

	Web Archiving	Project Name: Evaluating Twittervane	Date: 16 June 2013
		Document Title: Project Final Report	Version: Approved

[WAP035] [Evaluating Twittervane] Project Final Report

Web Archiving Mission Statement

The Web Archiving team collects, makes accessible and preserves web resources of scholarly and cultural importance from the UK domain.

Our mission is to:

- Implement non-print Legal Deposit by carrying out domain level and complementary crawls of UK websites.
- Develop multiple access routes to the web archive based on stakeholders' needs.
- Enable curation and ingest of archived websites for long term preservation.
- Ensure ongoing capability of archiving the evolving web.

Project Information

Senior Responsible Owner	Mary Pitt
Project Manager	Helen Hockx-Yu
Senior User	LoC, BNF, NLNZ
Senior Supplier	Andy Jackson
Project page location on wiki	https://intranet.bl.uk:8443/confluence/display/WAG/WAP035+Evaluating+Twittervane

Document History

Version	Date	Author	Status / change
0.1	18 Feb. 2013	Helen Hockx-Yu	Draft
1.0	16 Jun. 2013	Mary Pitt	Approved
Approved by Project SRO	Name Mary Pitt	Date 16 th June 2013	

	Web Archiving	Project Name: Evaluating Twittervane	Date: 16 June 2013
		Document Title: Project Final Report	Version: Approved

Table of Contents

<i>1. Executive summary</i>	3
1. Background.....	4
2. Aims and objectives	4
3. Project approach	4
4. Deliverables	6
5. Outcomes	6
5.1 Improvement of Twittervane	6
5.2 Evaluation by curators	7
6. Risk management	7
7. Project budget.....	7
8. Issues & lessons learned.....	8
9. Conclusions & recommendations.....	8
Appendix Evaluation Reports	9

	Web Archiving	Project Name: Evaluating Twittervane	Date: 16 June 2013
		Document Title: Project Final Report	Version: Approved

1. Executive summary

The Evaluating Twittervane project is funded by the International Internet Preservation Consortium (IIPC) to build on an earlier project, Twittervane. Twittervane is a prototype application capable of collecting and analysing Twitter feeds and outputs URLs mentioned in the Tweets. These URLs shared on the Twitter could potentially point to web resources relevant to web archive collections.

The main purpose of this project is to improve the prototype delivered by the previous project and evaluate the application by a wider range of curators independently to assess the validity of the Twittervane approach.

The planned development work were successfully carried out which improved the Twittervane prototype in many ways so that it could be deployed as a web service for the curators to evaluate. The source code and documentation of the Twittervane can be found in the Github repository as an open source project.

Curators from three National Libraries explored and tested the application and provided very useful feedback. Some of the feedback, where possible within the project's resource, was addressed while others have been logged as future requirements. Most curators taking part in the evaluation are positive about the Twittervane approach and see this as a complementary selection tool, especially for events-based collections.

The project was on time and budget, delivered all the high-level deliverables and met the acceptance criteria defined in the project proposal.

Twittervane is not a replacement of the curatorial process but has the potential to be a complementary tool, which may only be useful for events-based collections.

Further work need to take place to productionise Twittervane. However the question that needs to be answered first is whether the amount of processing that is required to produce the relative small amount of relevant URLs can be justified.

	Web Archiving	Project Name: Evaluating Twittervane	Date: 16 June 2013
		Document Title: Project Final Report	Version: Approved

1. Background

The International Internet Preservation Consortium (IIPC) funded the Twittervane project (WAP029) in 2012 for the British Library to develop a prototype application which is capable of monitoring and analysing Twitter traffic relevant to a given theme and generate a list of most frequently shared web resources. These websites can then be presented to curators as potential titles for web archiving, saving time and effort required for manual selection. WAP029 developed a prototype and piloted it at the British Library as an internal service to select additional content for the Diamond Jubilee special collection.

The Evaluating Twittervane project (WAP035) is a follow-on project of WAP029, also funded by the IIPC, to improve the prototype and make it usable for evaluation by member institutions of the IIPC.

2. Aims and objectives

The primary goal of the project is to evaluate the Twittervane prototype. It includes two strands of work: development work to improve the prototype by addressing some of the known issues and evaluation of the application by curators to assess the validity of the Twittervane approach.

The project aims to deliver the following high level deliverables:

1. A evaluation version of Twittervane application including improved usability and documentation covering installation and basic usage
2. Deploy Twittervane as a web service to enable evaluation
3. Evaluation of Twittervane by curators of three IIPC institutions
4. Final project report including the outcome of the evaluation and recommendations.

3. Project approach

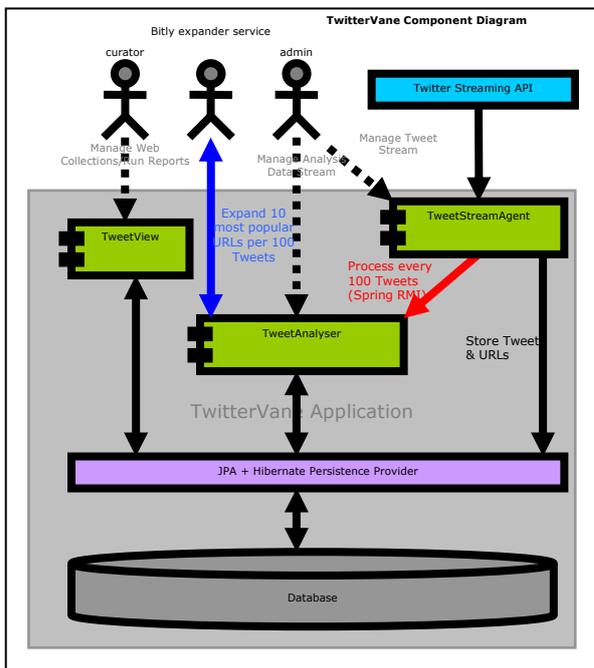
The project improved the Twittervane prototype by implementing the following changes:

Tasks	Details	Completion date
Better code management	Code and issues merged and made available into Github code base	03/01/2013
Debugging	Added analysis for every n-tweets (analysis is run after n-tweets are received - configured in the spring-servlet.xml file for the TweetStreamAgent component	22/01/2013
Process improvement	Service is deployed as 3 components: TweetView (Curator's UI), TweetStreamAgent and TweetAnalyser. Each component can be deployed to a separate Tomcat instance. Added URL expansion for the top 10 tweets (configured in the	23/01/2013

BRITISH LIBRARY	Web Archiving	Project Name: Evaluating TwitterVane	Date: 16 June 2013
		Document Title: Project Final Report	Version: Approved

	spring-servlet.xml file for the TweetAnalyser component).	
User interface improvement	New reports 'Tweet Summary By Collection' and 'Top URLs By Collection' added and validated against data collected.	25/01/2013
	New report 'URLs In Collection' added.	29/01/2013
	New report "Tweet Summary by Date" added	30/01/2013
Tweets in JSON files to provide better data availability or processability	Completed. Also added application level logging (configured in the spring-servlet.xml file for the TweetStreamAgent component).	04/01/2013

The diagram below shows the structural relationships between the components of TwitterVane.



TwitterVane has 3 service components:

1. The TweetView component provides the management and reporting features that curators use to create and report on Web Collections.
2. The TweetStreamAgent provides the UI and services for managing the inbound Tweet data from Twitter.
3. The TweetAnalyser performs URL expansion on shortened URLs associated with a Tweet, resolves a Tweet to a web collection, and manages the storage of Tweets.

	Web Archiving	Project Name: Evaluating Twittervane	Date: 16 June 2013
		Document Title: Project Final Report	Version: Approved

Curators of the National Library of New Zealand, the National Library of France and the Library of Congress evaluated the Twittervane. Some of the feedback by the curators has been implemented and a second evaluation version was deployed during the project.

4. Deliverables

Deliverable	Planned delivery date	Actual delivery date	URL (if applicable)	Comment
Management products (<i>project management documents: eg project plan, PID</i>)				
Project proposal	24/09/2012	24/09/2012		
Project plan			View document	
Project final report	28/02/2013	28/02/2013		
Specialist products (<i>those produced and delivered by the project</i>)				
Evaluation version	30/01/2013	29/01/2013	Twittervane	
Source code	28/02/2013	28/02/2013	https://github.com/ukwa/twittervane	
Documentation	28/02/2013	28/02/2013	Including System Installation Guide and User Manual	
Evaluation & report	31/01/2013 – 28/02/2013	31/01/2013 – 28/02/2013		

5. Outcomes

We are confident that the project has met the acceptance criteria which were proposed in the project proposal:

- The application is of sufficient quality in that it has the required functionality, is reliable, usable, efficient and easy to maintain.
- The application is tested with real users and properly documented.
- The methodology has been evaluated independently, by the three IIPC members mentioned above, and the results of this evaluation made publically available.
- The implementation is available as open source.

5.1 Improvement of Twittervane

The WAP035 project successfully carried out the planned development work and improved the Twittervane prototype so that it could be deployed as a web service for the curators to evaluate.

The most significant improvements are the data accessibility and scalability of the application, which are achieved by implementing the following:

1. Instead of running the entire application under one JVM as in the prototype, Twittervane now separates three distinct functions and implements them as three service components, improving scalability and making it flexible to deploy, depending on available machine resources.

BRITISH LIBRARY	Web Archiving	Project Name: Evaluating TwitterVane	Date: 16 June 2013
		Document Title: Project Final Report	Version: Approved

2. Data is stored in a stable database system (ie Postgres) which can be accessed easily (by system administrator).
3. Batch processing is implemented as part of the TweetAnalyser, which prevents the JVM from running out of memory when processing a large number of tweets.
4. Most of the reports are based on summary data. While analysing Tweets, the TweetAnalyser stores a set of summary data which are used for the report in the TweetView. This avoids generating reports on reading the full database.

TwitterVane is documented by a User manual and a System Installation Guide. And the source code is managed in the Github repository as an open source project.

5.2 Evaluation by curators

A common template was used for the evaluation. It contains a description of the main components of TwitterVane, a set of questions, and notes explaining the decisions / considerations which impact the reports produced by TwitterVane. The curators were asked provide guidance about these and help us understand their requirements. The template was returned with comments which summarise the curators' assessment and observations.

6 curators of the National Library of New Zealand, the National Library of France and the Library of Congress independently evaluated the TwitterVane methodology and provided their feedback. Curators had 3 weeks to use and test TwitterVane. They not only provided valuable feedback on the user interface and documentation, but also set up collections and assessed the relevance of the URLs reported by TwitterVane for their collections. Some feedback, where possible within the project's resource, was addressed while others have been logged as future requirements.

The general view is that TwitterVane could be useful for events-based collections, as it could reduce the time spent on web searching especially over a longer period of time (eg elections, Olympics). URLs reported by TwitterVane tend to point to news sites and online periodicals. Curators also found that only a small percentage of the URLs found by TwitterVane are relevant and can be accepted as valid selections (eg 20% ~ 30%). Some URLs lead to spam sites.

A workshop on TwitterVane has been proposed to the IIPC 2013 General Assembly programme committee. The outcome of the project will also be reported to the wider IIPC membership.

6. Risk management

For the TwitterVane to become a tool that curators use as part of their daily selection workflow, further work needs to be done to gather requirement and develop the application. There is a risk of this not taking place once the project has ended.

7. Project budget

The planned resource for the project includes one full FTE contractor developer and a part-time project manager as specified in the table below. The British Library is contributing the project management effort without requesting IIPC funding.

BRITISH LIBRARY	Web Archiving	Project Name: Evaluating Twittervane	Date: 16 June 2013
		Document Title: Project Final Report	Version: Approved

Resource	Costs	Notes
Full time developer	£450 x 40 days = £18,000	
Project manager – institutional contribution from the British Library	£400 x 8 days = £3,200	1 day a week including the time to produce project documents and reports
Total	£18,000	23,400 Euro

The contractor resource has been fully spent and was on target. There is a slight underspent of the project manager's time due to conflict of other commitment: 6 instead of 8 days has been spent. This however is compensated by technical effort by the British Library in providing technical guidance and infrastructural support to the project.

8. Issues & lessons learned

One curator pointed out that search terms are closely related to and impact the quality of the results produced by Twittervane. Unfortunately the project team wasn't much more experienced than the curators to provide more useful hints. Basic training including best practice about the use of search terms to obtain the most relevant tweets, seems an helpful area of future work.

The relevance and quality of the URLs expanded by Twittervane seem to raise the question whether they can justify the amount of processing required to produce the URLs. This may not only be related to the search terms used, but also to the nature of social networks like Twitter, that this approach may only be useful for very specific collections.

It could be that more extensive testing is required by curators over a much longer period of time which will enable them to become more skilled in using Twittervane and consequently reduce the noise in the results.

9. Conclusions & recommendations

The project improved the Twittervane prototype and made it available for curators to evaluate. Most curators who took part in the evaluation were positive about the Twittervane approach and saw this as a complementary selection tool, especially for events-based collections. However, Twittervane also points to a large number of URLs which are not relevant to the collections and cannot be used as valid selections. At times, they even point to spam sites and duplicates. This may be improved when curators are more skilled and establish best practice in using the most appropriate search terms for a collection. More testing is required over longer period of time to determine this. The issues related to data quality may also be addressed technically by for example removing duplicates and detecting spam sites but further investigations are required to achieve this.

Twittervane is not a replacement of the curatorial process but has the potential to be a complementary tool, which may only be useful for events-based collections.

Further work need to take place to productionise Twittervane. However the question that needs to be answered first is whether the amount of processing required to produce the small amount of relevant URLs can be justified.

BRITISH LIBRARY HISTORICAL	Web Archiving	Project Name: Evaluating Twittervane	Date: 16 June 2013
		Document Title: Project Final Report	Version: Approved

Appendix Evaluation Reports

Report 1 by the National Library of New Zealand

Curatorial

- Are the URLs found and expanded by Twittervane relevant to your collection?
Collections used were EQNZ and Sevens The URLs were relevant to the collection
- How many of the URLs found by Twittervane would you accept as a selection (for web archiving)
None
- Do URLs found by Twittervane lead to spam sites?
Mine were fine
- Would you have selected the URLs found by Twittervane if you were doing manual selection? **No**
- Do Twittervane URLs point to certain types of websites?
They tend to point to news sites while we're more interested in complete websites on a particular topic. Useful though for events once the newspaper site uses a consistent URL on a particular topic on their site so that those pages can be harvested.
- Is Twittervane useful? How does it aid / hinder selection in your view?

One problem is simply figuring out which are the best search terms/ hash tags to use in the first place to get the best search result. Some useful hints might be helpful. The Trends on Twitter are quite helpful but very limited.

Usability

- Is Twittervane easy to use? Is the UI intuitive? **Yes**
- How do you like the layout of the page? **Fine**
- What would you like to change / add to the UI?
A print option that allows you to print the URLs in the collection report
- What additional reports would you like to see: eg tweets grouped by tweeter – is that something you regard useful? **Could be if there are a lot of tweets.**
The top URL report didn't work. Top URL by collection report didn't work well for me either. **[I got to the URL list by looking at the Tweet summary by collection report](#)**
- Would you prefer just to see your own collections?
I'd like the option to "see all" as well as limit the view to my own collections. Seeing other people's collections can be useful.

System performance

- Is the response time quick enough? **Yes**

3. Notes

- You will come across "Unknown" in the reports: these are tweets which Twittervane cannot associate or assign to any collection – we could hide these.
- The processing of tweets are done in batches, set to 100 tweets currently – this is configurable.
- Twittervane also optimises URL expansions. Only the top 10 "most popular" (ie: the most frequently appearing) URLs for every processing run (set to 100 tweets). Again this could be changed.
The most popular sites are often news sites so some of the sites we're interested in might be further down the hit list, so haven't the ability to access the list would be helpful.
- The "tweets summary by collection" report
<http://194.66.239.180:8080/twittervane/reportView.html?report=tweetSummaryByCollection&sort=desc>

	Web Archiving	Project Name: Evaluating Twittervane	Date: 16 June 2013
		Document Title: Project Final Report	Version: Approved

contain both processed and unprocessed tweets – would like to see them separately? *I don't understand the difference between the two. Please explain!*

- Twittervane also has a processing and admin UI which we will evaluate separately.

Report 2 by the National Library of New Zealand

Curatorial

- Are the URLs found and expanded by Twittervane relevant to your collection?
For the most part (collections used were Novopay and Sir Paul Holmes)
- How many of the URLs found by Twittervane would you accept as a selection (for web archiving)
Around 20/30%. There were a lot of duplicate URLs that came up from different Twitter users.
- Do URLs found by Twittervane lead to spam sites?
I encountered a couple of spam sites. More of a concern was the amount of sites that weren't relevant. For our Sir Paul Holmes collection 25/93 sites were not relevant to that collection which seems really high. For the Novopay collection only 1 out of 29 wasn't relevant.
- Would you have selected the URLs found by Twittervane if you were doing manual selection?
Yes
- Do Twittervane URLs point to certain types of websites?
Yes, mostly news sites
- Is Twittervane useful? How does it aid / hinder selection in your view?
I think it could be useful as it could reduce the amount of web searching we do around event harvests especially over a longer time period e.g. elections/Olympics

Usability

- Is Twittervane easy to use? Is the UI intuitive? *Yes*
- How do you like the layout of the page? *It's fine*
- What would you like to change / add to the UI? *I'd like the ability to delete search terms in a collection. I'd like in the URLs by Collection report to be able to configure how many were seen at one time (currently only 10).*
- What additional reports would you like to see: eg tweets grouped by tweeter – is that something you regard useful? *Only the top domains report seems to work.*
- Would you prefer just to see your own collections? *The option would be good*

System performance

- Is the response time quick enough? *Yes.*

3. Notes

- You will come across "Unknown" in the reports: these are tweets which Twittervane cannot associate or assign to any collection – we could hide these.
- The processing of tweets are done in batches, set to 100 tweets currently – this is configurable.
- Twittervane also optimises URL expansions. Only the top 10 "most popular" (ie: the most frequently appearing) URLs for every processing run (set to 100 tweets). Again this could be changed.
- The "tweets summary by collection" report
<http://194.66.239.180:8080/twittervane/reportView.html?report=tweetSummaryByCollection&sort=desc>
contain both processed and unprocessed tweets – would like to see them separately?
- Twittervane also has a processing and admin UI which we will evaluate separately.

BRITISH LIBRARY	Web Archiving	Project Name: Evaluating Twittervane	Date: 16 June 2013
		Document Title: Project Final Report	Version: Approved

Report by the National Library of France

For two collections: *marriage pour tous* and *demission du pape*)

Curatorial

- Are the URLs found and expanded by Twittervane relevant to your collection?

Mariagepour tous: not all but some of them yes. They are relevant as they are in French most of the time and concern the chosen subject.

Démission du pape: I have found expanded URLs (518) but no relevant ones.

- How many of the URLs found by Twittervane would you accept as a selection (for web archiving)
Mariagepour tous: from the total, those in TLD .fr are already selected easily by BnF. About 6 for a total of 317 found urls could be selected:

<http://www.sourds.net>

<http://yagg.com>

<http://paritedanslemariage.com>

<http://infos-lgbt.centerblog.net>

http://lesalonbeige.blogs.com/my_weblog/

<http://www.immigrationjetable.org>

Démission du pape: None of them could be accepted because they are in other languages.

- Do URLs found by Twittervane lead to spam sites?
Mariagepour tous: No.
Démission du pape: No.
- Would you have selected the URLs found by Twittervane if you were doing manual selection?
Mariagepour tous: some of them yes, some of them no, about half of the relevant ones.
Démission du pape: No. Or some of them if i want an English language collection
- Do Twittervane URLs point to certain types of websites?
Mariagepour tous: yes, mostly online periodicals (Le monde, le Figaro, la Croix, Libération...) and magazines, those are not useful for my collection. And one same URL can be pointed out many times (around 50 over 317).
Démission du pape: Yes, online newspaper. And the same URLs come back any time (The independent, enenews, Daily news...).
- Is Twittervane useful? How does it aid / hinder selection in your view?
Mariagepour tous: yes, it helps but not much. Many urls found for very few relevant to my collection.

Usability

- Is Twittervane easy to use? Is the UI intuitive?
Yes
- How do you like the layout of the page?
It is OK
- What would you like to change / add to the UI?
About search terms, explanations for the choice and how to write them to obtain better results (with or without #). About "Tweet Summary by Collection", it is not intuitive to select "URL" to find the tweets details. Streamed tweets are not very useful as there are already enough results on the other screens.

BRITISH LIBRARY	Web Archiving	Project Name: Evaluating TwitterVane	Date: 16 June 2013
		Document Title: Project Final Report	Version: Approved

If possible, it could be better to add a tab language to guide collections and avoid the confusion of search terms (e.g: when I search of French term “pape” tends to bring URLs about the English term “paper”)

- What additional reports would you like to see: eg tweets grouped by tweeter – is that something you regard useful?

After selecting the report type, the titles are present but there is no number for “total tweets” and “total domains”

- Would you prefer just to see your own collections?
No

System performance

- Is the response time quick enough? Yes

3. Notes

- You will come across “Unknown” in the reports: these are tweets which TwitterVane cannot associate or assign to any collection – we could hide these.
- The processing of tweets are done in batches, set to 100 tweets currently – this is configurable.
- TwitterVane also optimises URL expansions. Only the top 10 “most popular” (ie: the most frequently appearing) URLs for every processing run (set to 100 tweets). Again this could be changed.
- The “tweets summary by collection” report
<http://194.66.239.180:8080/twittervane/reportView.html?report=tweetSummaryByCollection&sort=desc>
contain both processed and unprocessed tweets – would like to see them separately?
- TwitterVane also has a processing and admin UI which we will evaluate separately.

BRITISH LIBRARY	Web Archiving	Project Name: Evaluating TwitterVane	Date: 16 June 2013
		Document Title: Project Final Report	Version: Approved

Report by the Library of Congress

Curatorial

- Are the URLs found and expanded by TwitterVane relevant to your collection?
 - I created a collection for tweets related to the 2013 U.S. budget sequestration debate. Multiple requests of the *Top URLs* report for this collection produced only 1 page (of 8 pages) of results, with the pagination navigation missing. Of the 7 links listed there, 3 were spam sites, 1 was a 404, 1 was an unshortened URL by an unknown provider, 1 was the Google News home page, and 1 was a URL relevant to the collection. The *Top URLs by Retweet* collection displayed no URLs.
 - I looked at some of the other collections to see if mine was an outlier. The “guerre au mali” collection had many of the same URLs listed in the *Top URLs* report as in the “sequestration” collection, which suggests that spam is a major factor. The pagination worked, but none of the top 50 URLs was clearly related to the topic and the number of tweets pointing to any URL listed below this point (5 tweets) didn’t seem to indicate a clear convergence of Twitter users’ interest toward specific resources. The *Top URLs by Retweet* report is also empty for the “guerre au mali” collection.
- How many of the URLs found by TwitterVane would you accept as a selection (for web archiving)
 - We might accept the 1 relevant URL indicated in the *Top URLs* report for the “sequestration” collection.
- Do URLs found by TwitterVane lead to spam sites?
 - Yes; see above.
- Would you have selected the URLs found by TwitterVane if you were doing manual selection?
 - No; I don’t think we would have known that this URL was so widely circulated and, therefore, important.
- Do TwitterVane URLs point to certain types of websites?
 - It was difficult to discern trends based on the small number of relevant websites reported.
- Is TwitterVane useful? How does it aid / hinder selection in your view?
 - Conceptually, I still think it could be useful for event-based collections. From the few collections I’ve observed, I’d say that the signal-to-noise ratio is too low for it to be useful at the moment.

Usability

- Is TwitterVane easy to use? Is the UI intuitive?
 - Setting up a collection was very easy. I found the reports inconvenient to use. Every time I wanted to view a report for the same collection (what seems to me to be a common use case), I had to go back to the *Reports* interface, re-select the collection and select the *Report Type*. Additional clicking might be eliminated if the *Reports* interface were designed around the assumption that the user would most often be interested in seeing multiple reports about the same collection within a given session, rather than an arbitrary series of reports from any collection.
- How do you like the layout of the page?
 - Collections: I worried a little bit that the *Add New Collection* form might be too inconspicuous being “below the fold”, especially if there were a lot of collections.
- What would you like to change / add to the UI?
 - A collection-centric interface that provided access to reports.
- What additional reports would you like to see: eg tweets grouped by tweeter – is that something you regard useful?
 - Perhaps a report of co-incident hashtags? That might help to augment a given collection.
- Would you prefer just to see your own collections?

	Web Archiving	Project Name: Evaluating Twittervane	Date: 16 June 2013
		Document Title: Project Final Report	Version: Approved

- I think that would be preferable, though, in an institutional context, it's likely that more than one curator would want to be able to examine the same collection. Perhaps add a checkbox toggle to "show all collections"?

System performance

- Is the response time quick enough?
 - It seems ok in the likely case that I'd only want to peruse several paginated screens of results.

3. Notes

- You will come across "Unknown" in the reports: these are tweets which Twittervane cannot associate or assign to any collection – we could hide these.
- The processing of tweets are done in batches, set to 100 tweets currently – this is configurable.
- Twittervane also optimises URL expansions. Only the top 10 "most popular" (ie: the most frequently appearing) URLs for every processing run (set to 100 tweets). Again this could be changed.
- The "tweets summary by collection" report
<http://194.66.239.180:8080/twittervane/reportView.html?report=tweetSummaryByCollection&sort=desc>
 contain both processed and unprocessed tweets – would like to see them separately?
- Twittervane also has a processing and admin UI which we will evaluate separately.