SAGE Research Methods Cases

Selecting, Scraping, and Sampling Big Data Sets from the Internet: Fan Blogs as Exemplar

Book Title: SAGE Research Methods Cases
Chapter Title: "Selecting, Scraping, and Sampling Big Data Sets from the Internet: Fan Blogs as Exemplar"
Pub. Date: 20151120
Access Date: February 01, 2015
City: London
Print ISBN:
Selecting, Scraping, and Sampling Big Data Sets from the Internet: Fan Blogs as Exemplar

Learning Outcomes

By the end of this case study, you should
• Be comfortable with making multiple important decisions about selecting, managing, and analyzing a large data set
• Know how to scrape meaningful data sets from the Internet
• Understand the importance of sampling from big data sets
• Understand how to employ random sampling in multiple phases of data selection and analysis
• Understand that grounded theory and thematic analysis can be completed with big data sets drawn from the Internet

**Project Context**

*Mad Men* is a contemporary critically acclaimed, award-winning television drama. This original series, appearing on the US cable channel AMC (American Movie Classics) is a popular culture phenomenon:

Its actors appear on the covers of magazines and on the late-night talk shows; the series is reviewed by respected movie critics in widely-read newspapers, and fans as well as cultural commentators talk extensively about the series on the Internet. (Webb, Chang, Hayes, Smith, & Gibson, 2012, p. 226)

Scholars have examined the series in a variety of ways, producing multiple books and articles (e.g. Dunn, Manning, & Stern, 2012). Audience reception theory argues that to understand the meaning in a mediated message (such as a television series), scholars must examine the meanings of that message as decoded by its audience members. Therefore, our research team (Webb, Chang, et al., 2012) posed the question, How do *Mad Men* fans make sense of this pervasive and well-received cultural phenomenon? To investigate the audience's perspective on *Mad Men*, we examined commentary on 11 *Mad Men* fan websites.

We claimed that a fan can be distinguished from an audience member by two behaviors:
1. a fan actively attempts to decode, understand, make sense of, and/or interpret media messages rather than simply taking in the message as a passive observer.

2. a fan faithfully participates in fan communities associated with the object of fandom, in this case the television drama *Mad Men*, rather than simply observing and processing the message as an individual.

Audience reception theory argues that scholars must understand the audience’s perspective of mediated messages to know how such messages were received, the meanings ascribed to them, and the communication (meaning transfer) that has occurred. We content analyzed a random selection of posts on 11 popular *Mad Men* fansites to discover how fans interpret the ‘media text’ *Mad Men*. Random sampling was used to draw a representative sampling of the fan comments. Fans conversations centered on five elements of sense-making:

- attributions (statements about the causes of behavior)
- characterizations (providing a description of a person or object of interest)
- speculations (wondering whether a thing is so)
- predictions (projecting what will happen in future episodes)
- analysis and interpretations (offering a framework for understanding how a person or object of interest functions)

Our thematic analysis revealed that fans engage in multi-faceted interpretations commenting on the show as a whole, its characters, their relationships, the show’s plot, and *Mad Men* and the outside world. Such commentary on plots, the characters, and the relationships between them demonstrate that fans actually interpret viewed episodes of *Mad Men* within their fansite communities. In sum, our study demonstrates fans’ active and interpretative role in making sense of *Mad Men*.

In this case study, we explain our methods for studying fandom (hereafter called simply ‘the *Mad Men* project’), using it as an exemplar of three important research skills:

- how to select and scrape (transfer in its raw form) data directly from the Internet
- how to employ random sampling to select representative text for qualitative analysis
how to employ grounded theory analysis to analyze ‘big data’

Identifying the Research Design

Harvesting ‘big data’, or vast-sized data sets, from the Internet is a common procedure in studies of online text (Webb & Wang, 2013b). Big data typically share four traits: the data are unstructured, growing at an exponential rate, transformational,¹ and highly complicated (Webb & Wang, 2013b). Very large² data sets downloaded from the Internet fit this definition.

Free and inexpensive software is available to scrape big data sets from the Internet and also to analyze such data; nonetheless, a 2013 meta-analysis by Webb and Wang revealed that most data downloads and analyses are still done ‘by hand’. This case study describes an example of such a study (Webb, Chang, et al., 2012). A second analysis of the fan blog data with a differing set of research questions yielded a second meaningful set of conclusions (see Webb, Thompson-Hayes, Chang, & Smith, 2012). Both of these studies employed the qualitative methodology of thematic analysis, but quantitative analyses of big data sets from the Internet are also common (Webb & Wang, 2013a) and often yield very interesting and meaningful findings (e.g. Webb, Fields, Bouphe, & Stell, 2012).

Finding Appropriate Websites for Downloading

When downloading a big data set, the researcher can define the population quite narrowly such that a census or near census is possible. A census offers the advantage of including all known cases of the phenomenon under study in the potential data set. A list of such cases can serve as the sampling frame (i.e. the list of cases from which the researcher selects cases for inclusion into the sample).
In our Mad Men project, we defined the population as commentary from all fan websites. To obtain a census of such data, we identified appropriate web locations for obtaining the desired data using the following procedure:

- four online websites report Internet usage (e.g. The 20 Most Popular Websites 2009; Dosh Dosh Blog 2009; Saddington 2009; Top 20 Sites & Engines 2009). Together, these four websites identify and confirm the 20 most popular web locations used by the US population.
- using the simple search term ‘Mad Men’, we searched 7 of the 20 websites —specifically the websites that a user might rationally employ to search for fansites (i.e. Bing, Facebook, Google, Google blogs, Myspace, Yahoo, and YouTube).
- we then examined the results to discover web locations for data scraping.

Deciding the Number of Website Pages for Analysis

Our first challenge was deciding how many pages of output from each of the aforementioned seven websites we should examine in an attempt to identify all the Mad Men fan websites. No norms or standard practices have emerged regarding the number of website pages typically examined in a data collection project. Indeed, previous examinations of websites vary in the number of pages examined per website (Spyridakis, Mobrand, Cuddihy, & Wei, 2007; Yoon & Joseph, 2008). In our Mad Men project, we elected to examine only the first three pages of information from each of the seven websites (21 pages of text in total) for the following reasons:

- 1. the number of hits yielded by website searches varies according to the type of website (e.g. social networking websites vs large search engines); hence, not all websites produced more than three pages of results. Examining three pages of results did not over-represent the information yielded from large search engines (such as Google or Bing). Had we examined 10 pages, for example, from each website, only Google and Bing would have had 10 pages of data, and thus, the data from these two websites would have dominated the sample. Likewise, our procedure did
not under-represent websites with fewer pages of results (such as eBay or Myspace). When certain websites yielded fewer than three pages of hits, we examined the yielded pages.

- 2. around 90% of search engine users will click on information contained within the first three pages of search results (iProspect, 2006). Thus, information beyond the first three pages is not likely to be viewed by Internet users and would not contain information typically examined by the general public.

Focusing on the Object of Study

As we examined the three pages of output from each of the seven websites, our challenge became ascertaining which links led to Mad Men fan websites. In other words, we needed to decide upon a clear definition of our object of study, thus allowing us to eliminate Internet ‘clutter’ that should not be included in the data set as well as to recognize and include appropriate data in the study. For our project, the object of study was commentary on Mad Men fan websites. Therefore, at this stage of data collection, we were interested in identifying only websites that contained such fan commentary. Here is how we achieved that goal:

- we clicked on each link and examined its website to discover whether the website allowed fan comments.
- we retained only links to websites that allowed fan comments and had posted comments. This screening process yielded 30 websites.
- next, we more closely examined each of the 30 websites by clicking on the link and thoroughly examining the websites to discover the fansites. We did this through a process of elimination; specifically, we excluded the following: duplicate web addresses, websites where the television show Mad Men was not discussed, websites where fans commented on mass media stories about Mad Men rather than on the show itself, web addresses to blogs that focused on matters other than Mad Men and garnering fewer than 10 posters in response to the occasional blog entry on Mad Men.

After this process of elimination, only links to 11 websites remained. These 11 fansites included websites devoted exclusively to Mad Men fan commentary, discussion boards
devoted to *Mad Men* on larger websites, and blogs that focused primarily on *Mad Men* and attracted more than 10 posters. These websites constituted our sampling frame and are listed in Table 1.

**Table 1. Sampling frame and random selections for sample inclusion.**

<table>
<thead>
<tr>
<th>Website</th>
<th>Number of files</th>
<th>Number of pages/words</th>
<th>In the sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Basket of Kisses: The Mad Men Blog that Missed You <a href="http://www.basketofkisses.com">http://www.basketofkisses.com</a></td>
<td>5</td>
<td>71/25.436</td>
<td>5.5</td>
</tr>
<tr>
<td>3 <a href="http://www.3mp.com/spo/fan/mad-men">http://www.3mp.com/spo/fan/mad-men</a></td>
<td>1</td>
<td>2/285</td>
<td>3</td>
</tr>
<tr>
<td>5 <a href="http://www.3dsmashlounge.com">http://www.3dsmashlounge.com</a></td>
<td>5</td>
<td>83/29.541</td>
<td>5.2</td>
</tr>
<tr>
<td>6 <a href="http://www.madmenshow.com">http://www.madmenshow.com</a></td>
<td>5</td>
<td>54/15.888</td>
<td>6.3, 6.4</td>
</tr>
<tr>
<td>7 <a href="http://www.MSNBC.com/officias/AMC/website">http://www.MSNBC.com/officias/AMC/website</a></td>
<td>20</td>
<td>105/2325.226</td>
<td>7.1, 7.11, 7.2D, 7.3A, 7.4B, 7.5A, 7.5A &amp; C</td>
</tr>
<tr>
<td>10 <a href="http://www.WatchandTuneOnline.com/forum">http://www.WatchandTuneOnline.com/forum</a></td>
<td>1</td>
<td>7/539</td>
<td>10</td>
</tr>
<tr>
<td>Total</td>
<td>50</td>
<td>28,304 pages</td>
<td>18 files, 721 pages, 209,674 words</td>
</tr>
</tbody>
</table>


### Assigning Downloading Tasks

After locating the 11 fan websites, we developed a plan for downloading the data from these web locations. Like most research teams faced with the task of downloading large
quantities of data, we divided the task across our members. Random assignment of
downloading tasks avoids the introduction of bias as well as the appearance of bias.
The notion of bias in downloading may seem odd, but human beings rarely engage in
any human behavior in exactly the same way. Had only one research team member
completed all the downloading, then his or her way of completing the task would have
permeated the data set. Using four of the five research team members who may
have employed four slightly different ways of downloading minimized the possibility
of one method of downloading dominating the data set and thus consistently and
systematically introducing bias.

Here’s how we proceeded: in our Mad Men project, the senior scholar used a random
number table downloaded from the Internet to randomly assign the 11 fansites to the
remaining four research team members for data downloading. Random assignment
allowed individual differences in downloading to appear randomly rather than
methodically across the sample. This was the first of numerous times that we employed
random selection in our research methodology.

**Deciding What Specifically to Download from Websites**

Blogs and discussion boards contain vast quantities of posts (and accompanying
comments on those posts) that appear in reverse chronological order. The researcher
must decide exactly what text to download: every post and accompanying comments,
the most recent posts, the longest posts, posts on a certain topic of interest, and so
on. In the Mad Men project, we elected to download the longest strings of conversation
(the posts that garnered the most comments) because it offered the opportunity to
observe the longest and most detailed instances of fan interaction and discussion (i.e.
enactment of the fan community).
Scraping the Data from Websites

Days before the airing of the fourth season of *Mad Men*, in July 2010, we downloaded the five longest strings of conversation on the 11 *Mad Men* fan websites. Team members downloaded ‘by hand’ using the traditional procedure of highlight, copy, and paste to transfer an exact copy of information directly from the websites into Microsoft Word files. Team members created a separate file for each string, with two exceptions: If the entire collection of comments comprised fewer than five strings, then all strings were downloaded as one file. If the five strings were long, then the team member cut them into manageable chunks of text (approximately 50 pages) for saving into individual files for later analysis. Within 7 days, the team members cut and pasted the aforementioned web contents into Microsoft Word documents, thus creating a permanent copy of the sampling frame of the *Mad Men* fan commentary.

This procedure yielded a data set comprising 59 files available for analysis (see Table 1). This data set contained 28,301 pages of text (1,026,539 words). A data set this large demanded sampling if the team desired to employ ‘hand analysis’. Thus, we identified the data set as the sampling frame: ‘After a sampling frame is identified, the researcher can engage in “pure random sampling” meaning numbers can be drawn from a random number table and used as the basis for selecting incidences into the sample’ (Webb & Wang, 2013b, p. 101).

Selecting Data from the Big Data Files for Analysis

The goal of sampling is to select cases that accurately represent and portray the population of cases. A common and widely respected technique for obtaining a representative sample is simple random sampling. Researchers employ simple random sampling when the sampling frame is stable and can be examined at a fixed point in time, such as the *Mad Men* fan commentary days before the airing of the fourth season (the population for our *Mad Men* study).
In simple random sampling, each unit in the sampling frame (in this case, fan commentary) has an equal probability of selection into the sample. The researcher can offer two arguments for its representativeness:

- ‘selection biases are avoided’ (Kalton, 1983, p. 7).
- statistical theory predicts that the sample is likely to be representative of the sampling frame, if the sample is sufficiently large.

In our *Mad Man* study, four members of the research team each received five randomly selected data files from the downloaded data for coding and analysis only by that coder. This random assignment for coding was distinct from the random assignment for downloading and the second introduction of randomization into our methodology. Additionally, each coder was randomly assigned one file that overlapped with a randomly assigned second coder (a third introduction of randomization into our methodology), thus allowing for assessment of intercoder reliability. In total, each coder examined six randomly selected files. Coders were unaware of which files overlapped across coders.

Altogether, the coders received 18 of the 59 files (30.51%) for coding and analysis; these files contained data from all 11 websites and totaled 721 pages of text, containing 269,676 words. Table 1’s column on the extreme right displays the exact files in the random sample.

### Assigning the Coding Tasks

After each coder was randomly assigned six files, we faced the decision of whether the coders would examine everything in the files or a selection of text from each file. We deemed many of the files too large (i.e. over 50 pages) to ‘hand code’ all the text within each file. Given the size of the overall sample described above (721 pages of text containing 269,676 words), the coders randomly selected approximately 10 sequential pages from each file to code, thus examining approximately 50 pages per coder. This final random selection was the fourth time we employed randomization in our research methodology. The coded sections of each file were randomly selected using the following procedure:
• if files were 10 pages or less in length, then the coders read and analyzed the entire file.
• if the file was more than 10 pages, but less than 18 pages in length, the coder observed the page number of the last page of the file. If it was an even number, he or she coded the first 10 pages of the file; if it was an odd number, then he or she coded the last 10 pages of the file.
• if the file was longer than 18 pages, then the coders analyzed pages 9 through 18. We chose these exact page numbers based on two factors: drawing the number 9 from a random number table and our desire to sample approximately the same number of pages from each file, specifically approximately 10 pages.

Together, the research team coded a total of 180 randomly selected pages (.064% of the harvested data). All coding was completed within 1 week.

**Defining the Unit of Analysis and Themes**

A clearly defined unit of analysis allows coding to proceed smoothly and increases intercoder reliability. When both coders are analyzing units of the same size and type, they dramatically increase the odds of coding similarly. We defined our unit of analysis as the individual post or, for long posts, any individual idea expressed within a post. In accordance with grounded theory analysis (Glaser & Strauss, 1967), all themes emerged from the data, and we did not impose a priori categories on the data.

Coders identified common concepts (also called themes) that recurred across posters’ comments. Many researchers adopt Owen’s (1984) three criteria for a theme: recurrence (the concept appears across informants’ accounts), repetition (some and perhaps many individual informants mention the concept more than once), and forcefulness (informants state the concept in a way that emphasizes its importance to their understanding of the phenomenon under study). Like many researchers before us, we defined a theme as any idea that fit Owen’s criteria and was stated by three or more posters across one, two, or three files.
Enacting a Grounded Theory Approach to Thematic Analysis

We employed a grounded theory approach to coding (Glaser & Strauss, 1967). ‘Unlike its name would imply, grounded theory is not a specific theory of social science that describes or predicts the social behavior of human actors’ (Gibson & Webb, 2012, p. 160). Instead, the term ‘grounded theory’ references a genre of social scientific theory (specifically, theory derived from data); the term also references the research methodology used to develop such theory. Such a ‘methodology helps researchers conduct natural observations for the purpose of discovering emergent insights that can lead to the development of new theory of social behavior’ (O’Conner, Rice, Peters, & Veryzer, 2003, p. 353). Grounded theory research allows scholars to develop a basic understanding of a phenomenon by allowing informants' ideas to emerge from their accounts—in our case, fans' ideas about Mad Men emerge from their online commentary on fan websites.

In the Mad Men project, our coders located themes by reading the sampled portions of the files twice, taking notes on potential and discovered themes. During a third reading, coders formally noted and labeled themes, entering them on a coding sheet. Coders conducted a fourth and final reading to discover all instances of the emergent themes and any negative evidence to contradict the theme. Coders completed a coding sheet, listing their emergent themes and, for each theme, the following elements: characters, relationships, and story events mentioned by the fans, as well as the season of the discussed episode and the episode name, if that information was provided in the posts or on the fansite under analysis. This procedure yielded 76 themes.

Assessing Intercoder Reliability

Recall that each coder was randomly assigned to code one overlapping file with another coder. Within a week of completing coding, the pairs of coders met via video-chat to discuss their findings regarding the common file they analyzed. In each case, agreement exceeded 90%, and all disagreements were settled amicably via discussion.
Developing Supra-Themes

Grounded theory analyses often involve the development of ‘categories of concepts’ or ‘clusters of conceptual codes’ (Gibson & Webb, 2012, p. 165). This level of analysis is sometimes called axial coding and involves the development of supra-themes. The senior scholar on our Mad Men project who had completed no prior coding of the data set examined the 76 themes. Using a grounded theory approach, she developed two additional layers of axial coding across the 76 themes: 16 theme categories as well as five supra-themes. The results are displayed in Table 2.

Table 2. Topics of discussion on Mad Men fansites.
<table>
<thead>
<tr>
<th>Supra-theme</th>
<th>Theme category</th>
<th>Theme from extract</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Show as a whole</td>
<td>Praise for the show</td>
<td>Mad Men is my favorite show. Don is a great character. Discussions of the aesthetics for various sets. The episode ‘The Cypresses and the Hedges’ was so well written, amazing! Actor playing salesman January Jones has stage presence.</td>
</tr>
<tr>
<td></td>
<td>Discussion of artistic choices</td>
<td>Use of color on the show. The best line in the show is … Sets are true to the time period, with exception of IBM typewriter.</td>
</tr>
<tr>
<td></td>
<td>Time-line of the show</td>
<td>What happened between episodes? What about the missing years? Skipping ahead can cause inconsistencies and confusion. This show looks at a bygone era.</td>
</tr>
<tr>
<td>2 Plot development</td>
<td>Attributes</td>
<td>People are not always ready for parenthood; this explains the behavior of both Peggy and Betty. Don is getting increasingly reckless in his affairs. I think the characters have begun to work together for liberation; perhaps we are seeing the beginning of feminism on the show.</td>
</tr>
<tr>
<td></td>
<td>Speculations</td>
<td>Is Peggy pregnant? What will Betty do with the box? Is Betty’s baby really Don’s? Will Don and Betty’s divorce be disruptive to the children? Did Peggy’s sister adopt her son or is she just missing him? Was Betty justified in leaving Don?</td>
</tr>
<tr>
<td></td>
<td>Predictions</td>
<td>Don is good at what he does and loyal to his company; he will not be fired. I try to predict what will happen in an episode based on the title given in the TV Guide. Developments predicted for Season 4: What characters will be written out in the next season?</td>
</tr>
<tr>
<td>3 Characters</td>
<td>Attributes</td>
<td>Betty is seeking parental validation. Don destroyed the air force as a way to exercise power. Don has a guilty conscience.</td>
</tr>
<tr>
<td></td>
<td>Characterizations</td>
<td>Peggy is a mystery. Pete just doesn’t get it: he can’t recognize a good idea. Betty seems unable to separate dreams from reality. Don cannot be trusted. Betty is being irresponsible for smoking during pregnancy.</td>
</tr>
<tr>
<td></td>
<td>Analysis and interpretation</td>
<td>Like it in the US more than England because the US has fewer classes. Peggy is a self-made woman who knows advertising. Viewers would understand Betty’s motivations better if the show provided more information on her background and upbringing.</td>
</tr>
<tr>
<td></td>
<td>Speculations</td>
<td>Is Ken gay? Was Betty a prostitute with Olm? Is Don unfaithful? Is Suzanne attractive? Why does Don continue his sexual exploits? What physically about did Don suffer as a child? What secrets is Don hiding from a previous life (as deduced from the scene with the prison guard)? Can Betty continue to be a mother and a lover or will she need to get a divorce?</td>
</tr>
<tr>
<td>4 Relationships between the characters</td>
<td>Attributes</td>
<td>Betty is the product of her domineering father and invaliding mother. When Don drops his act and reveals a bit of his true self to Betty, it creates tension between the couple, as Betty needs to retain control of the situation. Betty gets nervous from being a mother and breast-feeding the baby.</td>
</tr>
<tr>
<td></td>
<td>Characterizations</td>
<td>Don and Betty are equally dysfunctional. Rogers is being more sensitive to Joan. The Betty-Henry...</td>
</tr>
</tbody>
</table>
Conclusion

This case encourages researchers who study online textual data to select narrowly defined populations of text for analysis. After deciding on a precise population, researchers can gather big data sets by scraping the entire population of the selected text from the Internet. Scraping can be accomplished using software designed for that purpose or by using the low-cost technique of simply cutting and pasting text from the Internet into Microsoft Word files. After capturing a population of data, the researcher can systematically sample the data set using well-established sampling techniques, such as random sampling, to draw representative samples. This case provides one model for assessing intercoder reliability across multiple coders and for conducting thematic analysis using grounded theory methods. Such techniques are useful for discovering how lay people in general behave on the Internet and how distinct populations of Internet users, such as specific fan blog communities, perceive objects of discussion as well as the specific topics they write about in such discussions. Additional analyses, not discussed in this case study, are possible using the same data set; such analyses include examining the interaction processes between users as well as discovering which topics draw commentary (and, therefore, might be perceived as important and contested) versus the topics that receive little or no discussion (and, therefore, might be perceived as unimportant, garnering little attention).

The Internet provides researchers with unfettered access to big data sets. By applying tried and true sampling techniques, researchers can select samples that represent the population under study and thus allow for broad generalizations. By using tried and true data-analytic techniques, researchers can draw accurate conclusions from their data. Together, such methods provide researchers with the tools to gain in-depth understandings of users’ online behavior.

Notes

1. By transformational, we mean that the format and structure of the data can change and evolve as new data are added across time.
2. Exact number of data points can vary, but we consider a data set ‘big data’ when its size exceeds 5–10 times the size of the data sets that are gathered via traditional means beyond harvesting data from the Internet.

## Exercises and Discussion Questions

- 1. The *Man Men* research team completed their data collection at one point in time, specifically a few days prior to the beginning of the fourth season of *Mad Men*. Given that the content of any website is in constant flux, what limited conclusions can be drawn from examining the content of blogs at a given moment in time? How might a study of blog content be improved with additional data collections across time but from the same websites?
- 2. Many audience members never log onto a fan website. Others visit such websites but never post; that is, they read without commenting. Given that the *Man Men* research team studied only comments from fans willing to post their opinions, how might the studies’ results be accurately framed in terms of their representativeness of audience members? What additional research would be necessary to understand how most or even typical audience members view this or any mass media message?
- 3. Given that the *Man Men* research team examined opinions of only audience members willing to post on fan websites, are the conclusions that can be drawn from such a study too limiting to justify the research? What valid conclusions can and cannot be drawn from the data collection described in this case?
- 4. What ethical concerns do you have, if any, about the *Mad Men* project, given that the researchers did not ask permission of the posters to analyze their commentary?
- 5. What aspects of blogs and discussion boards should or could be analyzed? The *Man Men* research team analyzed the longest strings, thus privileging controversial and/or complex topics. What knowledge can or should be gained from examining comments that never garner a response from fellow bloggers or that garner limited conversation? Conversely, rather than randomly selected comments or strings, what insights, if any, can be gained from systematically sampling comments, such a looking at every
comment that mentions a particular character or theme raised in the drama such as infidelity or workplace discrimination?

6. The *Man Men* research team developed a sampling frame by examining only three pages of search output from key websites. What additional justification could be added for this selection? Should they have reviewed more or fewer pages; why or why not?

7. What additional evidence of intercoder reliability could the *Mad Men* research team offer?

References


http://dx.doi.org/10.4135/978144627305014558403