“Designated Communities”: Through the Lens of the Web

Yunhyong Kim
HATII, School of Humanities
University of Glasgow

Abstract

The notion of a “designated community” has always been a rather elusive concept across the digital curation landscape. This paper is an effort to revisit the concept to stir up new discussions in the area. More specifically, this study offers a perspective on designated communities through the lens of the web, powered by developments in the last ten years in social media. The research presents a multi-faceted analysis of communities based on HTML content from online web pages to propose heuristics for defining designated communities based on the technology they adopt, properties of knowledge organisation, and how they link to each other. This impacts the building of quantifiable models of designated communities, estimating curation risks associated to the community and further, refining approaches to preservation strategies that meet the needs of the community.
Introduction

The notion of a “designated community” has always been an elusive concept within the digital preservation community. This term has come into wide scale use within the archival community since its use in the reference model for an Open Archival Information System (OAIS), published as an ISO standard in 2003 (ISO, 2003). The notion has propagated a high level understanding of the concept without a full discussion of the technical complexity locked within, and identifying community needs has been largely left as part of the workflow for policy development. The aim of this paper is to raise a new perspective on the designated community and to stir the digital preservation community into action to re-investigate this concept.

Ten years ago and in the years thereafter, digital preservation research has largely focused on object oriented engineering solutions to curatorial problems, with a tendency to avoid delving too much into the complexity behind designated communities and their needs. During the same time period, we have experienced the advent of social media, which has caused the explosion of visible online communities (e.g. O’Callaghan et al., 2013; Massimi et al., 2014). This phenomenon has opened up a window into the workings of communities (e.g. Coulon, 2005; Lampe, Ellison and Steinfeld, 2008; Rowson, Broome and Jones, 2010) in a way that was not possible in the first stages of digital preservation research.

The research reported here proposes that content analysis of web pages can provide insight when profiling designated communities to understand complexities that might arise in digital preservation processes for weblogs (e.g. for example the range of different types of material that need considering in the target information community). This investigation is only a lightweight study to open up approaches to analyse “designated community” as a complex system.

The concept of complexity is mentioned in many disciplines to refer to a number of different phenomena. The complexity under scrutiny here is limited to that related to emerging community behaviours regarding the variety of platforms and file formats adopted to disseminate information, the range of terms and concepts adopted to organise information, and the links established to communicate information. The investigation employs, as a proxy, a quantitative analysis of HTML markup, user-generated categories and topic tags, and link network to generate target community profiles.

The study makes three contributions to current research landscape:

- It presents heuristics for a comparative study of different online communities, on multiple levels, especially, highlighting those related to blogging communities,
- The work opens up the first steps towards automated processes for deriving a quantified model of a “designated community” that can be used by the web archiving community, and
- The discussion highlights an approach for how the results in this paper on online communities can be generalised to be applied to the general digital curation context.

---

1 For example, see definition from York University (2013).
2 For example, see Alliance for Permanent Access (n.d.).

IJDC | Peer-Reviewed Paper
Datasets

The analysis reported here was carried out on web home pages. The study was limited to home pages in order to make the study comparable across the datasets without noise propagated through page type variations. The analysis was performed using a fresh crawl of URLs from three datasets. The datasets are summarised in Table 1.

Table 1. Datasets used in the study. Right most column is included as a preliminary indication of the variation in the HTML source code.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>No. of URLs</th>
<th>No. of unique “&lt;!DOCTYPE&gt;” declarations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spinn3r</td>
<td>223 145</td>
<td>80</td>
</tr>
<tr>
<td>ClueWeb09</td>
<td>214 952</td>
<td>1420</td>
</tr>
<tr>
<td>BF16Cat</td>
<td>31 690</td>
<td>122</td>
</tr>
</tbody>
</table>

The first of these datasets comprise the crawl of 223 145 URLs from the “Spinn3r” dataset. The URLs of the Spinn3r dataset are collected as weblogs and social media websites and were valid links at the time the dataset was initially harvested, between August and October 2008.

The second dataset is a crawl of 214 952 URLs from the “ClueWeb09” dataset. The ClueWeb09 dataset represents general web pages, and were valid links when the dataset was collected between January and February 2009.

The third dataset is “BF16Cat”, consisting of 31 690 URLs of home pages across blogs in eleven subject areas and non-blog in five sectors. The URLs for BF16Cat were either collected from blogrolls of individual bloggers in the corresponding area, or through blog searching portals such as Technorati, mathblogging.org, scienceseeker.org, scienceblogging.org and Independent Fashion Bloggers (IFB). The sources for BF16Cat are shown in Table 2.

Table 2. Subcategories and/or domains associated to BF16Cat.

<table>
<thead>
<tr>
<th>Type</th>
<th>Subcategory</th>
<th>Size</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blogs</td>
<td>Computer Science (CS)</td>
<td>41</td>
<td>StackOverflow</td>
</tr>
<tr>
<td></td>
<td>Information Technology (IT)</td>
<td>138</td>
<td>Technorati “IT” category search</td>
</tr>
<tr>
<td></td>
<td>Entertainment (ET)</td>
<td>110</td>
<td>Technorati “Entertainment” category search</td>
</tr>
<tr>
<td></td>
<td>Fashion (FA)</td>
<td>164</td>
<td>Independent Fashion Bloggers</td>
</tr>
<tr>
<td></td>
<td>Game (GA)</td>
<td>7</td>
<td>University of Glasgow PhD student in Games</td>
</tr>
<tr>
<td></td>
<td>Health Blogs (HB)</td>
<td>130</td>
<td>Technorati “Health” category Search</td>
</tr>
</tbody>
</table>

3 Crawl performed in July 2012.
4 ICWSM 2009 Spinn3r blog dataset: http://www.icwsm.org/data/
5 ClueWeb09 dataset: http://www.lemurproject.org/clueweb09.php/
6 Technorati: http://technorati.com/
7 Independent Fashion Bloggers: http://heartifb.com
<table>
<thead>
<tr>
<th>Type</th>
<th>Subcategory</th>
<th>Size</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mathematics I</td>
<td>(M1)</td>
<td>110</td>
<td>Field’s Medalist Terry Tao’s Blog</td>
</tr>
<tr>
<td>Mathematics II</td>
<td>(M2)</td>
<td>552</td>
<td>Mathblogging.org</td>
</tr>
<tr>
<td>Music</td>
<td>(MU)</td>
<td>70</td>
<td>Technorati “Music” category search</td>
</tr>
<tr>
<td>Politics</td>
<td>(PO)</td>
<td>107</td>
<td>Technorati “Politics” category search</td>
</tr>
<tr>
<td>Science</td>
<td>(SC)</td>
<td>1071</td>
<td>Sciencesseeker.org and Scienceblogging.org</td>
</tr>
<tr>
<td>Non-blogs</td>
<td>Construction</td>
<td>27</td>
<td>National Building Specification website^8</td>
</tr>
<tr>
<td></td>
<td>Company (CC)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Fashion Company</td>
<td>61</td>
<td>Smashing Magazine Showcase of Beautiful Fashion Websites^9 and comments</td>
</tr>
<tr>
<td></td>
<td>(FC)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Funding Council</td>
<td>51</td>
<td>Search on Google</td>
</tr>
<tr>
<td></td>
<td>(FU)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Government</td>
<td>572</td>
<td>Politics Resources website^10</td>
</tr>
<tr>
<td></td>
<td>(GO)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>University</td>
<td>100</td>
<td>Sedghi and Evans (2012)</td>
</tr>
</tbody>
</table>

Examining HTML tags in the dataset, it was found that the total number of tags were 2,676,612, 139,176,708 and 145,535,289, for BF16Cat, Spinn3r and ClueWeb09, respectively. Despite the large number of tags, 22 distinct tags account for 93.79%, 87.96%, and 90.82% of all tag usages in BF16Cat, Spinn3r, and ClueWeb09, respectively. The datasets are profiled according to the frequency of each 22 tags relative to each other, in Figure 1.

![Figure 1. Relative distribution of the 22 most popular HTML tags across datasets.](image)

---

^10 Politics Resources website: [http://www.politicsresources.net/official.htm](http://www.politicsresources.net/official.htm)
Platform and File Formats in Use

In this section, we emphasise the differences between web pages across different domains of BF16Cat to draw out the distinctions between different communities. One distinguishing factor comes from the content management platform (e.g. WordPress\textsuperscript{11} and/or Joomla\textsuperscript{12}) the community chooses to adopt, and another from types of embedded content (e.g. file formats).

Communities Profiled by Content Management Systems They Use

The HTML code for the home pages belonging to the categories in BF16Cat have been analysed to produce the profile of each community in terms of the number of times each content management system (CMS) was adopted for the community home page. The results are displayed in Figure 2.

While Figure 2 shows many platforms being used by each community, it also suggests each community has a fairly strong preference for one or two platforms over others. For example both game blogs and mathematics blogs are predominantly deployed on WordPress. The figure illustrates substantial amount of variation across communities (e.g. government websites seem to have a strong preference for Joomla).

Indeed, in Figure 3, a hierarchical clustering of communities is displayed (based on Pearson coefficient (Pearson, 1895) distance measure and Ward’s method (Ward, 1963) for hierarchical clustering) which illustrates a clear division between the blogging websites (bottom sub-tree in pink) and non-blogging websites (top sub-tree in blue). The clusters include predictable pairings, for example, of funding councils and universities, of game blogs and entertainment blogs, of computing blogs and mathematics blogs, and of fashion blog and fashion company websites.

There are some curious connections that also emerged (e.g. mathematics I and music; health blogs and mathematics II) but further investigation is required to make any conclusions. Nevertheless, the clustering results suggest that platform selection might be influenced by the community of the blogger.

\textsuperscript{11} WordPress: http://wordpress.com/
\textsuperscript{12} Joomla: http://www.joomla.org/
Communities Profiled by File Extensions They Use

The HTML can also be examined for “file-extensions”. This information was extracted by extracting patterns consisting of a full stop followed by up to four characters from all the HTML tag attribute values. A hierarchical clustering of domains based on these “extensions” was performed to produce the diagram in Figure 4.

The result here shows that there is a clear division between blogging and non-blogging communities. It also shows some obvious connections between computer science and mathematics, and music and entertainment. Some of the earlier clusters that emerged with respect to platform selection re-emerge with respect to “extensions”, which suggests that the clustering might be influenced by “file extensions” associated to content management system attributes rather than content generated by blog/website authors.

Web Page Links

As this study was only intended to be a lightweight study, the concepts of betweenness, closeness, cliques, authorities and hubs of social networks have not been attempted. Instead the investigation was limited to a quick comparison of “referencing behaviour” of the communities.
Figure 5. Link structure derived from three blog and three non-blog home pages.

The network traffic can vary considerably across dataset. As an example, in Figure 5, we have included the network traffic of six of the collections in the BF16Cat dataset. Each node in the structure represents a webpage in the collection and the edges represent instances where a page has provided a hyperlink to another blog in the community using the “href” attribute field value for the HTML tag <a>. The representation in the figure is based on 100 random pages from each of the collections. It does not include references to their own websites.

The figure clearly shows the Maths network to be the busiest network among the six. However, more striking is the noticeable difference between blogs and non-blogs with respect to the number of connections made to external sites in the same domain. In contrast to this, the graph in Figure 6 shows that politics blog pages have the most number of links per page, on average.

Figure 6. Average number links per URL from each subcategory of BF16Cat.
This suggests that, while mathematics blogs tend to provide references to resources within the community, politics blogs point to resources outside of the politics blogging community. This observation motivates the general examination of average number of links, distinct links, external links and self-referential links. This is presented in Table 3. As might be expected, there seem to be more links on a blog web page on average than on a non-blog web page. However, non-blogs seem to be much more self-referential than blogs, that is they point to resource within their domain much more than to resources elsewhere in the same community (for example, universities tend not to link to other universities).

<table>
<thead>
<tr>
<th>Category</th>
<th>Average number of links per page (total no. links)</th>
<th>Distinct links</th>
<th>No. of non-self referential links</th>
<th>Self-referential links (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Construction Company</td>
<td>41.82 (1129)</td>
<td>843</td>
<td>125</td>
<td>0.8892825509</td>
</tr>
<tr>
<td>Computer Science</td>
<td>226.76 (9297)</td>
<td>6823</td>
<td>4773</td>
<td>0.4866085834</td>
</tr>
<tr>
<td>Information Technology</td>
<td>219.57 (30300)</td>
<td>21493</td>
<td>10313</td>
<td>0.6596369637</td>
</tr>
<tr>
<td>Entertainment</td>
<td>261.21 (28733)</td>
<td>19357</td>
<td>9960</td>
<td>0.6533602478</td>
</tr>
<tr>
<td>Fashion Blog</td>
<td>225.24 (36940)</td>
<td>28660</td>
<td>23513</td>
<td>0.3634813211</td>
</tr>
<tr>
<td>Fashion Company</td>
<td>105.57 (6440)</td>
<td>4926</td>
<td>1154</td>
<td>0.8208074534</td>
</tr>
<tr>
<td>Funding Council</td>
<td>71.84 (3664)</td>
<td>2803</td>
<td>503</td>
<td>0.8627183406</td>
</tr>
<tr>
<td>Game Blog</td>
<td>312.00 (2184)</td>
<td>1479</td>
<td>714</td>
<td>0.6730769231</td>
</tr>
<tr>
<td>Government</td>
<td>75.80 (43356)</td>
<td>31524</td>
<td>8464</td>
<td>0.8047790387</td>
</tr>
<tr>
<td>Health Blog</td>
<td>224.42 (29175)</td>
<td>21408</td>
<td>14054</td>
<td>0.5182862039</td>
</tr>
<tr>
<td>Mathematics Blog I</td>
<td>195.96 (21360)</td>
<td>15251</td>
<td>8977</td>
<td>0.5797284644</td>
</tr>
<tr>
<td>Mathematics Blog II</td>
<td>214.62 (118471)</td>
<td>83283</td>
<td>44349</td>
<td>0.6256552236</td>
</tr>
<tr>
<td>Music Blog</td>
<td>223.93 (15675)</td>
<td>11357</td>
<td>8959</td>
<td>0.4284529506</td>
</tr>
<tr>
<td>Politics Blog</td>
<td>361.08 (38636)</td>
<td>27709</td>
<td>20163</td>
<td>0.4781292059</td>
</tr>
<tr>
<td>Science Blog</td>
<td>233.10 (249652)</td>
<td>155420</td>
<td>129754</td>
<td>0.4802605226</td>
</tr>
<tr>
<td>University</td>
<td>89.83 (8983)</td>
<td>7369</td>
<td>2095</td>
<td>0.7667816988</td>
</tr>
</tbody>
</table>

**User Generated Category and Topic Tags**

User generated tags are largely associated with social networking media technology. The generation of topic tags is expected to be fairly active in the blogging communities but not so much so in non-blogging communities. Indeed there were no user generated categories and/or tags in the non-blog subcategory collections of BF16Cat dataset except for six categories (“tv”, “tourism”, “sports-news”, “somali-politics”, “somali-news”, and “business-news”) that were used in one of the government home pages.

Here we present some observations on how we might combine the analysis of category and topic tags generated by users with graph analysis to produce a
representation of domain knowledge. As a starting point, the harvested blogs were investigated to create term graphs. The steps for creating the graphs are as follows:

1. The home page address is the root node of the graph.
2. Each category and topic tag that appears in the home page is a node of the graph.
3. There is a directed edge from the root node to a category for every category found on the page.
4. There is a directed edge from category to topic tag for each category and topic tag that appear together in a post.

The initial result of taking these steps, a sample graph created from a game blog\(^\text{13}\) and a mathematics blog\(^\text{14}\), is presented in Figures 7 and 8.

![Figure 7. Graph constructed from user generated categories and topic tags for a sample game blog.](http://www.vg247.com)

![Figure 8. Graph constructed from user generated categories and topic tags for a sample mathematics blog.](http://terrytao.wordpress.com)

The proposal for the next step to refine the graph to reflect community knowledge is to layer the graph according to the URL frequency associated to each category and topical tag terms across the community. For example, the graph for the mathematics blog in Figure 8 shares the category “expository” with at least four other mathematics blogs. The “expository” node leads to 16 other topics, such as “Bolzano-Weierstrass-theorem” and “experimental-sums”. The term “experimental-sums” is also linked from

\(^{13}\) VG 24/7: [http://www.vg247.com](http://www.vg247.com)

\(^{14}\) Terry Tao Blog: [http://terrytao.wordpress.com](http://terrytao.wordpress.com)
the category “polymath” which is not shared as much in the community. This shared
topic suggests that there is a relationship between “expository” areas and “polymath”. In
fact, Polymath Projects\textsuperscript{15} has been established for “massively collaborative online
mathematical projects” of an experimental nature.

\textbf{Benefits to Digital Curation}

The topics in this paper highlight the potential of social media infrastructure for sharing
information when profiling a target “designated community”. The main objective of the
study was to better define “designated communities” by providing a quantifiable profile
of communities to complement labour-intensive qualitative processes. The profile could
be used to inform repositories right at the beginning of the collection process by 1)
adding community context to characterise the target materials to be collected, and 2)
providing first steps to assess the risks associated with the variety of adopted
technologies and formats, changes in concepts with respect to knowledge organisation,
and social impact of information loss given the community knowledge exchange
network. Studies of performing retrospective data analytics on collections (e.g. Jackson,
2012; Rimkus et al., 2014) unearth collection standards and trends but do not help to
determine whether or not these standards meet the needs of the community and/or
expose the technological culture that imposes obstacles to enforcing the use of
standards. By studying the community in action as information sharing takes place, this
gap can be bridged.

The study presented here was motivated in the development of policies for the
preservation of weblogs, and as such the target materials come with web and social
media platform infrastructure in place to make community profiling viable. The use of
the approach on material from a non-web environment may be, therefore, questioned.
However, this can be overcome through two immediate solutions: 1) by embedding the
repository within the social media information sharing workflow, and/or 2) by
developing a agreed approach for mapping the “designated community” to an online
presence (cf. Liu et al., 2012).

\textbf{Conclusions}

In this article we discussed the ways in which communities follow a recognisable
pattern with respect to: the content management platform they select to create their
websites, the file formats in use within the community, how they reference resources
within the community, categories and topics they use, and how they organise these
categories and topics. In particular, it has been demonstrated how these practices differ
between blogging and non-blogging communities, and between different types of sub-
domains associated to these. The article has demonstrated how these practices can be
studied using a automated process based on the HTML content of the home pages
associated to community website. This could be highly valuable for studying the
concept of “designated communities” by supplementing qualitative approaches already
used in the curation community. The study highlights the possibility that the maturity of
social media and other collaborative technologies may be leveraged to capture
designated community practices even when target materials are not created on the web.

\textsuperscript{15} Polymath Projects wiki: http://michaelnielsen.org/polymath1/index.php
Acknowledgements

The author would like to thank the European Commission for co-funding the BlogForever project (grant no. 269963) and the University of Glasgow for providing the funds and resources for carrying out this research. I would like to thank Seamus Ross for valuable feedback and Matthew Barr for directing me to the relevant game blogs.

References


