Trove crowdsourcing behaviour

Paul Hagon
@paulhagon
@TroveAustralia

Monday, 4 March 13
Crowdsourcing profile

Look beyond the numbers

Monday, 4 March 13

2 things I want you to take away today.
How does your perception of that number change when I say that 67% of our visitors spent less than 30 seconds on our site?
Monday, 4 March 13

Does it change further when we measure where people click & you start to get another idea of how your site is used. Doing studies like this lead to the recent redesign of our website.
What's new

Latest records

2935 new records

- Difficulty: Medium
  Status: Complete
  Master transcripts of proceedings before Commonwealth Courts and Tribunals, chronological series – J11
  Page 59 of 59

- Difficulty: Medium
  Status: Not started
  Master transcripts of proceedings before Commonwealth Courts and Tribunals, chronological series – J11
  Page 58 of 59

- Difficulty: Medium
  Status: Not started
  Master transcripts of proceedings before Commonwealth Courts and Tribunals, chronological series – J11
  Page 57 of 59

- Difficulty: Medium
  Status: Not started
  Master transcripts of proceedings before Commonwealth Courts and Tribunals, chronological series – J11
  Page 56 of 59

- Difficulty: Medium
  Status: Not started
  Master transcripts of proceedings before Commonwealth Courts and Tribunals, chronological series – J11
  Page 55 of 59

- Difficulty: Medium
  Status: Not started
  Master transcripts of proceedings before Commonwealth Courts and Tribunals, chronological series – J11
  Page 54 of 59

- Difficulty: Medium
  Status: Not started
  Master transcripts of proceedings before Commonwealth Courts and Tribunals, chronological series – J11
  Page 53 of 59

Hive from National Archives of Australia – http://transcribe.naa.gov.au
Help The New York Public Library improve a unique collection!

We’re transcribing our historical restaurant menus, dish by dish, so that they can be searched by what people were eating back in the day. It’s a big job so we need your help! Learn more.

Connect: menus@nypl.org | Twitter | Facebook

Are you ready for some football?

Sports menus to inspire your Superbowl tailgate!

Help review

Transcribed menus that need a second pair of eyes.

Explore

Browse the collection. 16,303 menus digitized and online.
British library georeference, where users can place maps over Google Earth to give coordinates – http://www.bl.uk/maps/
Welcome!

The key goals of The Commons on Flickr are to firstly show you hidden treasures in the world's public photography archives, and secondly to show how your input and knowledge can help make these collections even richer.

You're invited to help describe the photographs you discover in The Commons on Flickr, either by adding tags or leaving comments.*

Participating Institutions
FAQ
Rights Statement

Flickr Commons where institutions can upload photos for people to add comments & tags. – http://www.flickr.com/commons
And Trove from the National Library of Australia with their newspaper corrections. – http://trove.nla.gov.au/newspaper
A bit of background. When digitising text, if it just goes up, then it’s the equivalent of a browse interface of the physical object. In the case of newspapers you need to know the title, the date, the page etc. Not a good experience. So we need to search...
Search needs text so in this case we need to apply OCR over the digitised text. But OCR isn’t perfect so.
We can improve the OCR by adding in human correction – crowdsourcing.
In turn, this improves search
OCR correction levels. There’s a relatively high OCR correction rate for articles. Human correction is the “icing on the cake”
Monday, 4 March 13

These are a little out of date but they are the sorts of stats that we typically report on for things like the annual report. Fine for overall figures, but not really good at telling us exactly how users are using our resources and how we can improve our services based upon this. I’m really interested in the newspaper corrections.
We’ve estimated that if we had to employ staff it would have cost in the vicinity of $12 million. Massive benefit to the Library & to the community.
This shows what Trove is made up of. Journals, Archived websites & Newspapers are the resources with the most content.

This shows what is being used. 85% of Trove use is from newspapers. It starts to give us an indication of where we can focus time, energy & resources.

I'm really interested in newspapers & the activity surrounding that.
There’s more to Trove than newspaper corrections

One thing to keep in the back of your minds is there’s more to Trove than text corrections. Say after me....
One thing to focus on is newspapers & one of the appeals of newspapers is the correction. 85% of text corrections have been made by users that have created an account on Trove and are logged in. This is a commitment & an indication of having a relationship with the Library.
We have a leaderboard with 23,000 users that have made corrections. It’s not quite gamification and other studies have shown that competitiveness isn’t a main motivation for corrections. There’s a very small amount that have done a lot of corrections & then a super long tail of lots of users that have made very few corrections.
50% users < 100 corrections
75% users < 500 corrections
0.01% users > 1 million corrections

Now less than 100 lines of corrections isn’t a small amount of correction. The big numbers that you most commonly hear are being done by a very small amount of users.
Let’s look at it in a different way. We can’t track behaviour of non logged in users. Approx 23,000 logged in users have made 68,000,000 corrections.
Monday, 4 March 13

100 users have made 43% of corrections
The top 1000 users have made 81% of all corrections
The top 5000 users have made 96% of all corrections.
So we’re starting to see the top users are extremely important to the corrections program.
How do these patterns compare across other crowdsourcing activities? Hive from National Archives of Australia
Hive from National Archives. Much smaller numbers at 448 users, but the usage patterns are nearly identical.
Welcome!

The key goals of The Commons on Flickr are to firstly show you hidden treasures in the world's public photography archives, and secondly to show how your input and knowledge can help make these collections even richer.

You're invited to help describe the photographs you discover in The Commons on Flickr, either by adding tags or leaving comments.*

Participating Institutions
FAQ
Rights Statement

Let's look at the 6 Australian institutions that participate in Flickr Commons. Crowdsourcing their photographs using tags & comments.
Approx 1,005 users have added 31,026 tags

The top 100 users have added 89% of all tags
Flickr comments per user. Once again we start to see an identical pattern.
Monday, 4 March 13

Approx 12,753 users have added 26,173 comments

The top 1000 users have made 47% of all comments
Given the user behaviour, how can we encourage someone from the top 1000 to keep at it to reach the top 100. It’s a massive difference in the amount of corrections needed. To get there, you need to give up work and start text correcting full time.
Could this be ranked not just by the number of corrections but also incorporating how often the user returns, how efficient they are at correcting (not returning to the same article) or by how “difficult” the article might be (as a measure of the initial OCR accuracy). Let’s look at a few options.
Could we rank on accuracy or difficulty of article? We have an approximate OCR accuracy rate and we know the exact amount of characters corrected.
How often do users make corrections. Using same recency patterns as Google Analytics. Over 40% of the top 100 users return on a daily basis.
The top 1000 aren’t quite as dedicated. There’s a decrease in the immediate recency & an increase in the long term return rates.
Monday, 4 March 13

Looking at the pattern for the overall registered user base, 7% of users have made corrections within the past week. Nearly 70% of users haven’t made corrections in the previous 6 months & for nearly 45% of users it’s been more than 12 months since they last made a correction.
So the behaviour for the top 100 users is the opposite to the general behaviour patterns.
To give a bit of an idea of the patterns, 48,822 lines of correction on Christmas Day. Given that an average day will see in the vicinity of 120,000 corrections, it’s quite amazing.
Is there a burnout time, when people have enough of text correction? First time a user made a correction & the last time a visitor made a correction. There isn’t really a burnout.
There doesn’t appear to be any specific time periods that people are targeting (eg: First World War etc).
Monday, 4 March 13

Each article is classified according to the type of article it is: an article, advertisement, births deaths marriages etc. Trove newspapers are mostly articles. Not many articles that are Family Notices.

Once we look at what type of articles are being corrected, there’s some definite activity around family notices.
If we look at it a bit differently. As a percentage of the total article types, nearly 64% of the family notices have had some level of correction.
How can we use this information to dictate the future direction of Trove newspapers?
We keep adding articles to Trove. This isn’t going to stop. It’s increasing in a linear fashion.
The number of corrections that are happening each month isn’t increasing at the same rate.
Likewise the number of users making corrections isn’t increasing in a linear fashion. Are we reaching a plateau in what our existing users are capable of doing?
Let’s get back to the situation we faced at the start of the project. What’s going to happen over the next couple of years into the future? If we keep on putting more & more pages up – what happens when our correctors can’t keep up?
If articles keep getting added & the corresponding number of users aren’t joining or correcting, will search slowly become less effective?
Monday, 4 March 13

Do we need to improve OCR through automated terms. Improvements in OCR technology, general text pattern analysis.
Or improve manual corrections through marketing, promotion, incentives Do we need to change our API to allow write access so machines could programatically correct text?
Do we redesign the interface to highlight articles that have a low correction level? For instance, do we concentrate on years around 1880 or 1930 and not so much the years surrounding the First World War?
How can we get our passionate users doing high value tasks? Would other crowdsourcing activities like geo-spatial references be more valuable. Could they be set up doing specific tasks on uncatalogued material?
We have:

Great content
Passionate users
Family History

Monday, 4 March 13
We have: Passionate users who want to help us. We have niche interest groups like Family History. Getting all of these factors to align with our strategic directions.
Do you look at it the same way as you did 20 minutes ago?
Thanks

phagon@nla.gov.au
@paulhagon

@TroveAustralia