The MeAggregator: Personal Aggregation and the Life-Long Tail of the Web

M.P.Evans, S.A.Williams, P.Parslow, R.P.Ashton, K.O.Lundqvist
and E.Porter-Daniels

Odin Lab, School of Systems Engineering, University of Reading
e-mail: Michael.Evans@reading.ac.uk

Abstract

Content aggregation on the Web is now big business, with major sites such as YouTube, Flickr and Facebook aggregating billions of videos, images, and words from millions of users. Such sites exploit the Web 2.0 concept of the Long Tail in order to grow large, generate revenue, and gain competitive advantage through network effects. In this paper, we present a review of content aggregation, and introduce the MeAggregator, a Personal Aggregator we are developing that enables the aggregation and republishing of an individual's content across the Web. Initially designed for creating ePortfolios and other academic content, we show how the MeAggregator and other Personal Aggregators have the potential to transform the business of content aggregation by leveraging what we call the LifeLong Tail of a user's content.

Keywords

Aggregation, Web 2.0, Long Tail, LifeLong Tail, ePortfolio, eLearning

1. Introduction

Content Aggregators feature strongly on the Web, with sites such as Flickr, YouTube, Blogger, etc., aggregating a vast amount of User-Generated Content (UGC) in the form of words, images, videos, etc. However, most Aggregators currently operate as a walled garden, hosting content but not enabling the user to export it. For the user, this approach offers no control over their content, which, over the course of their lifetime, will become dispersed widely across the Web.

In this paper, we describe a *Personal Aggregator*: a service that aggregates all the content published by an individual, and enables its republishing to other Aggregators. In this way, the user can leverage what we have termed their *LifeLong Tail*: the Long Tail of content they will create in the course of their lifetime. A first step towards a Personal Aggregator is the MeAggregator, a JISC-funded project that we are currently developing. The MeAggregator was designed as an eLearning tool to give students control over their academic content, and to repackage that work as required for ePortfolios. However, the concept extends far beyond eLearning, and has implications for the future development of the Web. We aim to show the clear need for a Personal Aggregator, how such a tool could change the face of the Web, and how the MeAggregator can play a part in enabling the user to leverage the value from their LifeLong Tail of content.

The paper is presented as follows: Section 2 introduces the concept of aggregation in the context of the Web, and describes various content aggregators online today and the business model they employ. The section also describes how Long Tail Economics governs the strategies underlying today's Aggregators, and presents the case for the development of the Personal Aggregator. Section 3 presents the MeAggregator, our own Personal Aggregator currently under development, and compares it with other kinds of Aggregator in use today. Finally, section 4 examines the impact Personal Aggregators could have on the Web by freeing a user's LifeLong Tail of content.

2. Long Tail Economics and the role of Aggregation in Web 2.0

2.1. An overview of Content Aggregation

Adapting Anderson's definition, we define an Aggregator as any Web site that aggregates the long tail of content of a specific media type or contextual theme (Anderson, 2007). Examples of Aggregators and the content they aggregate include Flickr.com (images), YouTube.com (videos), Facebook.com (social data); and Blogger.com (online content in the form of blogs).

An Aggregator makes it easier for a user to find content, as it groups together content of a similar theme or media type in one place. For the Aggregator, aggregating more content than its competitors will enable it to dominate its niche. Both YouTube and Flickr, for example, dominate their respective niches of videos and images (ComScore, 2008). Aggregators are central to what Anderson calls Long Tail Economics, which describes the economic forces at play when resources are abundant, and is one of the foundations of the Web 2.0 model.

2.2. Web 2.0 and Long Tail Economics

Web 2.0 describes the new business models and technologies that have emerged since the end of the DotCom bubble in 2000. According to O'Reilly, Web 2.0 can be defined as the "network as platform", with Web 2.0 applications *"...consuming and remixing data from multiple sources, including individual users, while providing their own data and services in a form that allows remixing by others…and creating network effects through an architecture of participation."* (O' Reilly, 2005). It is this last part of O'Reilly's definition that focuses on the Long Tail.

2.2.1. The Long Tail and Power Laws

The term Long Tail comes from the shape of a power law curve, a scale invariant function such as the Pareto Distribution or Zipfs Law, which exhibits a long tail as the amplitude approaches, but never reaches, zero. Power laws accurately model the popularity of a range of products, with a small percentage of the product (the 'hits') making up a large proportion of the sales. In the case of Amazon, for example, Zipf's Law provides an accurate model of the popularity of the products sold (Brisco et al., 2006), but whereas a traditional retailer such as Wal-Mart would only have

shelf space for the top sellers, Amazon can stock millions of different products (Brynjolfsson *et al.*, 2003).

Aggregators such as Amazon, YouTube, Flickr, etc., are therefore able to sell or distribute items all the way down the tail (the 'niche products'). The total sales value of products in the tail can be worth as much as the value of the hits, giving rise to the idea of the Long Tail, "...a power law that isn't cruelly cut off by bottlenecks in distribution such as limited shelf space and available channels" (Anderson, 2006).

2.2.2. The three forces underlying Long Tail Economics

Long Tail Economics, according to Anderson, comprises three key forces:

1. *Democratize the tools of production* – the availability of cheap tools to produce UGC (e.g. digital cameras, blogging platforms, etc.) enables the creation of an enormous range of content.

2. *Democratize the tools of distribution* - By aggregating large volumes of UGC, an Aggregator makes it easy for content to be distributed and accessed widely, which in turn enables the Aggregator to benefit from the power of network effects (i.e. the Web site increases in value exponentially with increasing numbers of users and content).

3. *Connect Supply and Demand* - Each individual item of content will be relevant to someone, even if it is only once in that item's lifetime, which is why the Aggregators try to aggregate as much content as they can (Flickr, for example, currently aggregates over 2 billion images (*http://blog.flickr.net/en/2007/11/13/holy-moly/*)). However, users will only find the niche content they want if a filter is provided to filter out content that is irrelevant to them. For example, Amazon's Recommendation Engine uses choices made by previous customers to suggest products to new customers (Jacobi and Benson, 1998); iTunes's music filter directs users to other types of music from selections chosen by other users with similar tastes to their own; and Google's PageRank algorithm (Brin and Page, 1998) effectively filters the results of Web pages returned from a user's query, making the search engine's results more relevant to their information need. Most of these filters rely on the use of social data: that is, the previous choices made from millions of users (the "Wisdom of Crowds" effect (Surowiecki, 2005)).

2.3. Blogs as Aggregators

Long Tail Economics applies just as much to blogs as it does to Aggregators. Using data taken from the blog MobileMentalism.com, published by Evans since 2005, we find the long tail in the number of times each post in the blog has been viewed, ranked according to viewing frequency (figure 1a), and in the search terms used to find the blog (x axis) ranked according to their frequency of use (y axis) (figure 1b).

It is the long tail of search terms that makes search engines so important in driving traffic to an individual Web site or blog. For example, over 50% of MobileMentalism's traffic comes from external search engines, with the top 10 search terms appearing 12% of the time, and the long tail of search terms (26,345 in total) comprising the remaining 88%.

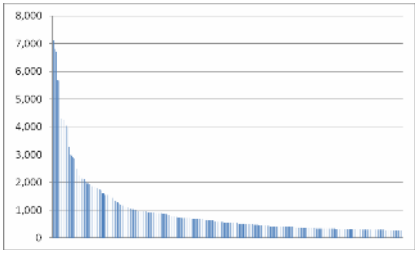

Figure 1a – Long Tail of posts, ranked by viewing frequency

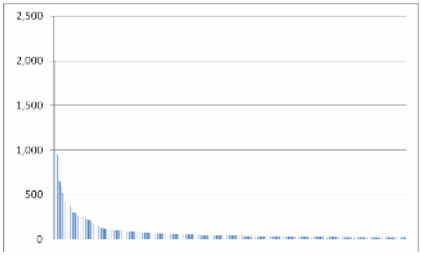

Figure 1b – Long Tail of search terms, ranked by frequency

Search engines therefore act as filters, connecting the supply of posts with the information demand from users, represented by their search queries. As more posts are written, so more search terms become relevant, which increases traffic to the blog, and with it, revenue, as there is a simple correlation between traffic and revenue earned from advertising.

2.3.1. Maximizing revenue through strong situational context

The forces of the Long Tail are not enough on their own to earn significant revenue, however. Although they may lead to a large number of users visiting a site, of far more value than the quantity of users is their quality: that is, users who are highly interested in a site's content, rather than browsing a site aimlessly. To attract high quality users, a blog needs to focus on one topic. Such *niche blogs*, as they are known, are focused in this way in order to provide a strong situational context for the user, the search engine and the advertiser. By focusing on one topic, the search engine is able to identify the topic more easily, and the set of relevant search terms will be specific to that topic. This, in turn, ensures that the blog's audience mainly comprises people who are interested in that topic, which encourages advertisers to pay a premium to advertise on the blog, as the audience is more likely to buy their products. Thanks to the nature of contextual advertising, in which ads are served based on the words contained in a page's content, all words now have financial value. Maximizing that value, however, relies as much on situating the words in the appropriate context as it does on the meaning underlying the words themselves.

Niche blogs, therefore, earn money thanks to the effects of the Long Tail combined with a strong situational context. However, this is actually the same strategy that the Aggregators employ. Aggregators may offer content that is dispersed across many topics compared to niche blogs, but they are able to funnel that content into individual niches through their own internal filters that connect supply with demand. For example, recommendation engines (Amazon), channels (e.g. YouTube), categories (e.g. Digg) and tags (e.g. Del.icio.us, Flickr, etc.) all act as filters, each of which serves to create an individual niche within the Aggregator that employs it.

The only differences between a blog and an Aggregator, therefore, is that a blog relies on an external filter (Google, in this case) to connect its supply with demand, whereas an Aggregator uses its own internal filter; and a blog must focus on one specific niche, whereas an Aggregator can focus on many, relying on its own internal filter to create niches that are themselves driven by user demand. A blog can therefore be thought of as an Aggregator in its own right, aggregating all posts from its author that are specific to one particular niche topic. It is here, however, that the current Aggregator model reveals its inherent unfairness from which it is under increasing pressure to change.

2.4. The changing nature of Aggregators

Long Tail Economics shows that an Aggregator needs a large body of content that can be created easily by its users, plus a filter to connect supply with demand. Both the UGC that the Aggregator aggregates, and the data from the users' interaction with this content, which the Aggregator uses to tune its filter, effectively become the Aggregator's most valuable assets. The Aggregator must leverage these assets to increase its competitive advantage over its rivals and ensure that its users remain loyal. However, this places pressure on the Aggregator to retain this content and data, and prevent them from being exported to its competitors. Effectively, the Aggregator becomes a walled garden, hosting content within its walls, but not letting that content be shared amongst other sites.

In a recent interview, Tim Berners-Lee said that "*...if you look at the social networking sites which, if you like, are traditional Web 2.0 social networking sites, they hoard this data. The business model appears to be, 'We get the users to give us data and we reuse it to our benefit. We get the extra value'" (Berners-Lee, 2008)*. This issue is slowly changing. Google, for example, has recently announced its OpenSocial platform (*http://code.google.com/apis/opensocial/*), while the Data Portability group is working on a set of standards to control "...the profiles, relationships, content and media we create and maintain, regardless of what platform they are hosted on" (DataPortability, 2008).

However, it is not just social networks that need data (i.e. content) to be portable. From a user's perspective, they should be free to aggregate their own content from across the Web, and distribute it where they see fit. This enables them to repurpose their content in different formats, for purposes such as ePortfolios, and republish it in different locations to leverage their LifeLong Tail of content (see section 4). In short, although aggregation as a business model has proved fruitful, there is a growing need for it to change. A new breed of Aggregator is needed, one that we call the Personal Aggregator, which has its origins in several life stream services that exist now, but which will ultimately be realised in new tools such as the MeAggregator, a JISC-funded project that we are currently working on that aims to facilitate the aggregation of personal content for academic use.

3. The MeAggregator – Personal Content Aggregation and Distribution

The MeAggregator is a JISC-sponsored project designed to facilitate eLearning through the use of social networking technologies and data portability. The key aim of the MeAggregator is to provide a seamless space that covers the general online world of the Web and the more focused world of academia, enabling a user to publish and manage their content from a central online space, and republish it at will across a variety of Aggregators and other services.

Specifically, the MeAggregator's novel approach enables the user to:

- import and export content to different Aggregators (both across the Web and within academic tools such as the Blackboard Virtual Learning Environment);
- remix and repackage their content for different purposes, such as eLearning portfolios, instructional articles, business reports, or personal use;
- share their content with other MeAggregator users, and form communities around items of content.

The initial plan for the MeAggregator is to enable the free exchange of content between the Blackboard Virtual Learning Environment used at Reading, Facebook, our own internal social network called RedGloo (Williams, et al., 2007) and a number of blog platforms. The goal is to provide a tool that students and staff can use to aggregate, publish and share information with each other, that can be reused in the future for ePortfolios, and which will ultimately lead to a body of information specific to the needs of a University that future students can learn from.

Given that the initial community comprises University staff and students, the set of content that emerges across all MeAggregators will be instructional and referential in nature. We plan to make this content searchable, thereby creating a resource for future students to work with, enabling them to share ideas and to see examples of best academic and professional practice. In addition, students and staff will be able to form communities around items of content, enabling them to discuss ideas or to help each other understand the topics being presented. The MeAggregator will therefore foster the emergence of communities of practice around specific topics.

3.1. MeAggregation is Personal Aggregation

Table 1 lists the results of a stakeholder analysis we conducted. As can be seen, the majority of use-cases centre around the MeAggregator as a Personal Aggregator: that is, a tool to give the user control over their own content, and to publish it and republish it as they wish. This enables the user to keep track of their content, to reformat it for new publication elsewhere (such as for use in an ePortfolio), and to aggregate all the knowledge they've learned over time and have it accessible to them whenever they need it. Equally, through the ability to share and tag this information

across a University, the body of knowledge from across the institution will be aggregated and made much easier to access.

Stakeholder	Use	Stakeholder	Use
Students	Using the MeAggregator to enable the representing of content across institutional and user owned technologies	Existing Aggregators	Broadening access to their products
Other learners	Personal use of the MeAggregator	Careers Staff	Facilitating students in creating ePortfolios. Plus personal use of the MeAggregator
Staff	Allowing students more flexibility in how they present and access material. Additionally, personal use of the MeAggregator	Larger community	Personal use of the MeAggregator
Alumni	Aggregating material, keeping in touch with their educational establishment	University of Reading	Empowering learners and teachers to work beyond the institutional boundaries

Table 1: Results of the Stakeholder Analysis for the MeAggregator

3.2. Aggregation in technical detail

The MeAggregator must work with a variety of technologies in order to facilitate the free importing and exporting of data from one Aggregator to another. Achieving this is non-trivial, however, as each Aggregator may offer different levels of data import/export according to the technology used. For example, some Aggregators, such as Flickr, provide full importing and exporting services using APIs based on Web services, whereas other Aggregators, such as Facebook, provide only minimal data export options through the use of RSS feeds.

Accordingly, the MeAggregator must adapt to whichever technology is used by the Aggregator the user wishes to work with. To achieve this, we have defined the MeAggregator Aggregation Stack (table 2), which categorizes the detail of data that can be imported and exported according to the technology used.

Level	Technology	Detail
Full	Access to services that provide an API, REST interface, XML-RPC or equivalent.	Importers and exporters are developed as plug-ins of the MeAggregator, which allow full access to relevant data from the different services.
Interpreted	Interpreted through HTML or RSS. (JavaScript versions are also envisaged, but are not the first priority.)	The user can manually aggregate data from websites. Set up and automated aggregation achieved through RSS using filtering.
Low	Picture or recording of content	The user can store the current data in a picture, sound recording or video.

Table 2: MeAggregator Aggregation Stack

When the user wants to aggregate from a source, the MeAggregator will first determine the highest level of aggregation that is available, and then provide the user with a suitable interface to perform the aggregation. Repurposing of the data for exporters (for instance changing data from one format to another: HTML to PDF, or Wiki to docx etc.) will be performed using a modified version of Apache Forest (http://forrest.apache.org/), which is a publishing framework for transformations between different formats.

3.3. Comparing the MeAggregator with other Personal Aggregators

Personal aggregation is a newly emerging field, with many new Web 2.0 sites claiming to aggregate all of a person's social networking content from around the Web. Current examples include *FriendFeed, Soup.io, Plaxo, Iminta, Spokeo, ProfileLinker, MyLifeBrand, Fuser, 30Boxes, Mugshot, Readr* and *Second Brain*. However, these sites are simply *Profile Aggregators*; that is, they import a user's profile and content from existing Aggregators (principally social networks), and present it on one Web page for easy viewing in what is termed a *life stream* (i.e. a personal news feed featuring content from you, your friends and family); there is no option to republish the content onto other Aggregators or Web sites.

In contrast, *SocialStream* (Clarke et al., 2007), a Google-sponsored project by Carnegie-Mellon researchers, is a true Personal Aggregator that not only aggregates content from a variety of other Aggregators, it also enables content to be republished. In this, SocialStream is similar to the MeAggregator. However, SocialStream only facilitates the exchange of content between Aggregators, whereas the MeAggregator approach also enables the reformatting and repackaging of content, extends aggregation into the academic space through integration with existing eLearning tools such as BlackBoard, and provides a new platform for the emergence of communities centred around topics of interest. In addition, the MeAggregator will add an extra dimension to the Personal Aggregation concept by fostering new communities around academic subjects, and leading to a new approach to creating, storing and retrieving academic material.

4. Personal Aggregation and the user's LifeLong Tail

A user's LifeLong Tail is the set of all digital content they create in the course of their lifetime. As all words now have financial value, much of this content, such as reports, email advice, recommendations to friends, etc., could be published on the Web, where it would have the same potential to generate revenue as any other Web content. However, in practice, this is difficult to achieve due to *weak situational context* and *valueless Aggregators*:

> *Weak situational context* - Much of a user's content could be republished in niche blogs, which provide strong situational contexts, but a user cannot set one up for every topic they write about. Consequently, content ends up being published either in weak situational contexts, such as a personal blog, or in part of the Invisible Web, the part of the Web that search engines cannot reach. This comprises:

- *the Deep Web*, in which content is buried in databases that are inaccessible to search engines. Approximately 302,000 Web sites fall into this category, with only a third of them indexed by the major search engines (He *et al*., 2007);

- *the Hidden Web*, in which content is hidden behind password-controlled interfaces. Facebook, for example, comprises over 100 billion stories for its newsfeed, with over 700 million added every day (McClure, 2007) – and none of them are indexed by the search engines.

- *the Outer Web*, in which content is published so far down a page, it is beyond the reach of a search engine's crawler (Yahoo!, for example, only indexes the first 210KB of Web page; Google, 520KB; and MSN, 1,030KB (Bondar, 2006)).

Other places where content may be stored, but which gives no financial value, include a company's intranet, email, or internal wikis, etc.

Valueless Aggregators - The content could, of course, be republished in a variety of Aggregators, which provide a much stronger situational context in the form of channels, tags or categories, etc., but these do not generally share their revenue, and so offer no value to the user for their content other than free hosting.

In contrast, Personal Aggregators such as the MeAggregator would enable the user to choose where to place their content according to the revenue it is likely to generate. Content would be free to move between niche blogs and Aggregators depending on which offered the user the greatest return. This would drive the Aggregators to offer a share of the revenue from the user's content in order to attract content no longer walled in. As the user continually generates content over the course of their lifetime, so more revenue will be earned as they become their own aggregator, increasingly leveraging the economics of their own personal LifeLong Tail of content.

5. Summary

We have presented the initial concept of our new eLearning tool, the MeAggregator, in the context of content aggregation on the Web. The MeAggregator is designed to facilitate the free distribution and reformatting of content, enabling content published elsewhere to be aggregated and repackaged for ePortfolio purposes, and enableing new communities of practice to emerge centred around academic topics of interest. In addition, we have described how the concept of Personal Aggregation goes much further, by viewing the content output of a user taken over the course of their lifetime. From this perspective, we described the LifeLong Tail of content, which has the potential to be extremely valuable. Ultimately, the MeAggregator offers the first step towards this vision of the Personal Aggregator. Future work will describe the design and finished solution, and assess its potential in helping bring this vision closer to reality.

6. References

Anderson, C. (2006), "The Long Tail: How Endless Choice Is Creating Unlimited Demand", Random House Business Books, July 2006. ISBN-10: 184413850X

Berners-Lee, T. (2008), Interview with Paul Miller, ZDNet, Feb. 26th 2008, http://blogs.zdnet.com/semantic-web/?p=105&tag=nl.e539

Bondar, S. (2006), "Search Engine Indexing Limits: Where do the Bots stop?", SitePoint, April 2006, http://www.sitepoint.com/article/indexing-limits-where-bots-stop

Brin, S., and Page, L. (1998), "The anatomy of a large-scale hypertextual Web search engine", In: Computer Networks and ISDN Systems, Volume 30, Issues 1-7, April 1998, Pages 107-117, Proceedings of the Seventh International World Wide Web Conference

Brisco B., Odlyzk A. and Tilly B. (2006) "Metcalfe's Law is Wrong", IEEE Spectrum, July 2006, http://spectrum.ieee.org/jul06/4109

Brynjolfsson, E., Hu, Y.J. and Smith, M.D (2003) "Consumer Surplus in the Digital Economy: Estimating the Value of Increased Product Variety at Online Booksellers", In: Management Science, Vol. 49, No. 11, November 2003, pp. 1580–1596

Clark, M., Crowe, B., Koranda, C., Jitkoff, J.N. and Treichler, L. (2007), "Social Stream", Carnegie Mellon University HCII, 2007, http://www.hcii.cs.cmu.edu/M-HCI/2006/SocialstreamProject/index.php

Comscore 2008, "Google Sites' Share of Online Video Market Expands to 31 Percent in November 2007, According to comScore Video Metrix", Jan 17th 2008 http://www.comscore.com/press/release.asp?press=2002

DataPortability, 2008, "The Data Portability Charter", http://groups.google.com/group/dataportability-public/web/charter

He, B., Patel, M., Zhang, Z. and Chen-Chuan Chang, K. (2007), "Accessing the Deep Web", Communications of the ACM May 2007/Vol. 50, No. 5

Jacobi, J.A, and Benson, B. (1998) "System and methods for collaborative recommendations", Amazon patent, number 6064980, 17th March 1998, http://www.google.co.uk/patents?hl=en&lr=&vid=USPAT6064980&id=sUMEAAAAEBAJ&oi=fnd&dq=amazon+recommendation+engine

McClure (2007) "Facebook Technology and Tasting Event," Palo Alto, California, Feb. 2007, http://flickr.com/photos/500hats/sets/72157594550002847/

O'Reilly, T., 2005, "What is Web 2.0 – Design Patterns and Business Models for the Next Generation of Software", http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html?page=1

Surowiecki, J. (2005), "The Wisdom of Crowds: Why the Many Are Smarter Than the Few", Published by Abacus, March 2005. ISBN-10: 0349116059

Williams S. A., Lundqvist K. Ø. and Baker, K.D. (2007), "Communities via a Learning Landscape", In: Proceedings of the Second iLearning Forum, Paris, France, 29-31 January 2007.