To my family—Bob, Dorian, and Quinn, all of whom contributed to the 2nd edition in their own way.

In memoriam—In memory of my colleague and friend Paul D. Skalski, PhD, whose contributions were many and whose spirit will never fade. His substantial abilities, enthusiasm, and support were essential to both editions of this book.

The Content Analysis Guidebook
Second Edition

Kimberly A. Neuendorf
Cleveland State University

SAGE
Los Angeles | London | New Delhi
Singapore | Washington DC | Melbourne
Content Analysis in the Interactive Media Age

by Paul D. Skalski, Kimberly A. Neuendorf, and Julie A. Cajigas

This chapter addresses the implications of interactive media for content analysis. In the years since the publication of the first edition of this text, interactive media technologies and applications have become a ubiquitous feature of daily life. They span interpersonal, organizational, and mass levels of communication. Their use is stimulated and enhanced by such popular manifestations as touch-screen mobile devices such as Apple’s iPhone and iPad, motion-controlled video games such as the Nintendo Wii and Microsoft Kinect, and social-media platforms such as Twitter, Facebook, Instagram, and Snapchat.

Interactive media are rapidly replacing traditional media and modes of communication such as newspapers, magazines, old-school television, and even the traditional telephone. For example, the hypothetical, personalized daily newspaper of the future, the Daily Me (compiled and received electronically, as predicted by MIT Media Lab founder Nicholas Negroponte in the 1990s), has pretty much come to pass with online newsfeeds and delivery systems that reinforce users’ specific interests. In the future, researchers may not content analyze stories in daily newspapers, but they will need to analyze the universe of individuals’ Daily Me news content.

Facebook has become the leading interactive media content generator following its dramatic ascension in popularity after being opened to the public in 2006. By the third quarter of 2011, U.S. Internet users were devoting more time to Facebook than to any other web site—a whopping 53.5 billion minutes a month (Nielsen, 2011). In 2013, Facebook accounted for more than 10% of the total time Americans spent online (Weigley, 2013). In 2014, Facebook still ranked number one, with around 71% of Internet users in the United States logging on (Duggan et al., 2015). Social media platforms may rise and fall in popularity, but interactive and social media appear to be here to stay, and they have had a dramatic effect on life globally. In the aftermath
of the Japanese earthquake of 2011, for example, the social media of the world followed and documented the crisis, aided in recovery efforts through online applications such as Google’s Person Finder, and, as noted on cable network MSNBC, “The survivors will never forget the terror of that day, and thanks to the growing power of social media, it will also become a part of the collective record preserved for future generations, 140 characters at a time” (Snow, 2011).

In order to content analyze interactive media content, it’s important to understand that interactive media users are more than just receivers or consumers, as they were with earlier media. They have an active role in adapting, altering, and even producing content. The interactive media revolution that began with video games in the 1970s, continuing with home computers in the 1980s and the Internet in the 1990s, has grown and evolved in the early 21st century into what has been dubbed the Web 2.0 revolution.

The term Web 2.0 is attributed to technology guru Tim O’Reilly (2005), who used it to refer to changes happening to the Internet after the dot-com investment bust in 2001. During this time, sites such as Wikipedia, Google, and personal blogs burst onto the scene, revolutionizing content generation on the Internet. Web 2.0 platforms drastically changed the ability of the average user to generate online content by including interfaces that require little or no technical knowledge. The result was an immediate shift from professional writers, editors, and other gatekeepers as the sole producers of mediated content toward the empowerment of nonprofessionals to use sophisticated media for the promulgation of messages. Potter (2011) nicely sums up what most see as the core characteristics of Web 2.0:

Web 2.0 is a perspective about the Internet that fosters a social dynamic where people have the freedom to share their work through all sorts of open web sites. People are free to access all of these sites, use what they want, create their own messages, and make their messages available to anyone. The easy availability of these collective resources celebrates open participation, and this results in an enormous increase in creative activity. (p. 213)

Corresponding to the increase in creative activity described by Potter is an enormous increase in mediated content. Web 2.0 and other digital advances, coupled with the wide diffusion of high-speed wireless Internet connectivity, give today’s media user rapid access to and unprecedented power over content. Importantly, the various message functions outlined in Chapter 1—for individual, interpersonal, group, organizational, and mass purposes—are all readily found online, in unprecedented volume, and with unprecedented access. Private interpersonal communications, once studied only in the laboratory or via participant observation in the field, are now archived and available through the content of social media. Public organizational messages, such as corporate responsibility statements, once only
available to and targeted to a select group of employees and stockholders, are found openly on virtually all corporate web sites. And mass messages such as TV programs and commercials, once available to a researcher only if recorded at the time of airing, are now heavily archived for on-demand online access and are immediately accessible even “on the run” via mobile devices such as smartphones and tablets.

Indeed, content has never been so readily available. The Internet has swelled to contain literal libraries of information, and not just of printed content, but of audio and visual material as well. In 2015, YouTube reported that an average of 300 hours of video was uploaded to its site each minute, and more than 1 billion people in 75 countries are users (www.youtube.com, 2015). Streaming audio and video services now account for approximately two thirds of Internet traffic in North America during peak usage periods, led by Netflix, which alone gobbles a third of all traffic (Reisinger, 2012). Print media also continue to thrive in an interactive media age. According to a 2014 Pew Research Poll, 42% of U.S. adults owned tablets, and 32% owned dedicated e-readers, such as Kindle Fire. Of these adults, 42% had read an e-book or listened to an audiobook in the past year (Zickuhr & Rainie, 2014). The Google Books project has digitized more than 20 million titles (Howard, 2012), with the goal of digitizing all of the books in the world by the end of the decade (Jackson, 2010).

In addition to traditional media content such as books and videos, the Internet also contains interactive messages. Much of this newer content appears in ways that make its analysis and even capture difficult. Whereas traditional media content came in “fixed” forms such as the newspaper page, the television episode, or the motion picture, the dynamic or fluid nature of interactive media content, ranging from how users play a video game to what they choose to post on Facebook, makes it much less tangible. Web pages and other interactive content may also be updated moment by moment by their creators. Snapchat and Periscope are platforms built to maximize live and ephemeral communications. Snaps, the slang term for messages on Snapchat, last mere seconds and then self destruct. Should a user attempt to capture them with a screenshot on their mobile device, the user who sent the message is notified. Periscope, a live streaming app that allows users to broadcast live events worldwide, archives videos for only 24 hours. Users can save the videos to their mobile device, but in terms of public viewing, the clock is ticking. In addition, Web 2.0 platforms and templates themselves are often changing or evolving, sometimes more rapidly than users can adjust. This poses considerable challenges to the method of content analysis, a problem recognized at the turn of the 21st century by Sally McMillan, who wrote about the difficulties of applying the “microscope” of content analysis to the “moving target” of the World Wide Web (McMillan, 2000).

With the proliferation of online content, researchers are now faced with massive data sets generated by user behavior. As noted in Chapter 5, big data is a term used to describe data sets that are too voluminous or complex
for traditional methods of analysis. In the realm of content analysis, big data often take the form of information produced by human behavior and collected and archived by the programs behind social media platforms, websites, and mobile media applications (Lewis, Zamith, & Hermida, 2013). We now have the ability to search, aggregate, and cross-reference large data sets from a variety of interactive platforms, giving researchers the ability to overcome traditional sampling and coding limitations (boyd & Crawford, 2012). However, by definition, big data implies that the data are too big and complicated to handle or even be fully conceived by humans—computer power must be employed to collate, massage, and analyze. Thus, big data are removed from human experience, so only gross summarizations of the outcomes of analyses can be comprehended, making the implications of findings on big data rather abstract and not always directly applicable to human experience.

There are ethical implications in scraping and analyzing data about human behavior and communication, as well as challenges in acquiring and analyzing data sets that can’t be managed by human action. Scientists in a variety of disciplines have expressed concern over big data due to the fact that its use may ignore principles of representative sampling and of the deductive process of scientific investigation (big data analyses typically take an inductive, data-dependent approach to things). Lewis et al. (2013) argue for a hybrid approach that blends computer text analyses and manual methods in order to “preserve the strengths of traditional content analysis, with its systematic rigor and contextual awareness, while maximizing the large-scale capacity of big data and the efficiencies of computational methods” (p. 47). This notion is consistent with the observations in Chapter 5 that CATA and human coding can and should be used in complementary fashion; this seems particularly apropos for dealing with big data. The criticisms of big data should be kept in mind when considering content analyses of social media and other subscription-based media (e.g., Twitter) in that many such investigations do use “very big data.”

The shifting nature of interactive media does not make their content analysis impossible, however. There are a number of parallels between traditional and interactive media content and ways to deal with the differences between the media forms, making content analysis methodologies applicable to even the most challenging of messages. Since the release of the first edition of this book, a great variety of types of content analyses of interactive media has been conducted; some interesting and wide-ranging examples are itemized in Box 7.1. Moreover, newer interactive media empower the content analyst in exciting, never-before-possible ways, at the stages of creating, acquiring, archiving, and coding content. This chapter reviews considerations for each of those stages and offers recommendations for the content analysis of interactive media and other new content forms, as well as for the use of interactive media to facilitate content analyses of all types.
Box 7.1 Interactive Media and Content Analysis

The following are examples of interactive media topics that have been studied using content analysis in the first years of the 21st century, demonstrating the range of new topics that have been addressed:

**Web Sites**

- The rise of event-driven news (e.g., stories captured with a videophone) between 1994 and 2001 (Livingston & Bennett, 2003)
- Position-taking and issue dialogue on campaign web sites during the 2002 election cycle (Xenos & Foot, 2005)
- Rationales provided on web sites posting DVD decryption software (Eschenfelder, Howard, & Desai, 2005)
- Media profiles of living and dead public intellectuals on the Internet and in traditional media (Danowski & Park, 2009)
- Internet coverage of college basketball’s March Madness (Kian, Mondello, & Vincent, 2009)
- Health promotion appeals on U.S. and Korean web sites (Baek & Yu, 2009)
- Privacy policy statements on health web sites (Rains & Bosch, 2009)
- Campaign information as unmediated messages on candidate web sites (Druckman et al., 2010; Druckman, Kiwer, & Parkin, 2010)
- New forum comments on TV and newspaper web sites (Hoffman, 2015)
- [Corporate] web site localization strategies (Wu et al., 2015)

**E-Commerce**

- Online video game auctions on eBay (Wu & Neuendorf, 2011)
- The gender marketing of toys on the Disney Store web site (Auster & Mansbach, 2012)
- Online travel agencies as a source of hotel information (Peterkin, 2014)

**Social Networking Media**

- Social interaction on Usenet newsgroups (Turner et al., 2005)
- Emotion in MySpace comments as a function of gender and age (Thelwall, Wilkinson, & Uppal, 2010)
- Nonverbal displays of self-presentation on MySpace (Kane et al., 2009)
- Level of personal information posted by adolescents on MySpace (Patchin & Hinduja, 2010)
- Political Facebook groups during the 2008 presidential election (Woolley, Limperos, & Oliver, 2010)
- Fortune 500’s Facebook pages (McCorkindale, 2010)
- NCAA organizational sport pages and Big 12 Conference athletic department pages (Wallace, Wilson, & Miloch, 2011)

(Continued)
Companies cultivating relationships with publics on SNS in China and the United States (Men & Tsai, 2012)
Online social networking in discussions of risk on Facebook (Ledford & Anderson, 2013)
Facebook pages dedicated to moms (Kaufmann & Buckner, 2014)

Non-Social Networking Media

- Blogs as retrospective sources of public opinion and reactions to news (Thelwall, 2007; Thelwall, Byrne, & Goody, 2007)
- Blog use by libraries and librarians (Aharony, 2009; Bar-Ilan, 2007b)
- Postings on health and medical blogs (Buis & Carpenter, 2009)
- Visual motifs in Jihadi and Cholo videos on YouTube (Weisburd, 2009)
- Twitter during the 2009 H1N1 pandemic (Chew & Eysenbach, 2010)
- Sentiment in Twitter events (Thelwall, Buckley, & Paltoglou, 2011)
- The August 2011 riots in England on Twitter (Procter, Vis, & Voss, 2013)
- A cross-cultural content analysis of Twitter and Weibo (Ma, 2013)
- Childhood obesity on Twitter (Harris et al., 2014)
- Twitter during the Tunisian and Egyptian revolutions (Hermida, Lewis, & Zamith, 2014)
- Direct-to-consumer pharmaceutical ads on YouTube (Muncy, Iyer, & Eastman, 2014)
- World Cup 2014 and U.S. sports fans' tweets (Yu & Wang, 2015)
- Pinterest as a resource for health information on chronic obstructive pulmonary disease (COPD) (Paige et al., 2015)

Asynchronous Communication

- Mapping the communication behavior of the social network of Usenet (Turner et al., 2005)
- Responses to use of pseudonyms in email communication (open-ended coding; Heisler & Crabill, 2006)
- Chat reference service use by public library patrons (Kwon, 2007)
- Trust dynamics in work-oriented virtual teams (Kuo & Yu, 2009)
- Peer tutor behavior in asynchronous online discussion groups (De Smet et al., 2010)

Other Online Activity

- Link analyses of web sites (i.e., the patterns in which web sites are hyperlinked and the meaning assigned to these links; Harries et al., 2004; Thelwall, 2006)
- The impact of Google bombing over time (Bar-Ilan, 2007a)
- “Media catching”—emails and Twitter requests by news reporters seeking information from PR specialists (Waters, Tindall, & Morton, 2010)
Chapter 7  Content Analysis in the Interactive Media Age

Considering Interactive Media

Definitions of interactivity abound. During the time the Internet was emerging as a mass medium, attempts to define the concept among academics began en masse with work by scholars such as Sheizaf Rafaeli (e.g., Rafaeli & Sudweeks, 1997) and Sally McMillan (e.g., McMillan & Hwang, 2002). Some have argued that interactivity is a perceptual variable (e.g., Bucy & Tao, 2007), while others have focused on how it represents a transaction between a user and system (e.g., Sundar & Kim, 2005). McMillan (2002) delineated among three types of interactivity: user-to-system interactivity (as in video gaming and using a search engine), user-to-document interactivity (as when navigating hypertextual content online), and user-to-user interactivity (as on social networking sites), all of which have implications for the application of content analysis (Ramasubramanian & Martin, 2009).

In this chapter, we subscribe to a “functional view” of interactivity, which defines it in terms of the functions provided by a media interface, including features, attributes, processes, and outcomes (Sundar, Kalyanaraman, & Brown, 2003). Based on this definition, researchers have operationalized interactivity in terms of functional features such as email links, chat rooms, and the ability to download audio or video (e.g., Massey & Levy, 1999). Adopting the functional perspective grounds our interest in interactivity in the realm of media form and content rather than users, whose (perceived) interactivity would be more appropriately studied through survey, experimental, or participant observation methods. Referring to interactive media (instead of just interactivity) also cements our approach to the concept on the medium/message component of the communication process. There is still, however, much to be learned about users through interactive media, particularly since they play a major role in creating its content.

Content Creation in the Interactive Media Age

Historically, the process of message production has not been of direct concern to the content analyst. But in the interactive age, the creation of mediated

---

**Gaming and Simulations**
- Violence in popular video games (Smith, Lachlan, & Tamborini, 2003)
- Decision-making in a computer-simulated microworld (Elliott et al., 2007)
- Hypersexuality in video game characters (Downs & Smith, 2010)
- Advertising in popular video games (Lindmark, 2011)
- Racial stereotypes in video game magazines and covers (Burgess et al., 2011)
messages has become an issue of critical concern. Although the integrative model discussed in Chapter 2 recommends an empirical look at message producers or sources, this is technically a step beyond the methodology of content analysis itself, which strictly examines the content of messages. Regardless of who created a traditional media product, whether it was a single book author or a team of thousands working on a major Hollywood film, an underlying assumption was that the end product (e.g., the book or the film) existed in a fixed, objective form that could be documented. Further, the notion of message source had historically been stable. Whether a single author of a series of diary entries, a dyad in a doctor/patient interaction, a team of employees creating a corporate culture statement, or a reporter on a local TV news story, the sources of messages were clearly identifiable.

This assumption of easy identifiability of source and message no longer holds in the arena of interactive media, in which audience members have the “ability to shape their media environment” (Ramasubramanian & Martin, 2009, p. 114). Much of content creation today depends in part on choices made by a user while interacting with or within a medium. For example, the content of a video game depends on how each individual user plays the game. With a violent video game, some players may fire away and recklessly aggress against foes, while others may adopt a more strategic, stealthy approach. This makes coding for a construct such as “violent activities” more complicated in a video game than with traditional media.

We should note the parallel between this trend in interactive media and a key choice faced even when content analyzing traditional media. As described in Chapter 3, defining a population and drawing a representative sample from that population of mediated content may follow one of two philosophies: (a) an availability-based approach or (b) an exposure-based or use-based approach (i.e., “what's actually attended to or used”). An availability approach to the study of video game content would attempt to inventory and then sample from all possible gaming sessions/outcomes, a huge and perhaps doomed undertaking. An exposure/use-based approach, sampling from users’ actual gaming sessions, is the method that has been almost universally employed (Pieper, Chan, & Smith, 2009). Similarly, an availability approach to the study of sexually explicit web content would need to attempt to identify all such web sites. An exposure/use-based approach might do as Salazar et al. (2009) did in their panel study of 530 U.S. teens: They used proprietary software to collect and store all of the teens’ web traffic for a 30-day period, then identified which pages contained sexual content, and then executed a content analysis of those pages.

Other forms of interactive media pose similar challenges. How users experience information on a web page, for example, depends on where they choose to navigate, how frequently they click, other interactive choices they make, and their past online behaviors (e.g., when past surfing affects the insertion of tailored web ads). And a related challenge stems from the rise of Web 2.0, which puts users in control of content creation and deletion and has
resulted in a flood of media units and products available for analysis. The popular web site Facebook alone has, as of this writing, more than 1.5 billion active users, each with a personal page akin to a traditional media unit of data collection such as a newspaper article or television program. Within this personal page are a number of smaller units that might be analyzed, including wall posts, photo albums and their captions, reposted content from elsewhere on the web, personal information, and much more.

Another challenging complexity that researchers face is how to determine what content has been created by the profile owner using content generation tools and what content has been reposted from elsewhere on the web. With the proliferation of multimedia content generation tools and applications, it has become increasingly difficult to determine whether or not photo, video, and other multimedia content was created by the profile owner. Given widespread access to photo, video, and multimedia editing programs, which are generally geared toward allowing the unsophisticated individual to create high-quality content, it is not easy to assess whether or not particular users have the ability necessary to generate the content that might be found on their profile. Users often repost photos, memes, and other content that was created by other sources rather than linking to the content. Popular content tends to spread rapidly via social networking sites, making it increasingly difficult to determine the original source of any given update or post. In the end, when everyone becomes a creator and sharing is encouraged, the choices a content analyst must make can be overwhelming.

The challenges do not end there. The content of a web page may also change depending on what a user searches for, or even the user’s location, due to technology that detects the location of the user. This applies in particular to advertising content. Targeted ads are sent to web surfers and video game players alike. Dynamic video game advertising received widespread media attention in the midst of the 2008 presidential election campaign, during which billboards promoting Barack Obama were placed in the racing game *Burnout: Paradise* (Simons, 2008). Targeted ads varying by location, demographics, and time zone are the norm in certain games, which update the messages regularly through the Internet connections of players (Kaufman, 2006). With a very robust ad creation platform, Facebook enables marketers to target advertising based on location, age, sex, interest key words, and connection status. The lack of fixed content in computer-based new media eliminates the certainty that once existed with particular types of content.

**The Special Nature of Content in the Interactive Media Age**

To bring some order to this huge range of content creation options, content analysts should distinguish between *user-generated content*, *user-selected content* (including *user-curated content*), and *interactive media output*. All are
forms of content creation in the interactive media age that are ripe for content analysis, but all challenge our traditional notions of both source and message content. This three-part typology of interactive media content deserves further elaboration and exemplification:

1. User-generated content (UGC): User-generated content is often considered synonymous with Web 2.0 and marks a shift away from corporate-provided media content (Lanchester, 2006). It refers to content created and provided by users, through easy-to-use (and frequently corporate-owned) tools (Potter, 2011). The wealth of UGC online has been noted in popular media (e.g., Time magazine’s 2007 recognition of “You”—the online user—as “Person of the Year”), and its proliferation has been scrutinized for legal and policy implications by such bodies as the International Organization for Economic Cooperation and Development (OECD) (see Geist, 2007, and www.oecd.org/dataoecd/44/58/40003289.pdf).

Examples of UGC include Facebook profiles, Vine videos, Instagram photos and videos, Snapchat stories, web logs (blogs), emails, microblogging via services such as Twitter, Wikipedia pages, and many YouTube postings. In this chapter, we distinguish between UGC and user-selected content, defined in the following paragraph. Again, it should be remembered that interactive media accommodate all communication functions. That is, a significant portion of UGC is interpersonal in nature (e.g., emails and Facebook messages), other UGC is “personal” but available more widely (e.g., social networking activities, blogs), and some UGC is intended to be “mass” (e.g., Wikis and most YouTube videos). UGC is sometimes referred to alternatively as user-created content (UCC).

2. User-selected content (USC): User-selected content (or user-collected content) refers to existing media products posted and shared online, such as when someone uploads a clip from a TV show on YouTube or shares music on Bit Torrent. USC is distinct from UGC, according to our conceptualization, in that the nominal media source is merely an intermediary or broker between content and an interactive medium rather than the creator of content for the interactive medium. Another way that users can select, collect, and share content is through and across Web 2.0 applications. The phenomenon of “viral” content, such as viral videos, is the direct result of the ability to share and repost media content such as videos, photographs, and blog posts via Web 2.0 sites. For example, users browsing videos on YouTube can quickly and easily share a video with their social network by reposting it to their Facebook page or their Twitter feed or by designating it as a favorite video on YouTube itself. In addition, much of the content posted to blogging and microblogging Web 2.0 sites can be collected and shared using a content aggregator, such as an RSS feed.
What is unique about USC is that it is reflective of an individual’s message choices rather than message content generation. We are just beginning to consider the implications of the analysis of USC as an indicator of how people navigate and make sense of the tremendous volume of online content.

A particular type of USC is user-curated content. Rather than content that is selected based on the user’s enjoyment and reposted for others without much thought, curated content is selected and aggregated for the express purpose of meeting the needs of an audience. For example, Flickr, Tumblr, and Pinterest allow users to curate and present content that they have individually reviewed and selected with the intent of creating a useful, beautiful, or otherwise attractive content collection. Lin et al. (2009) analyzed curated content on Flickr, looking at group photo streams, which are connected by agreed upon themes. Etsy.com, a peer-to-peer e-commerce site, features the use of the “Treasury,” a curated selection of craft items for purchase created by thousands of Etsy.com users. A “treasury” is usually themed around a color or other thematic thread and showcases a user’s favorite items, almost like the editorial picks in a style magazine. These curated collections are categorized by Etsy.com based on “hotness.”

3. Interactive media output (IMO): Interactive media output refers to content created as media consumers navigate or use interactive media. Examples of IMO would include web-surfing patterns and how a player interacts with a video game. This type of interactive media content requires more advanced and active capturing methods than the other two in order to prepare for content analysis.

The content analyst must be clear on whether it is assumed that the message source is the media user (e.g., video game player) or the creator of the interface (e.g., video game designer). Either conceptualization is possible, and the choice will dictate the population under investigation. The content analyst should also consider the motivations users might have for sharing content in order to best understand their relationship to that content.

Practical Suggestions

Considering the challenges of working with dynamic interactive media content, what follow are some practical suggestions:

1. Be aware of standardized content on Web 2.0 pages when the goal is to analyze user-generated content. Users control only certain portions of Web 2.0 sites, despite the illusion of creative control. On YouTube, for example, users can post videos and control selected information such as tags, but other content, such as the related links, may be added by the site operators. The same holds for certain aspects of Facebook pages. Facebook users have no control over the ads that
appear on their home or profile page, even though such content may be attributed to them. Content analysts should also distinguish between content posted by a user to their page and content posted on the page by other users, as in the case of a Facebook Timeline or YouTube comments section.

2. On a related note, tags and other “meta” content on a page can be useful tools for content analysis. These may be created by the user or the site itself and can help identify similar content for coding. For example, many digital cameras embed photographs with metadata, which can include the type of camera, camera settings, and geographic location where the photo was taken. In another example, blog users may create tags to help categorize their postings into topical groups, allowing readers or content analysts to easily navigate their content.

3. Remember to distinguish between user-generated, user-selected, and interactive-output content. These distinctions become important, for example, when one wishes to assume that the poster of information on a social media web site is indeed the message “source,” which may be possible with UGC and IMO but is less likely with USC.

4. Finally, it should be noted that the templates provided by web sites may change over time, necessitating fluid codebooks to match the fluidity of interactive media content. In one ongoing analysis of YouTube comedy, the researchers were forced to revise the codebook repeatedly over time due to changes in the layout and options of the social networking system (Neuendorf et al., 2016).

**Acquiring Content in the Interactive Media Age**

Content has never been easier to acquire. The mass diffusion of the Internet has ushered in a “golden age” of content acquisition. The first edition of this book listed content archives that were often real-world locations, with physical media such as videotapes. Since then, a vast amount of content has been digitized and posted on the Internet, allowing for the easy uploading, downloading, and viewing of audio/video content alongside the text content that dominated the medium in its early days. These range from the venerable LexisNexis, which systematically and exhaustively collects digitized legal documents and print media content such as newspapers and magazines, to YouTube’s haphazard collection of audio and video content to the Internet Archive Wayback Machine, a collection of “classic” web pages dating back to 1996 (Waite & Harrison, 2007).

In addition to these online sources, other interactive technologies for acquiring content have emerged, including subscription-based streaming video services such as Netflix and Hulu and digital video recorders (DVRs). Another
major source for digital audio and video content is iTunes, which was created by Apple to complement the iPod music and video player device and has since blossomed to provide digital media for many platforms. Content purchased from the iTunes store is not streamed live, but is rather digitally archived on the hard drive of the user’s computer, which may offer advantages when compared with streaming media. Beyond archives of traditional media, there is also a large body of Web 2.0 content, which includes web logs, microblogging content, and social-media content. This content is created in the digital environment and easily acquired and indexed. The availability and access to digitized content is constantly increasing due to advances in technology. In fact, the available digital acquisition technologies are fast replacing physical acquisition media such as videocassettes, CDs, DVD-Rs, and handwritten journals; some are even being developed with the content analyst in mind.

One example is SnapStream, a server-based system for acquiring television content that includes networked-DVR capability for multiple work stations and “TV search engine technology.” The system is marketed explicitly as enabling content analysis. SnapStream users can record thousands of hours of video content, with cloud-sharing available, and can then search the digital archive by topic or key word through the closed captioning in the recorded sample without having to view the content. Simon (2011) used SnapStream to (a) record every local news television broadcast over a one-month period preceding the 2010 Ohio general election, (b) search for every instance of a news story that included mention of one of the seven state and county elections that were the focus of his study, and (c) download the closed captioned text for the relevant segments for use in CATA analyses. He devised his own set of search dictionaries, employing the open-source CATA program Yoshikoder to tabulate occurrences (see Resource 1).

Another example of a digital acquisition technology is Evernote, a software that facilitates the collection of a wide variety of types of messages, typically via use of a mobile device such as a smartphone. Box 7.2 describes this application in more detail. Evernote might be seen as perhaps the most tailor-made for content analysts, but it is certainly not the only application that might be of use in finding, screening, or even coding messages for content analytic studies. For example, a number of mobile apps are available for the instant recognition of music (e.g., Soundhound, Midomi, and Shazam). Apple’s iPhoto photo organizer, Google’s Picasa image organizer, and Facebook itself all have facial recognition apps.

The vastness of available content presents some challenges at the acquiring stage, particularly with regard to sampling. How can the content analyst sample from such an overwhelming pool of messages? For example, how should a researcher navigate individual users’ privacy controls on Facebook to acquire the content they need? And how much of a video game should be sampled, given that a play session does not have set time frames like the half-hour sitcom or two-hour movie? Following are some issues and suggestions for sampling common types of content in the interactive media age.
Box 7.2 Technologies in Archiving: Evernote

The challenges of working with rapidly changing media can be somewhat mitigated with technological advances in archiving technologies. Created in 2008, the popular Evernote software (www.evernote.com), available for desktop and mobile devices, is an example of such an advance in message archiving. Evernote allows users to collect content in various ways, including entering text, taking a photograph of handwritten or printed materials, taking a snapshot of a web page, and recording an audio clip. The content is then processed, indexed, and made searchable. If the content is a photograph of written or printed text, the program interprets that text and makes it searchable. This software has many potential applications in content analysis. For example, the web capture portion of Evernote allows the researcher to take a snapshot of a single tweet or Facebook update or the entire web page. It then catalogs that snapshot or page clipping with a date, time, and title and allows the user to enter key words for additional organization. Thus, the clipping can be retrieved and organized into a data set relatively easily.

Another example of the usefulness of such a program can also be seen in the following example:

A researcher wishes to analyze the content of billboards in her or his city. Evernote allows said researcher to take a photo of each billboard. The photos are not only processed for easy categorization and searching, but they are also embedded with location data that will allow the researcher to create a virtual map of the signage as it appeared around the city and calculate distances from other points and points of view. Further, Evernote is accessible from the desktop to the mobile device and synchs all archived content automatically.

Acquiring Video Game Content

Video games present challenging sampling issues at several levels (Pieper et al., 2009). As Schmierbach (2009) points out, video game content analyses not only require the selection of games for inclusion in a sample, as one might do with movies or TV shows, but also the selection of units within a game. Video games have no fixed time frame or time frames that extend far beyond traditional media experiences. Rockstar Games has estimated that their hit game Grand Theft Auto 4, for example, takes about 100 hours to finish (Yin-Poole, 2008). And within those 100 hours are undoubtedly widely varying types of content. Moreover, time frames may vary from player to player. A game session for a new player of the classic arcade game Donkey Kong may last only a few minutes, while world-record holders Billy Mitchell and Steve Wiebe typically play for hours at a time (Cunningham & Gordon, 2007).

The decision concerning how to select units within games has implications. Haninger and Thompson (2004) modeled the relationship between length of game play and type of content observed (e.g., violence, profanity, drug use, etc.) and found that a 10-minute sample of play likely leads to missing one
or more content types 40% of the time, while playing for an hour leads to missing only 10% of the time. Although this suggests that longer segments more validly represent a game, Schmierbach (2009) argues that the issue is more complex. Sampling the first hour of a game, he says, may misrepresent the frequency with which certain content types happen. Violent acts, for example, may increase as a game character acquires more powerful weapons or fights more enemies as a game progresses. Furthermore, important game modes, such as the multiplayer matches that are popular for players of series like Halo and Call of Duty, may be missed if only the default mode is selected. The game-sampling problem is further compounded by varying skill levels of players whose interactive output is typically captured as the “content,” and the players’ choice of character, which many games offer. A fighting game character with a spear, for example (e.g., Scorpion in the popular Mortal Kombat franchise), may commit more bloody and violent acts than one who uses hand-to-hand moves to battle foes (Smith, 2006).

Just as TV shows have Nielsen ratings and movies have box office figures, games have associated sales data that can aid in sample selection. Assuming the goal is to sample game content reaching the greatest possible number of players (i.e., a use-based approach to population definition and sampling), one can use sales data and select the top 20 or 30 games across different popular consoles, as some content analysts have done (e.g., Smith, Lachlan, & Tamborini, 2003). The NPD Group, a North American market research firm, provides game sales data. If the goal of a study is to generalize to the population of all games, however (i.e., an availability-based approach), it would be preferable to get a complete list of games and randomly sample, with perhaps some stratification across consoles (e.g., 20 games from Wii, 20 games from PS3, 20 games from Xbox 360) or based on other variables that address study questions.

The more difficult sampling decision concerns what within-game units to sample. If the goal of a content analysis is to capture what an entire game is like, then sampling only the first 10 minutes or some other early game play segment does not adequately represent what exists. A more valid approach might be a stratified sample of random time segments (e.g., 10 or 20 minutes) from the first third, middle third, and final third of a game. Or if a game has fixed levels (such as the four different boards in the arcade version of Donkey Kong), an appropriate approach would be to sample times from each to get a clear picture of what happens across noteworthy changes in a game. Determining what segments of a game are important might require the researcher to become immersed in the content (i.e., do a lot of playing of the game).

There may be differences between console gaming and computer gaming that are important to the content analysis task; a console functions differently than does a computer. Although with the increasing generations of consoles (e.g., as of this writing, new consoles are of the eighth generation—PS4, Xbox One, Wii-U), it looks as if developers are trying to make multipurpose console
entertainment systems into almost user-friendly computers. This factor should be taken into account when planning a study of video games and will have important implications for sampling (e.g., will the study be of one generation console only, or a comparison across generations?).

Other important considerations for video game content analysis include character selection, player skill level for the participants who will generate the game play sessions, and related message content.

1. Since different characters may have different powers and abilities affecting the interactive media output they produce, there are several options. One is to choose the most popular characters, since they are the ones players are most likely to experience. Popularity data may be available online for games, or it can be determined through a presurvey. An alternative approach is to randomly sample from the available characters, making sure at least 20% are represented. If the study is about character attributes (such as sex, race, body type, etc.), a larger sample is recommended.

2. As for the player skill question, experienced players are most likely to approximate how content would manifest in the population of players. If unskilled players are used, they should at least go through a training period before their content is captured, or they are unlikely to generate much usable content. There are alternate views on this, however, including the notion that a random sample of people from a population of interest should be engaged. Newman (2004) recommended tracking novice players longitudinally as they become better. Smith (2006) suggested having 10 experts and 10 novices play a game and looking at between- and within-group variability. Schmierbach (2009) recommended using a large number of players to produce content, with careful training for novices.

3. If games have multiple play modes, such as single-player and multiplayer campaigns, we recommend sampling content from the most popular modes, which can again be determined through online data or a presurvey. As Schmierbach (2009) notes, sampling, for example, only Halo 3’s single-player campaign misses a lot of the game’s content that would be experienced by players, since much of the title’s popularity stems from its multiplayer modes.

4. For online computer gaming, it is possible to record game play data and also messages that are routinely exchanged among multiple game players during play. This may be achieved by buying server space for a game or by obtaining gaming data from game hosters. As with other online activities, there may be terms to joining the server or online service, so consent must be secured from the players in order for the play and message data to be used for research purposes. The intergamer interactions seem to be a promising source for future content analyses, especially
when linked with game play data. This is demonstrated by Lehdonvirta et al.’s (2012) study of players of UWO, a Japanese massively multiplayer online game (MMOG) in which the researchers collected user-to-user communications along with player data. They found that when using male avatars, players were less likely to receive sought-for help but more likely to receive indirectly sought help than when using female avatars. Since these differences were independent of actual player sex, the researchers concluded that men “overcome their inhibition for help seeking when using female avatars” (p. 29).

Ultimately, we agree with Sherry (2007), Smith (2006), and others in believing that the choices a researcher makes should be driven by the theoretic processes under investigation. Sampling is probably more difficult with video games than with any other type of content, given the sheer number of considerations and possibilities, but carefully weighing options based on solid theoretic principles and expert advice should lead to sound decisions.

Acquiring Web 2.0 Content

As mentioned earlier, the Web 2.0 revolution has brought with it a deluge of content. While exciting in some ways, it also may seem overwhelming from the standpoint of sampling. A study of Facebook profiles, for example, has hundreds of millions of sampling units to choose from.

Or does it? The first and most important consideration to make when sampling and acquiring content from a Web 2.0 site is to consider the various platforms and their capabilities. With the breakneck pace at which Web 2.0 evolves, it would be impossible to give a comprehensive guide to every platform and program that the content analyst might like to study, especially into the future. But regardless of the platform or the program, there are a number of questions researchers should ask themselves before sampling, unitizing, and acquiring content. First, what is the function of this platform? Is it intended to broadcast a message to a large, undifferentiated audience, or is it meant to share personal messages with a smaller-known network (of friends, family, etc.)? Second, how does the Web 2.0 platform achieve its goal? Does it use public messages only, or are there interpersonal and private message functions? Third, does the platform allow users to create content, to share collected content, or both? Fourth and finally, has a similar platform been studied in the past, and if so, how was it studied at that time? The answers to these questions are invaluable for guiding content acquisition, especially in new or emerging media.

When discussing content acquisition in Web 2.0, it is important to note that Web 2.0 can be broken into several subcategories. Mazur (2010), for example, distinguishes between social networking sites and blogs and presents content analysis issues related to each. We agree that social networking sites are a distinct type of Web 2.0 content; however, blogs are just one of
many other types of social media. In an attempt to be as comprehensive as possible, we divide Web 2.0 content into two primary types: social network sites (SNS) and those that are not SNS (non-SNS). While these categories are not mutually exclusive, the distinction is helpful in the discussion of content acquisition:

1. Social network sites (SNS): SNS were defined by boyd and Ellison (2008) as “Web-based services that (1) allow individuals to construct a public or semi-public profile within a bounded system, (2) articulate a list of other users with whom they share a connection, and (3) view and traverse their list of connections and those made by others within the system” (p. 210).

Similar to other Web 2.0 sites, SNS allow users to publish both UGC and USC. Unlike other Web 2.0 sites, SNS are not designed to reach a large, undifferentiated audience, but instead, have highly specialized privacy tools, which allow users a great deal of control over who can and cannot access their content based on their connection status. Investigations of social network sites such as Facebook, MySpace, and LinkedIn have been growing in popularity as the focus of inquiries using neutral loci for the study of unfettered, natural communication (Compton, 2008; Grasmuck, Martin, & Zhao, 2009; Jones et al., 2008; Kaufmann & Buckner, 2014; Kim, Klautke, & Serota, 2009; Kobayashi, Spitzberg, & Anderson, 2008; Ledford & Anderson, 2013; Neuendorf & Skalski, 2010; Waters et al., 2009).

2. Non-social network (non-SNS) Web 2.0 sites: These sites include the same ability to publish UGC and USC, but can be distinguished by their lack of (or lack of emphasis on) articulated networks with mutually formed connections. Privacy and limited access are harder to control in many of these Web 2.0 sites due to the lack of a list of users with whom connections are shared. And as the goal for many of these types of sites is to reach the largest audience possible, this eliminates the need for the privacy controls present in SNS. In general, if users of non-SNS sites want to keep their broadcasts private, they can require subscription-for-viewing permission or can employ password protection. Non-SNS Web 2.0 sites include personal blogs, podcasts, microblogging services such as Twitter, wikis such as Wikipedia.com, and media-sharing sites such as YouTube. The content analytic study of such 2.0 content has been wide-ranging (e.g., Abbasi et al., 2008; Birch & Weitkamp, 2010; Habel et al., 2011; Lieberman & Goldstein, 2006; Ma, 2013; Neviarouskaya, Prendinger, & Ishizuka, 2009; Oh, Agrawal, & Rao, 2011; Waters & Jamal, 2011; Weisburd, 2009; Yu & Wang, 2015).

**Acquiring Content From Social Networking Sites.** A major challenge in acquiring content from SNS is the precision and wide use of privacy tools
for gatekeeping by SNS users. Whereas the goal of Web 2.0 sites like blogs and wikis is often to reach the largest audience possible, SNS are geared toward providing personal content to a network of “friends” or connections that are mutually agreed upon. Because content on SNS may only be accessible to “friends” (as is often the case on Facebook) or other approved entities, it can be difficult to sample from the general population. Recent studies have done a respectable job of overcoming the privacy challenge in acquiring content from SNS. Carr, Schrock, and Dauterman (2009), for example, first recruited students from a class for participation in a Facebook study and then, after gaining their informed consent, told them to “friend request” a page created by the researchers so that their status messages could be viewed for coding. This interesting, creative approach alleviates concerns some may have about human subjects violations from content analyzing messages set to private. Assuming that all data are kept confidential, we do not believe there are significant ethical breaches in sampling from social networks like Facebook if the user gives the researcher access to their information by friend requesting them or if the posted information is open to the public.

It should be noted that some SNS members have variable privacy settings and may make only a portion of their content available to select members, which should be apparent when certain content is searched for and not visible. If this happens, it may necessitate the elimination of that content from further analyses. We recommend that researchers take an in-depth look at the privacy settings for each SNS and use it to help decide the ethics of using information found outside the profiles of consenting research subjects.

When there are no or few restrictions on a population of messages from social networking sites, we recommend looking for site-provided tools to aid in drawing a sample. For example, both Lunk (2008) and Kane (2008; Kane et al., 2009) sampled MySpace profiles by using the site’s “advanced browsing” function available at the time of their studies. This tool allowed them to display a random sample of profiles based on set criteria. Despite having millions of potential cases, they were to draw 3,000 profiles at once, displayed 300 at a time. A small number were blocked, but the advanced browsing function still allowed them to draw solid, representative samples. There were also settings that could have narrowed their searches to include only certain types of users, selected by sex, age, relationship status, and/or other criteria. Other SNS sites have similar features that can be used by the content analyst.

While archived private messages (similar to personal emails) on Facebook, MySpace, and Twitter allow researchers an unprecedented opportunity to study interpersonal communication outside of the laboratory or field setting, these messages are not readily available to the researcher. Researchers will need to acquire the private content from one half of the dyad and then consider the ethical dilemma of whether or not permission from both halves of the dyad is needed in order to include the content analysis of the private message in a study. Overall, methods of getting at difficult-to-sample content like
SNS profiles or private messages are fraught with limitations that should be acknowledged by researchers.

*Acquiring Content From Non-Social Network Web 2.0 Sites.* Whereas SNS are used to share information within a limited network, non-SNS sites are generally used to broadcast information to as large a group as possible, serving a true mass communication function. Because these non-SNS sites do not rely on mutually agreed upon connections for gatekeeping their content, the challenges in content acquisition for these sites are very different. Because of their function, it is rare to find a blog, microblogging feed, or wiki that is not public. When this does happen, a personal blog or Twitter feed, for example, might offer only limited access to friends and family as a repository of personal thoughts (Nardi, Schiano, & Gumbrecht, 2004). Privacy on these types of sites can be achieved by requiring subscriptions that must be approved by the site owner in order to view or by password protecting the entire site. Because the majority of these sites are completely public, the real challenge in content acquisition is sampling, specifically in the case of blogs and microblogging feeds, which are discussed in detail in the following paragraphs.

While there are many ways to locate weblogs on the Internet, sampling them presents a number of unique challenges. Li and Walejko (2008) outlined several of the pitfalls in the study of blogging, including spam blogs, abandoned blogs, access-restricted blogs, and nontraditional blogs. The most problematic are (a) spam blogs, which are also called *link farms* and are created to boost the page rank of web sites in search engines and (b) abandoned blogs, which may make up as much as two thirds of blogs. Access-restricted blogs, although rare as previously discussed, will tend to be underrepresented in studies because they are difficult to locate. This must be taken into account when designing research and addressing study limitations. Finally, there are nontraditional blogs, which include sites with a user-generated content function that would not traditionally be considered a blog. As Li and Walejko (2008) point out, popular SNS sometimes provide users the option to blog within the bounds of the site. MySpace, for example, has an integrated blogging platform, and researchers should remember those nontraditional blogs, which may not appear in blog searches.

In order to adequately sample blogs, the content analyst must successfully navigate all of these pitfalls and create sampling frames that allow them to select, find, and archive their sample. Li and Walejko (2008) examined 24 blogging studies to uncover four strategies for sampling bloggers and blogs on the Internet: (1) a self-selected or convenience sample, (2) sampling through blog hosts, (3) sampling with the assistance of blog aggregators or indexing web sites, and (4) sampling for ready-published lists of blogs. Self-selection is achieved by soliciting bloggers the same way one would solicit participants for any kind of study—“Students who just started a romantic relationship needed,” for example. This type of sampling has all of the advantages and disadvantages of traditional convenience sampling. Researchers may choose to use a blog host, such as blogger.com, to construct their sampling frame.
Often these blog hosts have accessible blog lists or “find-a-random-blog” functions, which may help researchers in sampling. The downside of sampling by blog host is that it excludes self-hosted blogs, which limits the generalizability to other blog hosts and platforms.

Blog aggregators collect, organize, and publish blogs and blog posts. Some blog aggregators allow users to create their own blog aggregation, while others create the aggregation and order using algorithms. Technorati, for example, uses a ranking algorithm and lists blogs in order of their popularity. Li and Walejko (2008) explain the two major limitations of using blog aggregators to create sampling frames as, first, no one aggregator encompasses 100% of blogs, and second, aggregators are constantly changing, potentially making it difficult to replicate the sampling in future studies. Finally, a researcher can use blog lists and rings (or connections between bloggers who write about similar topics) to construct a sampling frame. Li and Walejko (2008) suggest using blog rings as a starting point to create larger samples. The important thing to remember when choosing a method to sample and acquire blog content is that the generalizability of the research depends on the population and the sample. If one randomly selects 500 blogs from Technorati, the population is not “blogs”; it is instead “blogs listed by Technorati.” Keeping this in mind will help in the proper interpretation of such studies.

Acquiring content from microblogging services presents its own unique challenges. As of the writing of this book, the most popular and ubiquitous microblogging service is Twitter. Launched in 2006, Twitter has grown to more than 320 million active users (“Twitter Usage,” 2015; see also Fiegerman, 2012), who generate more than 500 million “tweets,” or short status updates of 140 characters or fewer, per day (“Twitter Usage Statistics,” 2013; see also Terdiman, 2012). One challenge in acquiring content from a massive service like Twitter is simply conceptualizing the medium. The widespread use of the metadata hashtag (#) to organize mass messaged tweets can aid the researcher in reviewing and sampling tweets. However, Twitter also contains interpersonal and “follow” functions that make it similar to social network sites. In addition to the mass communication function of tweets, which are broadcast, Twitter allows users to directly (through “retweeting”) and indirectly (through manual reposting) repost the content of others. It also contains an interpersonal function (using the @ symbol to allow microbloggers to manage interpersonal conversations), which can be started by all Twitter users, regardless of whether or not they follow one another. In addition to @ replies, there also exists the ability for users to private message only those who follow them, echoing, though not equaling, the articulated mutually agreed upon networks in SNS. Upon this closer examination, we see not only that microblogging is extremely platform specific in many ways, but also that there is a large amount of different kinds of content to analyze in this rich medium.

Published studies of Twitter have employed several methods for sampling, using various tools. Honeycutt and Herring (2009) studied collaboration.
and user-to-user exchanges on Twitter. In order to create a representative sample, they attempted to collect tweets from Twitter.com’s public timeline (a feed of all tweets available to all users in real time). Their sampling strategy attempted to collect tweets in four one-hour samples gathered at intervals. What they realized was that the public timeline was refreshing too quickly to capture all posted tweets, and they ultimately employed a “scraper” program that was able to collect 20 messages at a time in intervals of 3 seconds (Honeycutt & Herring, 2009). Thelwall, Buckley, and Paltoglou (2011) also used a program built to scrape tweets in order to capture them from the public timeline.

Using the public timeline, however, is not the only way to acquire content from a microblogging site. Chew and Eysenbach (2010) used Twitter’s own application programming interface (API) to create an “infoveillance” program, which gathered publicly available tweets containing their key words of interest. These researchers archived over two million tweets containing references to swine flu in order to examine trends in the discussion of the global pandemic. Scanfeld, Scanfeld, and Larson (2010) used Twitter’s built-in search function to acquire tweets that mentioned antibiotics. “Each individual search returns results from a variable time frame (approximately 1 week), depending on the storage capacity of Twitter’s Database,” explain the researchers (p. 183). They conducted two searches per week for several months, generating a list of 52,153 status updates or tweets mentioning at least one of their search terms.

External search tools can also be used to acquire content from a microblogging site. Binder (2012) used the Advanced Google search engine to identify relevant messages on Twitter for a study of U.S. commentary about nuclear risk following the 2011 emergency at Japan’s Fukushima-Daiichi nuclear power station. The researcher’s search criteria returned a sampling frame of 2,359 tweets, and a systematic-random sampling technique was then used to extract a sample of 124 messages for human coding.

Sieben (2014) had a bigger challenge than Binder (2012) in collecting content about the key period of the uprisings in Egypt and Syria in 2011, during the Arab Spring movement. Sieben (2014) was attempting to analyze tweets based on the presence of key words about the Arab Spring movement from more than a year in the past. Because selling data sets has become part of Twitter’s business model, the Twitter API limits the ability of users to create large data sets from information tweeted in the past. If a researcher is not collecting at the time of the event, it becomes increasingly difficult to find the information. Ultimately, Sieben borrowed two data sets from other researchers who had been collecting tweets during the Arab Spring movement and used cURL, an open-source command line tool, to access the Twitter API and collect 1% (cURL is limited to 1% by the Twitter API) of the tweets with a given hashtag to round out his sample. A second challenge that Sieben faced was the size of the population. His borrowed data sets, along with his own collection, amounted to more than 400,000 tweets. Sieben (2014) was able to cut down the sample using a random sampling method. Researchers who
wish to study world events on Twitter, especially those that happened in the past, may experience similar challenges.2

Another way to sample and acquire content on Twitter is by selecting a population that contains a select group of Twitter profiles to analyze. In their 2010 study of Fortune 500 companies’ use of Twitter, Rybalko and Seltzer (2010) selected 170 active Twitter profiles maintained by Fortune 500 companies, identified the set as their population, used the list as their sampling frame, and drew a random sample. They utilized two sampling units for their study: Twitter profiles and individual tweets. In their study of Twitter coverage of the 2011 Mumbai terrorist attack, Oh, Agrawal, and Rao (2011) chose to examine only one Twitter profile, www.twitter.com/mumbai. Their content analysis examined each individual tweet as a unit.

Unfortunately, much archived Web 2.0 content is either private or proprietary (Karpf, 2012). In 2014, microblogging platform Twitter had a pilot project called Twitter Data Grants, a competitive program for researchers to gain access to Twitter data sets and the assistance of Twitter’s engineers and researchers. This and other collaborative projects between online providers and academic researchers are one way for content analysts to apply their skills to proprietary big data message collections.

Acquiring Online News

Hester and Dougall (2007) offer sound guidelines for acquiring online news content, based on the results of a study they conducted comparing different sampling methods and sizes. The researchers collected 7,438 news articles on Yahoo! News during a 183-day period and then had them coded for the types of content appearing each day to identify population parameters. Then, they drew 50 samples in each of several different manners to test for the representativeness of each approach. Their results indicate that constructed-week sampling, which involves identifying all Mondays, Tuesdays, and so on and then randomly selecting one Monday, one Tuesday, and so forth to construct a composite week, works substantially better than simple random sampling or consecutive-day sampling. Their findings also suggest that a minimum of two constructed weeks are needed and that as many as five may be needed for certain variables (i.e., those with high variability). Hester and Dougall’s research not only contains good advice for sampling, but it also serves as an example for how empirical evidence may be generated to guide decisions about what acquisition methods work best for answering particular content analysis questions.

Content Acquisition by Interactive Media

In addition to options for acquisition of interactive content, there are also options for selection of content by interactive media. Well-worn search engines such as Google, Ask Jeeves, and Bing might be the most salient tools
for accessing content through interactive media, but there are other options. Web crawler programs, which systematically browse the web, have the potential to search for and acquire content of interest. For example, Wolfram Alpha is a “computational knowledge engine” that uses a vast collection of expert knowledge to answer questions from users (www.wolframalpha.com). Although this technology has been used for data mining, it could also be programmed to retrieve message content of interest (i.e., text mining). There are also different kinds of content aggregators, including FlipBoard, which allows the collection of multiple types of content, Rotten Tomatoes and Metacritic for film reviews, a variety of feed readers including NewsBlur and Feedly for RSS feeds, Apple News and Google News for news, and a variety of applications that search and track heavily discussed topics on social media. These technologies harness the power of interactive media for automated, and often automatic, content retrieval or collection and are promising options for content acquisition.

Archiving Content in the Interactive Media Age

Archiving content for analysis can be a challenge with interactive media. It does not come “pre-packaged” and ready for analysis. Saving video game play, web content, and other forms of interactive output requires a few additional steps, for which there are fortunately a number of viable options. Getting a “snapshot” of this type of content is essential due to its changeability and updating. This section reviews options for archiving interactive content and then shifts to a discussion of how interactive media itself can aid in the archiving of traditional and new media content.

Archiving Video Games

Unlike traditional media, video games have to be played in real time and simultaneously recorded in order to be captured and archived for later analysis. A typical procedure for this had been to have research assistants play games selected for inclusion in a sample and then use video recordings to capture the play (e.g., Lindmark, 2011; Smith, Lachlan, & Tamborini, 2003). Having content in digital form has the advantage of being easily duplicated and shared with other coders for a reliability check. It can also be archived for future analyses and/or shared with other researchers.

Archiving Web 2.0 (and Other Web) Sites

Web pages are more difficult to archive now than they were in the Internet’s early days, when they were simple HTML constructions. Current sites have many disparate elements that may not be captured by using the
“save page” function in a browser. However, there are several options for archiving pages that work with varying degrees of success, depending on the type of page and content. For example, Shelton and Skalski (2014) used Adobe Acrobat to archive Facebook profiles and photo pages by converting them to PDF files, creating virtual “snapshots” of the sites. From these, the researchers were able to code for both pro-academic behavior and anti-academic behavior, in most cases. The only problem they encountered was that the photos in some archived pages were too small to see certain variables, in which cases they instructed coders to return to the profiles and look at the full-sized photos.

Evernote is also useful for archiving social media messages. Evernote, a software that facilitates the collection of a wide variety of types of messages (see Box 7.2), can capture more information than just text or the basic content of the message. It allows the researcher to take a screenshot or photo of the message and archive it on Evernote’s cloud server as well as on the researcher’s desktop. The messages can be kept private or made accessible to other researchers and become searchable through Evernote’s use of optical character recognition (OCR) to convert the text of photographed messages into searchable text, allowing the researcher to search individual words and phrases in a message sample gathered from SNS.

Kane (2008) used an alternative technique for archiving MySpace profiles for her study of self-presentation on social networking sites. She saved the page files in her sample using MHTML, which preserves all HTML, text, and photos in a single file and thereby maintains the integrity of the original display. It does not save all content, however, such as audio or video. Much software has been developed for the express purpose of saving entire web sites, such as TeleportPro. Simon (2011) used HTTrack Website Copier to archive political web sites in his comprehensive study of news, online, and public foci on political races in the 2010 Ohio election.

For Internet filtering, security, and computer protection, services such as IBM’s Internet Security Systems (ISS) use algorithms to detect faces and objects and analyze content for objectionable features (e.g., nudity, illegal activities). Such facilities might be adaptable to content analysis needs for searching and saving.

Traditional Media Archiving With Interactive Technology

Despite some challenges with preserving certain forms of interactive content, newer media also offer a number of useful options for archiving traditional media content. Content analysts no longer have to hunt through dusty physical libraries for magazines, videotapes, or other coding content. Or even if they do, they are able to snap digital photographs of old print ads, for example, posting them online for sharing by the entire research team (Dixit, 2016). Content can now be archived digitally using interactive technology, which saves space and speeds up the rate at which the content may
be pulled for analysis. Options for archiving audio and video content include programs such as Annotape and askSam. As mentioned earlier, SnapStream is another option for storing and maintaining a digital archive of moving image content.

Ample space also exists for storing digital content now, thanks to increasingly large personal hard drives (frequently in the multi-terabyte range) and cloud storage services, which are booming in popularity. Cloud storage is space on a huge remote computer server that users access via the Internet (Walton & Fendell Satinsky, 2013). Examples include Dropbox, Google Drive, Microsoft Skydrive, and Apple iCloud. Most cloud storage services offer some free space and charge a fee for additional space. From our experience, the free space of multiple cloud storage accounts provides enough room for storing most types of content analysis data. Cloud spaces also allow for the easy sharing of content and coding materials among multiple coders. Coders can be granted access to a single cloud server with everything they need. They can even do their coding “on the cloud” (versus downloading a digital coding sheet to their hard drive) and have their work immediately saved and inspected by the content analyst. Cloud computing introduces a host of possibilities for content analysis that we are only beginning to realize.

Coding Content in the Interactive Media Age

Coding interactive content involves the usual steps in a content analysis, as discussed in this text, with some new challenges and issues at the stages of unitizing, codebook and coding form creation, and coding itself.

Unitizing Interactive Content

Units of data collection in interactive media often have parallels to traditional units in content analysis. It is up to the researcher to decide which units are appropriate for answering particular hypotheses and research questions of interest. Lunk (2008), for example, was interested in comparing communication patterns of U.S. and Hungarian users of social networking sites, so she sampled 300 comments left on MySpace profiles of U.S. and Hungarian users. Each comment was deemed a codable unit. This unitizing parallels the traditional analysis of sentences, utterances, or turns in real-world communicative interactions (e.g., Bales et al., 1951). Similarly, Martins, Williams, and Harrison (2008) used 368 adult female game characters as units of data collection in their study of women’s body type portrayals in popular video games, which parallels work that has looked at body imagery in print media and on television (e.g., Byrd-Bredbenner, 2003; Greenberg et al., 2003).
The analysis of online discussions might be informed by studies of face-to-face interactions. For example, whether to select the post, the thread, or even the social network as the unit of data collection for research on chat rooms or social media is a perennial question (De Wever et al., 2007; Strijbos & Stahl, 2007). These decisions are analogous to the selection of the turn, the verbal exchange, the discussion, or the dyad or group in research on face-to-face interactions. Further, as Strijbos and Stahl (2007) found, “unit fragmentation” often occurs, which refers to fragmented utterances by a single author spanning multiple chat lines. These fragments make sense only if considered together as a single utterance. Thus, reconstruction of the interaction is necessary.

In other instances, the unitizing task for interactive media does not have a clear counterpart in traditional research. The choice of what unit to select within interactive media output (IMO) content has developed via convention among researchers who focus on each type of content. In a typical example, in their study of the computer simulation “Networked Fire Chief,” Elliott et al. (2007) selected 20 five-minute scenarios for each of the 20 participants. Video game researchers face a similar decision process, as described earlier.

### Codebooks and Coding Forms

Content analysts have capitalized on technology advances even in the very construction of their codebooks. Electronically produced codebooks can easily include images, diagrams, and links to instructive online materials (such as specific exemplars, should the researcher choose to introduce them). The capability of researchers to create codebooks that include pictorial elements has increased over time. Figure 7.1 shows a typical “demo” page from a codebook, indicating where measured variables (explicated in words on other codebook pages) can be located in the targeted content (eBay auction pages; Wu & Neuendorf, 2011).

Most content analysts have switched from traditional paper coding forms to electronic coding forms, such as Excel files. Coders can leave these files open on a screen beside the medium displaying their content (or on the same screen, in separate windows) and code more quickly and easily than they would on paper. Coding directly into electronic files also makes intercoder reliability checks and subsequent analyses easier since the data are already in an appropriate format. Both Lindmark (2011) and Brown (2011) used Excel files for coding, in slightly different ways. In his study of advertising in popular video games, Lindmark’s coding forms were Excel files with two tabs that allowed for switching between game-level and advertisement-level variables. In her study of disrespectful behaviors on children’s television, Brown had a tab for each show in her sample and coded all program-level variables within each. With careful backups of electronically coded data, electronic coding forms are an obvious advantage.
Coding Interactive Content

Coding of interactive content should follow the advice given earlier in this text, including training, pilot coding, and intercoder reliability checks. Given the complicated nature of some new media content, specialized coding procedures may be required in some cases. Shelton and Skalski (2014), for example, encountered low reliability between certain coders for some variables in their study of Facebook profiles. They therefore made the decision to have the more carefully trained, more reliable coders code certain complicated content and have the other coders handle easier content. This approach is not ideal, but in some cases, it may be the only option to obtain an acceptable level of intercoder reliability.
The Use of Technologies for Medium and Message Management, Message Preparation

There always has been a need for the content analyst to understand the nature of the medium in which the targeted messages are found and the operation of equipment for delivery of the messages. In the past, this could include knowing how to find specific news articles in a newspaper’s morgue, learning how to record and play back audiotape or videotape or knowing how to use transcription equipment. But with the proliferation of options in electronic and digital media, new considerations have come into play for the preparation of messages for coding and for message handling during coding.

Advances in technologies for the automatic processing of messages have stemmed largely from the areas of computer science and machine learning, generally without clear applications to the needs of social/behavioral scientists and others who are interested in the research applications of these technologies (Divakaran, 2008). What follows are some examples of key processes that digital technologies have made available for possible message preparation, handling, and even coding.

1. Image measurement: In an inventive study of food portions over time, Wansink and Wansink (2010) used a CAD-CAM system for the simple process of measuring the “food-to-head” ratio in a series of Last Supper paintings spanning the last millennium. And Pettijohn and Tesser (1999) used the PhotoMagic software to help in the measurement of facial features of film actresses in a study of facial attractiveness over time.

2. Speech recognition/automatic transcription: There are many systems that provide the support function of the transformation of spoken messages electronic text form (i.e., automatic speech recognition or ASR). This capability has been delivered into the hands of anyone with a computer or smartphone; for example, the Dragon NaturallySpeaking software by Nuance is available in six different languages. Speech recognition has been used for the transcription of speech samples (e.g., Gottschalk & Bechtel, 2005) and broadcast news (e.g., Gauvain, Lamel, & Adda, 2000) for quite some time, in the latter case showing superiority over the use of on-the-fly closed captioning. Oger, Rouvier, and Linares (2010) have used speech recognition to automatically transcribe a variety of moving image content, resulting in the correct identification of seven types based solely on linguistic content: cartoons, movies, news, commercials, documentaries, sports, and music.

3. Multimedia content analysis: This term has been appropriated by information technology (IT) specialists seeking ways of filtering, indexing, and retrieving video, audio, and digital images. As noted by Dimitrova (1999, p. 87), “Data compression coupled with the
availability of high-bandwidth networks and storage capacity have created the overwhelming production of multimedia content. . . . For content producers, advertisers, and consumers, there will be increased availability and increased challenges to manage the data.” A number of systems for retrieval of multimedia content have been developed, such as the products of Streamsage, Inc., which include tools for searches and retrieval of “timed media” content, such as streaming audio, streaming video, timed HTML, animations such as vector-based graphics, and slide shows (Streamsage, 2012, p. 1).

Two 2012 volumes provide the state of the art of this type of so-called “content analysis” for video (Kompatsiaris, Merialdo, & Lian) and for audio (Lerch). The techniques presented in these volumes include practices for the annotation, retrieval, organizing, and quality control of video and the extraction of metadata and other information from audio signals. The main point of all of these techniques is to automate the indexing and retrieval of video and audio archives. While they do not execute content analysis in the sense defined in this book, these algorithms might be useful adjuncts to the execution of a full content analysis.

Specific functions which seem most applicable to content analysis include the following:

1. Video segmentation: By programming a computer to recognize segmental markers, such programming segments as separate news stories (Boykin & Merlino, 2000; Kubala et al., 2000) or commercials may be identified. SRI International’s Scene Segmentation and Classification (SSC) process provides the real-time indexing and retrieval of individual shots in videos (“Advanced Manipulation and Automation,” n.d.). Further, most video editing systems have built-in shot detection/segmentation. However, Cutting, DeLong, and Nothelfer (2010), in a careful investigation of the utility of such algorithms, note problems with their accuracy to date.

2. Image, facial, object, and behavior recognition: A number of systems have been developed that facilitate computer recognition of static or moving images. Content-based image retrieval (CBIR) is the term applied to the process of retrieving desired images from a large collection on the basis of syntactical image features such as color feature, texture, shape, luminosity, and edges (Rorissa, 2007). CBIR avoids the limitations of metadata-based systems, which require human intervention in the description of images in a database. Rather, CBIR uses the image content itself as the data for comparison with other images. The free online system TinEye (www.tineye.com) is an example of such applications; the system will search the Internet for occurrences of a particular image or somewhat-altered versions of that image.
Facial-recognition software is becoming more present in consumer software programs. For example, HP’s Automated Publishing can identify faces on web sites. And Apple’s iPhoto and iMovie programs both contain facial recognition features. A user could, for example, archive 1,500 photos of political rallies in iPhoto and begin tagging specific attendees. iPhoto would then learn to associate people’s names with their photos and search the entire corpus of photos for those people, tagging them when their faces are recognized by the software. Though this is clearly a commercial application, it may have some utility for the content analyst.

Non-research applications of this type of technology show the potential power of the technique while raising significant privacy and civil liberties issues. The surveillance system FaceTrac was used by the Tampa Bay police to scan the fans at the 2001 Super Bowl, comparing their facial characteristics to mug shots of known criminals (Grossman, 2001). No arrests were made that day at what has been called the “Snooper Bowl” (Singel, 2010), and only after $8 million in system upgrades were the police able to use the system effectively on the street. Tampa police used digital cameras to take pictures of citizens at traffic stops, compared them against a database of 7.5 million mugshots, and made nearly 500 arrests (Singel, 2010).

Human interaction recognition has been the goal of some research, with only modest success to date. For example, Patron-Perez et al. (2010) developed algorithms for the detection of four discrete behaviors in the context of video retrieval: handshakes, high fives, hugs, and kisses. However, the optimal outcome they achieved was a 64% correct identification rate, well below what would be expected via human identification and coding.

3. Streaming video: Streaming video allows Internet users to play video content “anytime anywhere, thereby eliminating the spatial and time constraints of traditional media forms” (Dupagne, 2000, p. 11). The technical deficits to online video noted over a decade ago (Dupagne, 2000; Neuendorf, 2002) have been somewhat overcome, although temporal interruptions to the streaming of films from Netflix and other services remains a user complaint (Campbell, 2012). Further, a content analysis comparing streaming and DVD-based film viewing revealed that distinct differences exist between the quality of streaming films and the DVD-presented film, including aspect ratio, color and sound quality, and picture clarity (Campbell, 2012). Nevertheless, content analysts now routinely conduct their coding of film and television content and of recorded human interaction behaviors streamed directly from online archives. This type of coding has obvious advantages: the elimination of costly materials, such as recording media (e.g., videotapes, DVDs) and playback equipment, and the efficiency of multiple coders having immediate access to the same pool of messages.

4. Metadata: Perhaps no other class of innovation holds greater promise for advances in automated content analysis that does the application
of metadata. Metadata means “data about data” and could be anything that summarizes information about a set of data, such as a message. The hypothetical uses of metadata have far exceeded the real-world applications to date, particularly when it comes to moving image metadata.

Perhaps the most familiar use of metadata is in the annotation of computer files, in which date of creation, the creator’s user identity, file type, file size, and other descriptive information are embedded within the file. Digital cameras routinely attach important metadata to still photographs, including make and model of the camera, shutter speed, aperture value, focal length, and ISO.

In the video realm, there have been much contestation and many changes in recent years. Older video systems (such as the American standard NTSC) allowed a small amount of metadata to be carried, such as closed captioning and DVS (descriptive video service, an oral description of action provided for the visually impaired) carried in one line of the vertical blanking interval, the area between video frames. The contemporary digital systems have greatly expanded the opportunities for metadata, which in a digital environment may be stored anywhere in the signal throughout the entire duration of the content. Television engineers noted that with the diffusion of digital content, “the number of distinct varieties of Metadata” became “potentially limitless” (EBU/SMPTE Task Force for Harmonized Standards for the Exchange of Program Material as Bitstreams, 1998) and early on they offered such suggestions for metadata as copyright information, author(s), and origination date and time and such technical indicators as color correction parameters, time code, and edit decision lists used to produce the video.

However, to date, no widely accepted standards for metadata have been adopted for digital video. MPEG-7, released in 2001 by the Motion Picture Experts Group (MPEG), is a multimedia content description standard that provides a rich set of tools for the attaching of metadata to photos, audio, and video. For several years, MPEG-7 annotation software (e.g., IBM’s Annotation Tool) showed promise for the embedding of a wide variety of information as metadata—not just technical specifications, but also data about the content (e.g., “exterior scene,” “water,” “animal-deer”). However, the software was not widely adopted, and alternative applications have left the situation fuzzy. MPEG-21, introduced in 2004, is a standard that provides further metadata capabilities but was developed primarily as a “rights expression language” intended to manage restrictions on digital content usage. The MPEG-21 standard is designed to incorporate metadata on ownership and licensing of digital content and so far has not been used for metadata that might be of significant interest to content analysts.

Thus, metadata is currently viewed having only industry applications, such as copyright protection or for professionals to receive credit on a media product (Robair, 2015). And to complicate things further, metadata
may be thought of as primarily focused only on “production” or “distribution,” and therefore might not be carried through from content creation to audience reception.

In the future, metadata capability may readily be used to store information about message substance and form, such as human characters in the filmic frame, the dialogue being spoken (i.e., the script), and type of shot and transition at each point in time. This type of metadata could pave the way to a fully automated form of content analysis for the moving image.

**Automated Technologies and Measurement on the Moving Image**

All the systems described in the last section provide assistance to the content analyst but stop short of fully automating the coding process. As described in Chapter 5 and Resource 1, CATA options do provide fully automated coding of text, including both coding with a priori schemes (internal and custom-designed dictionaries) and with “emergent coding” approaches. But computer coding is still essentially limited to CATA—the promise of computer-automated coding of moving image content (e.g., “Teaching Computers to Watch TV”; Evans, 2000) has not yet been fully realized. However, several initiatives have provided indications that certain types of computer-driven measurements might be useful.

First, the full range of “mining” functions ought to be considered. The original notion of “data mining,” the aggregate, strategic, and nondirected (i.e., atheoretic) application of statistical techniques to large sets of data in order to find meaningful patterns, has been expanded to other informational forms. “Text mining,” as has been mentioned, is a broad-brush application of emergent-coding CATA techniques (e.g., Thelwall, Wilkinson, & Uppal, 2010). And “video mining” may prove to be a useful set of tools for the wholesale identification of moving image patterns in form and content (Rosenfeld, Doermann, & DeMenthon, 2003).4

To date, some researchers have successfully harnessed the capabilities of computing for the automatic content analysis of specific moving image material. For example, Kearns and O’Connor (2004) used Autonomy Virage video search software in their extension of the notion of entropy for the moving image (originated by Watt, 1979, and Watt and Welch, 1983, as a specific application of the notion of message complexity). Entropy is the degree of randomness or unpredictability in a set of elements such as words, numbers, or [television] program production elements (Watt, 1979, p. 59). Watt (1979) provides six different formulae for the measurement of visual and auditory entropy. The higher the entropy, “the less predictable is the appearance of any unit, and the more complex is the message” (p. 59). Greater viewer attention occurs with “a greater element of surprise, greater information, and greater entropy” (Kearns & O’Connor, 2004, p. 146). Thus, the researchers were
concerned with measuring entropy at multiple points in a video presentation and used the Autonomy Virage module to extract key frames. In other words, the program grabbed video frames at points of detectable change in the data stream (p. 153), with those points being the units that needed to be identified for systematic coding of entropy.

In a unique combination of technology and critical/cultural film analysis, Anderson and O’Connor (2009) analyzed a small segment of the Alfred Hitchcock film *The Birds* in order to compare electronically deduced “structural” characteristics with a close semiotic analysis by film scholar Raymond Bellour. Using an AVI file version of the sequence, they extracted 12,803 JPG image files, each of which generated an RGB (red-green-blue) histogram, from which a Gini coefficient was calculated for each of the frames. Each Gini coefficient represented the color distribution of the video frame. These coefficients were graphed in several ways, including an analysis of frame-to-frame change, providing a ready comparison of these “hard data” on color distribution with Bellour’s shot-by-shot critical analysis of the film sequence. As Anderson and O’Connor note, their technique “demonstrates the validity of this approach to numerical and graphical representation of filmic structure” (2009, p. 39). They suggest further analyses of film soundtracks and other moving image elements.

However, Cutting et al. (2010) provide evidence that even the simple task of dividing moving image content into separate shots is not yet fully achievable via automated methods. They note,

> We were unimpressed with purely digital methods. Cut-finding algorithms often confuse motion across frames within a shot with spatial discontinuities across shots. They also do poorly with fades, dissolves, and wipes . . . [with] hit and false alarm rates of about 95% and 5%. (p. 2)

Cutting et al. (2010) were interested in content analyzing the editing rhythm of 150 Hollywood films released between 1935 and 2005. They used a combined coding process of a MATLAB-based identification of candidate cuts/transitions, with human coder inspection of the 12 frames surrounding each candidate transition to confirm computer-identified transitions and check for missed transitions. Their findings, by the way, identified a trend over the 70 years toward greater correlation of shot length with adjacent shot lengths, with recent films more closely matching a human tendency for changes in attention to follow a spectral patterning known as $1/f$. They conclude that contemporary films more naturally “help harness observers’ attention to the narrative of a film” (p. 1).

There have been noteworthy advancements in the digital/electronic measurement of message features for one particular type of content—music. In an interesting example using both text analysis of lyrics (specifically, via the General Inquirer CATA program) and automated computer analysis of
acoustic attributes of the music (using the WEKA package for machine learning), Yang and Lee (2004) applied psychological models (e.g., Tellegen, Watson, & Clark, 1999) to study negative emotion in alternative rock music. The General Inquirer measures of hostility, sadness, guilt, love, excitement, pride, attentive, reflective, and calm best distinguished the song lyrics, while such acoustic attributes as beats per minute and timbral features such as spectral centroid and spectral kurtosis were used to correlate with volunteer assessments of the music’s emotion intensity.

Some scholars have been exploring machine-learning applications to music information retrieval (MIR; www.ismir.net/), the interdisciplinary field that is concerned with extracting information from music. The goals range from the automatic generation of a musical score to the machine measurement of features such as timbre, chords, harmonies, and rhythms to the automated categorization of music genre. Others have applied similar models to the continued development of alternative algorithms for the automated detection of emotion in music (e.g., Trohidis et al., 2008), culminating in a full volume devoted to “automatic music emotion recognition (MER) systems” (Yang & Chen, 2011; see also Lerch, 2012).

Almost Content Analysis

Processes that are almost (but not quite) content analysis have become widely available to the average interactive media consumer, ranging from simple text-to-visual converters/compilers such as Wordle and Leximancer to interactive big data text mining compilers and summarizers like Radian6. Some applications provide personalized message analytics, such as ThinkUP’s ability to provide individual social media users with such information as whether they retweet more men than women and how much they talk about themselves. Other applications exemplify the power of the web to convey summaries of message information on virtually any topic—for example, based on true scholarly research on logos and their strategic importance to organizations, James I. Bowie’s Emblematic web site (www.emblematic.com/) reports on trends in logo design, using quantitative data from the U.S. Patent and Trademark Office. Analyzing more than 1.2 million logos dating back to 1884, Emblematic presents patterns over time and across industries, including new styles, deaths of old trends, color use, and geography.

While not meeting the criteria of a scientifically motivated investigation of messages that is content analysis, these message analytic services and adjuncts provide us with new ways to summarize messages, often using big data sources. They can provide the content analyst with the grounding to develop future investigations. (They are also, we find, a good way to get students interested in content analysis!) Box 7.3 gives examples of these popular, not-quite content analysis options.
Conclusion

Clearly, there are numerous content analysis capabilities, limitations, and challenges that have emerged in the interactive media age. All continue to proliferate. For updates, visit *The Content Analysis Guidebook Online* (CAGO).

**Box 7.3 Not-Quite Content Analysis**

With the proliferation of online content and messages, the analysis of content—or in many instances its close cousin, text mining—has become big business. Depending on your background, you may be familiar with robust social media analytics software such as Radian6 or Sprout Social, which are able to pull content data sets from platforms including Twitter, Facebook, YouTube, blogs, online communities, and more. These programs are geared toward business users rather than social scientists, but they can offer ways to capture and analyze content that are helpful to researchers and professionals alike. In addition to often expensive, business-oriented services like these, free tools that claim to analyze content are also becoming readily available. Open access to social media platform APIs (application programming interfaces) by developers and users alike has caused a proliferation of free or nearly free apps that offer to analyze everything from post sentiment and user influence levels to “your personality.”

With all of these new online interfaces, however, the question “Is it content analysis?” still applies. In many cases, these programs are specialized applications that mine text, or use atheoretic- and emergent-coding schemes to analyze content. Some of these new apps, however, are beginning to use more established coding schemes. Whether scholastically sound content analysis or something else, these apps, which are popping up left and right as data become more accessible, are prime examples of the power of computing in the swift analysis of massive quantities of messages.

*What’s an API?* You may have heard the term API when discussing the analysis of new media content. An API is an application programming interface, or a way for developers and users to interact with social media platforms. Most developers interact with a platform’s API in order to create an application for that program. Periscope and Meerkat, for example, are two live video-streaming applications that interact with Twitter’s API to allow users to tweet while they are broadcasting. APIs also function to allow third parties access to the data gathered by the platform (Wang, Callan, & Zheng, 2015). In order to pull data from Twitter’s stream API, a researcher will either need to use an application created for that purpose or create one using a programming language such as Python. Luckily, there are many applications that allow users some level of access to social media data. In fact, Twitter has made its own analytics package available to users, though it doesn’t complete any true content analyses at the time of this writing.
From Twitter’s trending topics to the capabilities of Google Analytics, there are many places where content is being quantified and repackaged with new meaning for end users. Since it would be impossible to create a list of all such applications available, what follows is a more in-depth look at a few examples in which content is analyzed in some fashion online today:

1. Google Trends is an analysis tool that allows online inquiries into the data of the billions of Google searches that users execute each month (www.google.com/trends/). These trends can be examined by time period, topic, and geographic location of users. Google Trends allows users to search not only individual topic trends but also to search for multiple topics in order to create useful comparisons. In the following figure, the names of three early presidential candidates for the 2016 election have been searched along with the word *president*. Upon a cursory review, it’s evident that Hillary Clinton has been associated with a presidential run longer than the other two candidates, and her name trends along with *president* in 2008, at the time of the last presidential election with no incumbent. In 2012, when President Barack Obama was running for a second term, we see the emergence of Donald Trump and *president*, as he was considering entering the field of Republican candidates at that time. It’s clear to see how Google Trends provides a useful analysis of search term content. However, it’s important to note that the numbers displayed on the graphs are not absolute search volume, but representative numbers assigned by Google based on an algorithm. Google Trends is free to use, but there are additional premium features and options.

Screenshot of Google Trends™ captured August 20, 2015

(Continued)
2. Part of Salesforce’s marketing cloud, Radian6 (www.exacttarget.com/products/social-media-marketing/radian6), primarily functions as a social media monitoring tool that collects and analyzes data from multiple social media platform APIs, including Facebook and Twitter, as well as from blogs, online communities and forums, mainstream news sites, and more. Radian6 is part of a large group of similar cloud- and software-based applications that includes Sprout Social, Lithium, and others. The application offers key word analyses, word frequencies, and a sentiment analysis using natural language processing. Unlike Google Trends, the data analyzed by Radian6 comes from social media conversations rather than search terms. The user can select which of the platforms each analysis draws from and a time frame for the analysis as well.

The following screenshot shows several different analyses on the topic of the University of Akron. In 2015, the university experienced significant public relations challenges, and the analyses here help to put them into perspective. The sentiment analysis shows that even though there have been a number of negative posts, as can be seen in the “news river,” a widget that offers a live stream of mentions, 77% of the posts about the university are still classified as positive. The comparison of the terms olive jar and LeBron James show that while students are posting regularly about an expensive olive jar in the university president’s remodeled home, the announcement that pro basketball star LeBron James would partner with The University of Akron is generating far more mentions and is also potentially responsible for the positive sentiment analysis.

Radian6 will also allow researchers and professionals to download and export collected messages, providing large samples of social media and news media content for analysis. The one downside of the platform is cost. There is no free version, and the application is priced at the enterprise level.
3. AnalyzeWords (www.analyzewords.com) is a novelty program that allows you to input a Twitter handle for anyone and generate a “personality” analysis that includes the source’s emotional style, social style, and thinking style. Below, Katy Perry’s Twitter handle has been analyzed. We can see that she scores high for Depressed under emotional style, and high for Spacey/Valley Girl under social style. While an application of this nature certainly might not at first blush seem scientific, James W. Pennebaker and Roger J. Booth, who together created the Linguistic Inquiry and Word Count (LIWC) computer program, are the development team for Analyze Words and have used LIWC as the engine for the analyses it creates. Booth and Pennebaker also note on the site that they are keeping data from analyzed handles to create a larger set for future analysis.

4. Twitter Trends are divided into two types: location trends, which are determined by the user selecting a geographic location, and tailored trends, which are created by an algorithm that claims to identify topics that are popular now based on each user’s location, followers, and other factors (“FAQs About Trends on Twitter,” 2015; Wang, Callan, & Zheng, 2015). The display of these trends has become more integrated into the

(Continued)
Twitter platform over time, and Facebook has followed suit, offering trends of its own that are also integrated into the basic user experience. Twitter trends appear in the bottom left-hand corner of each user’s main profile screen. Users can switch between location trends and tailored trends easily and can select any location they would like to monitor. Another addition to Twitter Trends is “promoted trends,” where companies place a sponsored trend in the trend space.
5. Users who want to further analyze trending terms can use third-party sites such as Topsy (www.topsy.com) to compare trends and view trend performance over time. Some of these sites offer limited free analyses, but most, like Topsy, have a premium or pro version where users can get more detailed data and analyses. The following analysis compares tweets per day with the key words Donald Trump, Hillary Clinton, and Bernie Sanders. Topsy.com displays the tweets that garnered the most attention during peak trend periods, like the following mashable tweet, which caused Bernie Sanders to trend when Black Lives Matter protesters interrupted one of his campaign events. The large peak for Donald Trump occurred during the first Republican debate.

![Tweets per day: Donald Trump, Hillary Clinton, and Bernie Sanders](Screenshot from Topsy.com captured August 20, 2015)

### Notes for Chapter 7

1. The occurrence of unexpected crossovers of functions is worth noting and perhaps deserving of future study. For example, the family of one of this chapter’s authors posts home videos on YouTube for viewing by relatives in distant locations. Nevertheless, one of these videos has over 500 views—clearly people other than family members have found the video worth a look. Although intended for family–interpersonal communication purposes, the video has ended up with a “mass” (large, undifferentiated) audience.

2. Commercial firms such as Crimson Hexagon have begun to provide samples of archived content for a fee from such 2.0 sources as Twitter and Facebook.

3. Although seemingly a ready source of message information, closed captioning has not proved to be an attractive source because of frequent inaccuracies. Even classic films often have closed captioning that is grossly abbreviated or simplified, not accurately representing the spoken dialogue. And real-time captioning is full of errors (Dellinger, 2000); for example, when Meryl Streep won the Academy Award for *Sophie’s Choice*, the word *Holocaust* appeared on-screen as *holly cost*. On April 19, 2013, closed captioning for a broadcast by *Fox News*
declared that the suspect in the Boston Marathon bombing was “19-year-old Zoey Deschanel.” Exceptionally bad closed captioning has attracted the attention of collectors of “found” humor—to wit, a 2009 local weather broadcast in Cleveland, Ohio, attracted over 12,000 YouTube viewers with the forecast, “My cats got weeded down again, other and said they don’t get what it down. Lick-ing here again at a fairly isolated pattern . . . the Duracell could pop up camel’s clouds begin to build it in Akron can area. Kinsman have a little benders night” (“Fox 8 News,” 2009).

4. Non-research applications of video mining include the unobtrusive and automatic surveillance of shoppers in order to detect types of in-store behaviors (e.g., www.videomining.com).


References


Brayack, Barbara. (1998). A content analysis of housing messages targeting the elderly (Unpublished master’s thesis). Cleveland State University, Cleveland, OH.


Campbell, Rachel M. (2012). *Film viewing in the interactive age* (Unpublished master’s thesis). Cleveland State University, Cleveland, OH.


Casey, Mary. (2016, expected). *Our community online: A look at local community web sites* (Unpublished master’s thesis). Cleveland State University, Cleveland, OH.


References


Fan, David P., & Shaffer, Carol L. (1989, November). Opinion survey using open ended essays and computer content analysis: College students’ knowledge of AIDS. Paper presented at the annual meeting of the Midwest Association for Public Opinion Research, Chicago, IL.


Fox 8 News (Cleveland) and their closed captioning. (2009, July 27). Retrieved from https://www.youtube.com/watch?v=7MzmiiWo-tk


Franke, Michael. (2000). Social perception and attribution of responsibility in news magazine coverage of the Manson family. Unpublished manuscript, Department of Communication, Cleveland State University, Cleveland, OH.


Hayes, Andrew F., & Krippendorff, Klaus. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures, 1*, 77–89.


The changing portrayal of adolescents in the media since 1950 (pp. 105–131). Oxford: Oxford University Press.


Kane, Carolyn L. (2008). I’ll see you on MySpace: Self-presentation in a social networking web site (Unpublished master’s thesis). Cleveland State University, Cleveland, OH.

Kane, Carolyn L., Maguire, Katheryn, Neuendorf, Kimberly, & Skalski, Paul. (2009, November). *Nonverbal displays of self-presentation and sex differences in*
profile photographs on MySpace.com. Paper presented to the Human Communication and Technology Division of the National Communication Association, Chicago, IL.


References


Lindmark, Pete. (2011). A content analysis of advertising in popular video games (Unpublished master’s thesis). Cleveland State University, Cleveland, OH.


References

cluttering of television. Paper presented to the Mass Communication Division at the annual conference of the International Communication Association, Montreal, Canada.


Mason, Michael J. (2010). Attributing activity space as risky and safe: The social dimension to the meaning of place for urban adolescents. *Health & Place*, 16, 926–933.


Miller, M. Mark, Boone, Jeff, & Fowler, David. (1992, November). The emergence of greenhouse effect on the issue agenda: A news stream analysis. Paper
presented at the annual meeting of the Midwest Association for Public Opinion Research, Chicago, IL.


References


Robb, David. (2000, February 29). Blacks get lots of prime-time roles, but half are in sitcoms, study finds. Cleveland Plain Dealer, p. 6E.


Short, Jeremy C., & Palmer, Timothy B. (2008). The application of DICTION to
content analysis research in strategic management. Organizational Research
Methods, 11, 727–752.

Short, Jeremy C., Payne, G. Tyge, Brigham, Keith H., Lumpkin, G. T., & Broberg, J.

Shoukri, Mohamed M. (2011). Measures of interobserver agreement and reliability
(2nd ed.). Bota Raton, FL: CRC Press.


Sieben, Johannes. (2014). Twittering the #ArabSpring? An empirical content analysis

Pastiche, literary style and Raymond Chandler. Computers and the Humanities,

Dale Kunkel, Jennifer Manganello, & Martin Fishbein (Eds.), Media messages and public health: A decisions approach to content analysis

Signorielli, Nancy, & Bacue, Aaron. (1999). Recognition and respect: A content
analysis of prime-time television characters across three decades. Sex Roles, 40,
527–544.

Signorielli, Nancy, McLeod, Douglas, & Healy, Elaine. (1994). Gender stereotypes in
MTV commercials: The beat goes on. Journal of Broadcasting & Electronic
Media, 38, 91–101.

Simon, Adam F., & Jerit, Jennifer. (2007). Toward a theory relating political dis-

agendas as a predictor of voter choice in federal, state, and local elections
(Unpublished master’s thesis). Cleveland State University, Cleveland, OH.

Teaching of Psychology, 23, 220–226.

www.newstatesman.com/blogs/culture-tech/2008/10/usa-vote-obama-game

Simonton, Dean K. (1980a). Thematic fame, melodic originality, and musical zeit-
geist: A biographical and transhistorical content analysis. Journal of Personality
and Social Psychology, 38, 972–983.

Simonton, Dean K. (1980b). Thematic fame and melodic originality in classical
music: A multivariate computer-content analysis. Journal of Personality, 48,

Simonton, Dean Keith. (1981). The library laboratory: Archival data in personality
and social psychology. In Ladd Wheeler (Ed.), Review of personality and social

Simonton, Dean Keith. (1984). Melodic structure and note transition probabilities: A
content analysis of 16,618 classical themes. Psychology of Music, 12, 3–16.

Simonton, Dean K. (1987). Musical aesthetics and creativity in Beethoven: A com-

Guilford.


Spicer, Jeffrey. (2012). *The changing face of the Western: An analysis of Hollywood Western films from director John Ford and others during the years 1939 to 1964* (Unpublished master’s thesis). Cleveland State University, Cleveland, OH.


References


Thelwall, Mike. (2016, in press). Heart and soul: Sentiment strength detection in the social web with SentiStrength. In Janusz Holyst (Ed.), *Cyberemotions*.

References


Tov, William, Ng, Kok Leong, Lin, Han, & Qiu, Lin. (2013). Detecting well-being via computerized content analysis of brief diary entries. Psychological Assessment, 25, 1069–1078.


Manganello, & Martin Fishbein (Eds.), *Media messages and public health: A decisions approach to content analysis* (pp. 233–245). New York: Routledge.


References


Zhu, Junhuan, You, Quanzeng, Luo, Jiebo, & Smith, John R. (2013, December). Towards understanding the effectiveness of election related images in social media. Paper presented to the IEEE International Conference on Data Mining (ICDM), Workshop on Domain-driven Data Mining. doi:10.1109/ICDMW.2013.112


