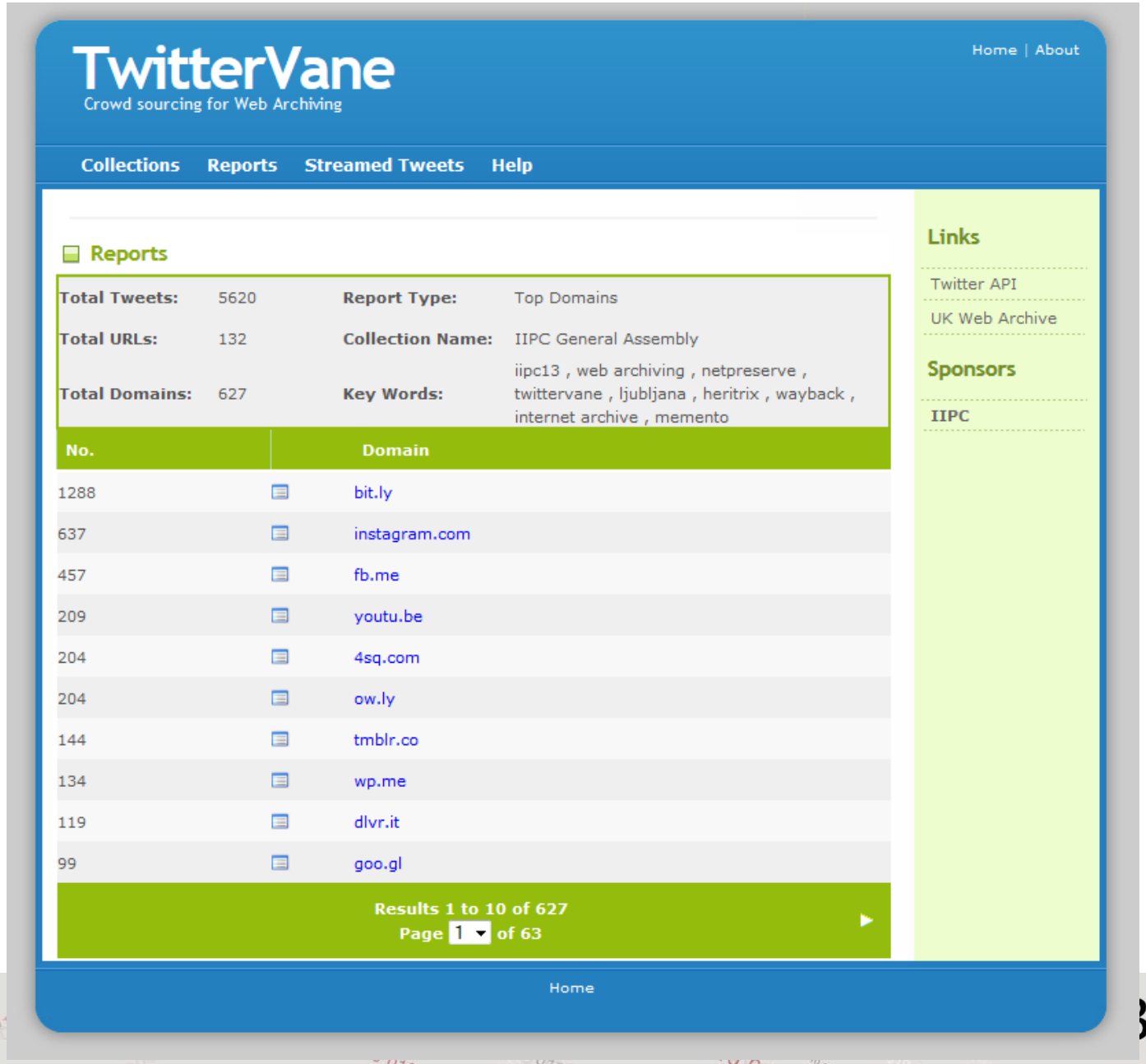
IIPC

Übersicht über bestehende Collections

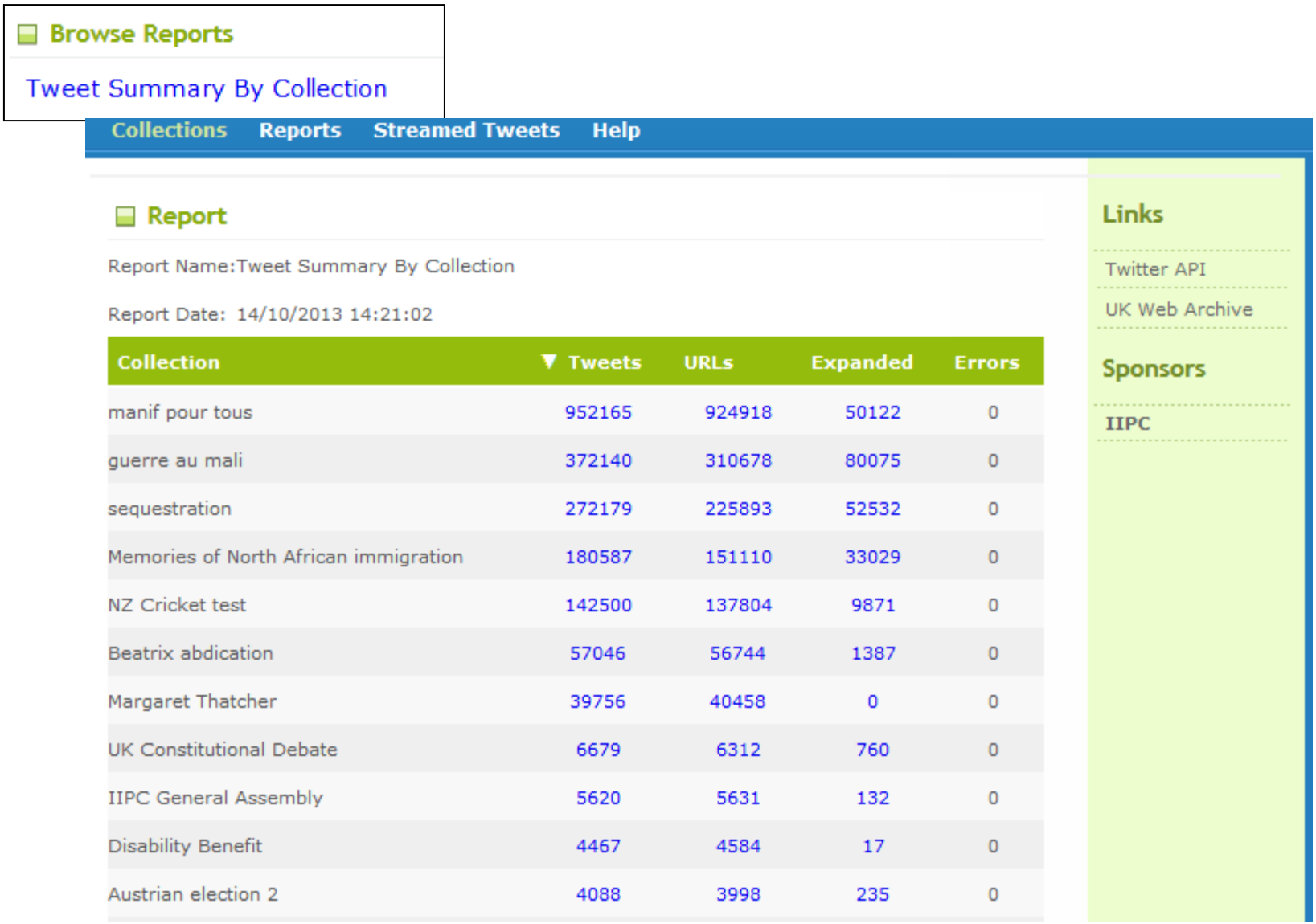


Anlegen neuer Kollektionen


TwitterVane
Crowd sourcing for Web Archiving
Home | About
Collections
Reports
Streamed Tweets
Help
Reports
Collection
#etiquette
Report Type
Top Domains
Filter URL
Filter Domain
Generate Report
Browse Reports
Tweet Summary By Collection
Top URL By Collection
Reports
Collection
#etiquette
#etiquette
Art History
Austria Election3
Austrian election 2
Bavarian election 2
Bavarian state election
Bavarikon
Beatrix abdication
Chinese Food
Christchurch earthquake anniversary
Cricket test match
Disability Benefit
FSM2013
IIPC General Assembly
Lærerlockout
Margaret Thatcher
Memories of North African immigration
NZ Cricket test
Non Print Legal Deposit
Oil Painting
Report Type
Top Domains
Top Domains
Top URLs
Top URL by Retweet
Links
Twitter API
UK Web Archive
Sponsors
IIPC
Home



Report nach Top Domains


Browse Reports
Tweet Summary By Collection
Collections
Reports
Streamed Tweets
Help
Report
Report Name:Tweet Summary By Collection
Report Date: 14/10/2013 14:21:02
Collection
Tweets
URLs
Expanded
Errors
manif pour tous 952165 924918 50122 0
guerre au mali 372140 310678 80075 0
sequestration 272179 225893 52532 0
Memories of North African immigration 180587 151110 33029 0
NZ Cricket test 142500 137804 9871 0
Beatrix abdication 57046 56744 1387 0
Margaret Thatcher 39756 40458 0 0
UK Constitutional Debate 6679 6312 760 0
IIPC General Assembly 5620 5631 132 0
Disability Benefit 4467 4584 17 0
Austrian election 2 4088 3998 235 0
Links
Twitter API
UK Web Archive
Sponsors
IIPC


TwitterVane
Crowd sourcing for Web Archiving
Home | About
Collections Reports Streamed Tweets Help
Report
Report Name: Expanded URLs In Collection
Report Date: 15/10/2013 11:43:44
Collection Name:Christchurch earthquake anniversary
URL Tweeter Tweet
http://tweetedtimes.com/d... Mr. DIY Diy Food And Wine News Daily http://t.co/Zee5ogRZXB - top stories by Bourdain, chefludo, FrenchChefWife
http://www.youtube.com /pl... Henry 22 February Commemoration by Christchurch City Council #Videos http://t.co/dLaNuCOtm5 #eqnz #Chch
http://bit.ly/X46lZu 648 Ancaster Community Food Drive a success - CHCH News: Ancaster Community Food Drive a successCHCH NewsIt was a ... http://t.co/lYbnuLI8oj
http://www.stuff.co.nz/th... Ben Stanley RT @secondzeit: Very nice state of the quake feature by @cdlanderson #Chch http://t.co/NQ9xFyL6ku
http://tinyurl.com/afhtjq... YMCA Ham/Burl /Brant #hamont-Watch @CHCHTV 6 pm News for interview with @ymcahbb employment specialist on employment report out today: http://t.co/WSRE8yw1Bl
http://songsforchristchur... FESTA Free gig in central #chch today 12-6pm at Re:START, supports FESTA - they truly are Songs for Christchurch: http://t.co/S0u1OSh9rz
http://bit.ly/UUExIm Canterbury Forum New Post: Prime Minister Key to account for civil defence failure #NZ #CDEM #EQNZ #Chch #quake #... http://t.co/emvvv9d5pP #christchurch
http://ibnlive.in.com/new... Naresh kamboz RT @Bollywood__News: Vote Aishwarya The Best Actress of 100 Years http://t.co/j3WbR34IaZ @BachchanLover @amit_pc? http://t.co/FG79ZWAah1 -
http://goo.gl/nYeb Naresh kamboz RT @Bollywood__News: Vote Aishwarya The Best Actress of 100 Years http://t.co/j3WbR34IaZ @BachchanLover @amit_pc? http://t.co/FG79ZWAah1 -
http://ibnlive.in.com/new... Aish fan RT @Bollywood__News: Vote Aishwarya The Best Actress of 100 Years http://t.co/j3WbR34IaZ @BachchanLover @amit_pc? http://t.co/FG79ZWAah1 -
Results 1 to 10 of 40
Page 1 of 5
Links
Twitter API
UK Web Archive
Sponsors
IIPC
Home

Nutzen

- Event-basiertes Harvesting, z.B. Wahlen, aktuelle Ereignisse
- Schnelle Reaktionszeit
- Reduziert ggf. Zeitaufwand, um archivierungswürdige Websites zu finden
- Ergänzt die manuelle Auswahl durch Experten
- Sammeln von Websites, die stark rezipiert werden

Fragen / Schwierigkeiten

- Viele URLs zu Zeitungsartikeln und Online-Zeitschriften, wenige komplette Websites zu einem Thema
- Nur ca. 20-30% der URLs relevant
- Spam
- Lohnt der Aufwand (anlegen von collections, Auswahl von Suchbegriffen, Selektion der URL-Liste) für relativ wenige relevante Websites?
- Was sind geeignete Suchbegriffe?

Fazit

- Ersetzt nicht den Auswahlprozess, aber kann als zusätzliches / komplementäres Tool zur Auswahl von Websites dienen
- Besonders geeignet für Event-Harvesting
- Optimierungspotential vorhanden, z.B. Verbesserung der Ergebnisse durch automatischen Entfernen von Spam-Websites und Duplikate

Quellen

- Twittervane:
  http://www.webarchive.org.uk/twittervane/
- Project Final Report:
  http://netpreserve.org/sites/default/files/resources/ProjectFinalReport_Twittervane_Approved.pdf
- User Manual:
  http://netpreserve.org/sites/default/files/resources/TwitterVane%20User%20Manual%20v1.1.doc
- Administrators Guide:
  http://netpreserve.org/sites/default/files/resources/TwitterVane%20Administrators%20Guide%20v1.0.doc
- System Installation Guide:
  http://netpreserve.org/sites/default/files/resources/TwitterVane%20System%20Installation%20Guide%20v1.0.doc
- https://github.com/ukwa/twittervane

Vielen Dank für Ihre Aufmerksamkeit!