Jurnal Teknik Industri, Vol. 24, No. 2, December 2022
ISSN 1411-2485 print / ISSN 2087-7439 online

DOI: 10.9744/jti.24.2.105-116

Content Analysis and Its Application with Dynamic Online Content: A Case Study

Toong Tjiek Liauw[1]*

Abstract: Content analysis is a well-established and widely used research method. In its early form, it was used extensively in the quantitative analysis of newspapers, and its applications later evolved to include electronic media such as radio and television. It has recently been applied to digital media, including the Internet. However, the use of content analysis in analyzing online content has been chiefly applied to static content, such as 'static' websites, in the early days of the Internet. Studies that involve its use in analyzing dynamic Internet content—for example, content that resides behind databases—are relatively much less common. This article is not written as a research paper per se. This article will instead discuss reflections on the efficacy of content analysis as a research method when applied to dynamic content such as DRs by using a previous study, which has applied content analysis to the dynamic content of digital repositories (DRs), as a case study. The previous study used as the basis for this article had applied content analysis to several DRs using manual counting by the researcher. In the process, several idiosyncrasies in terms of the way institutions populate their DRs with digital objects and the user metadata to facilitate discoverability of those digital objects were encountered that have introduced some 'complication.' This article will focus on how content analysis, as a research method, can be adapted to account for those idiosyncrasies to produce better results. This article will also identify the limitations and challenges of content analysis in dynamic online environments and offer some suggested approaches.

Keywords: Content analysis, quantitative method, digital repositories, dynamic content.

Introduction

As a "systematic analysis of text," content analysis (CA) has a long history dating to the 17th century with church-related studies (theology), which evolved into "quantitative newspaper analysis" in the mass communication era at the beginning of the 20th century (Krippendorff [1]).

Besides the classical debates, which will be discussed in the next section, CA has also evolved to accommodate new developments, including being applied to 'new media,' such as the Internet. Neuendorf [2] considered the use of CA as a means of analyzing Internet websites and cited several instances of the emergence of this method of research. With its dynamic nature, the Internet has introduced challenges in applying CA as a research method. As Rice and Rogers [3] have argued, "the natural contexts of new media may limit how faithfully traditional research designs and methods may be applied" and "the nature of new media themselves may create limitations, as well as new opportunities, for the kinds of research typically applied to mass media."

A previous study by Liauw [4] faced the challenges discussed above. The study was a research project that investigated several DRs in Indonesian higher education (HE) institutions. It applied content analysis by manually counting samples of digital objects in the DRs being investigated. Some of the challenges encountered were the various idiosyncrasies in how institutions have populated their DRs with digital objects, the use of metadata to facilitate the discoverability of those digital objects, and the dynamic nature of online content.

This article intends to reflect on the issues and challenges faced by CA – as a research method – when applied to dynamic content, using the previous study mentioned above as a case study. Consequently, this article – not being a research article per se – does not have the usual components, such as research questions, research objectives, etc. These components were present in the original research published by Liauw [4], as discussed above. Instead, this article is intended as a scholarly reflection focusing on some aspects of the adaptation of how CA has been applied to dynamic content on the Internet. The adaptation has been essential to applying CA to dynamic online content to minimize the undesirable effects caused by the unique nature of the content. Besides describing the challenges, this article will also try to offer some recommendations for a better application of CA in investigating dynamic online content.

[1] Faculty of Industrial Technology, Graduate Study of Industrial Engineering, Petra Christian University, Jl. Siwalankerto 121-131 Surabaya, 60238, Indonesia. Email: anugraha@petra.ac.id

* Corresponding author

Methods

Content Analysis as a Method

Content analysis has been defined as a "technique for the objective, systematic, and quantitative description of the manifest content of communication" (Berelson [5]). Although there have been multiple definitions of content analysis, they have usually "reveal[ed] broad agreement on the requirements of objectivity, system, and generality" (Holsti [6]). Holsti [6] described 'objectivity' as "rules and procedures" on which the research must be performed; 'systematic' as referring to the impartiality of the research process based on the formulated rules, and 'generality' as indicating the theoretical relevance of the findings.

Two main aspects of content analysis have been widely debated. Firstly, the relative merit of the method when used for quantitative and qualitative measurements (Krippendorff [1]). Lasswell *et al.* [7] stated that "[t]here is clearly no reason for content analysis unless the question one wants to be answered in quantitative." However, Holsti [6], by referring to earlier work by Lazarfeld and Barton [8], suggested that the method can also have qualitative applications, as "measurement theorists are generally in agreement that qualitative and quantitative are not dichotomous attributes, but fall along a continuum." George [9] has contributed to the discussions on CA by introducing the "non-frequency" content indicators, which he defined as "the mere presence or absence of a given content characteristic or a content syndrome within a designated body of communication," and which he regarded as "the non-quantitative or non-statistical variant of content analysis."

Secondly, it is the issue of manifest versus latent content. This distinction has been raised by previous researchers when they have noted that "content analysis analy[s]es not only the manifest content of the material" (Mayring [10]), and that in addition to the primary content (subject matter) of a work, there is also the latent content (contextual information) provided by the metadata (Becker and Lißman [11]). Berelson [5] has argued that there is no guarantee that different readers will comprehend the same manifest content and that "[t]o some degree the argument goes, every reader takes his peculiar meanings away from the common content." Neuendorf [2] suggested that a latent construct can be measured by using one or more manifest variables and provided an example by citing a previous work on the study of Internet websites by Ghose and Dou [12], where "the latent variable, 'interactivity'… was represented by 23 manifest variables that are easily measurable, such as presence or absence of a key word search, electronic couponing, online contests, and downloading of software." Schneider and Foot [13] referred to the "structural and feature elements of websites, hypertexts, and the links between them," as elements that potentiate and mediate "the relations between producers and users of web materials." Herring [14] agreed that besides referring to "the thematic meanings present in text or images" CA can sometimes refer to the structural or feature analyses of websites as mentioned by Schneider and Foot [13] above.

Besides these 'classical' debates mentioned above, the emergence of the Internet and the World Wide Web (WWW) has also introduced new challenges, albeit also new opportunities, for content analysis. Newhagen and Rafaeli [15] described "five defining qualities of communication on the Net" that are different from traditional mass media: "multimedia, hypertextuality, packet switching, synchronicity, and interactivity." Arguably three of these qualities (multimedia, hypertextuality, and interactivity) have had a significant impact on how content analysis can be applied to web-based media: the use of "mix multiple media including text, audio, graphics, animation, video, and even tactile and olfactory messages;" while hypermedia links of the WWW "has broken the shackles of linearity" and "overthrown the tyranny of author over reader" (Weare & Lin [16]). As a result, this new media has provided "the reader/user choice over the sequence and context in which material is consumed" and a degree of interactivity that "empowers users to become dynamically involved with the media," thereby giving reader/user "control over the program through which they are navigating, and consequently, the Internet moves from an author-centered to a user-centered, or decentered, the structure of information exchange" (Weare & Lin [16]). Along the same line, when elaborating on their "Web sphere analysis," Schneider and Foot [13] also talked about the inclusion of "analysis of the relations between producers and users of web materials."

Sampling and Coding

As with research in any field, it is virtually impossible when using CA to examine the whole universe (population) of any research object (Krippendorff [1]). This raises the issue of sampling, which has two functions. Firstly, sampling is essential in reducing the data that needs to be collected. In the context of CA the first step is "to list all members of the class of documents about which generalizations are to be made" (Holsti [6]). Tools such as lists, indices, directories, etc. can be utilized to define the sample. Secondly, sampling helps researchers to define the limit to which they can make generalizations based on the data gathered. However, an "adequate sampling design is a necessary but not a sufficient condition for validity" (Holsti [6]).

Krippendorff [1] defined three kinds of units in 'traditional' CA, namely: "sampling units, recording/ coding units, and context units," which he then described as follows: (with original emphases)

· "Sampling units are units that are distinguished for selective inclusion in an analysis."

· "Recording/coding units are units that are distinguished for separate description, transcription, recording, or coding."

· "Context units are units of textual matter that set limits on the information to be considered in the description of recording units."

In terms of sampling in the web-based media, Weare and Lin [16] cautioned that although "the Internet eases data gathering, its sheer size and mutability complicate the development of scientifically random samples." Due to the mutability aspect of the Internet, researchers have even argued that "selecting a truly random sample may be next to impossible" (Bates and Lu [17]). Weare and Lin [16] recommended "some popular method[s] to develop a sampling frame," including to use of search engine lists from collector websites and the most popular websites on the subject(s) under investigation, each of which has its advantages and disadvantages. At the same time, they also stressed that "news in an electronic, digital environment can be customized, or personalized, in a way not possible in other media. Organizations and individuals are usurping the editorial function by aggregating articles and other information on a common topic for specialized groups" (p. 283), a phenomenon that might challenge the validity of the research at hand. Weare and Lin [16] also stated that "[t]he majority of existing studies … define their sampling unit as a single Web site."

Based on her analysis of nineteen studies applying CA techniques to the WWW, McMillan [18] concluded that there are three types of coding units: "content categories" (the most common), "structural features of the Web site (e.g., links, animation, video, sound, etc.)," and "[the] 'demographic" characteristics of sites such as country of origin and type of institution that created the site … [or] the nature and/or purpose of the sponsoring organization in more detail." The studies she analyzed, however, did not produce any standard list of content categories, and she concluded that, "content categories seem to be specifically related to the goals of the given study" (McMillan [18]). Neuendorf [2] seemed to be in agreement when she argued that "[a]lthough the content analyst should consult both scholarly literature and commercial research and use theory as a guide whenever possible, he or she is, in fact, the boss, the final authority on what content needs to be examined and what variables ought to be taped," and that "variables to be included in a content analysis must reside in the message rather than the source or receiver." Weare and Lin [16] advocated "forsak[ing] exclusive reliance on the categorization of manifest message attributes" and instead "employing judgmental scales of Web site content" to enable a researcher to "measure holistic reactions of the audience that may be impossible to reduce to several manifest attributes." However, they also conceded that "[t]here are concerns about the reliability of judgmental measures." These coding units are the ones that will be used for comparisons, analyses, summaries, and the basis for inference-making (Krippendorff [1]).

Context units, on the other hand, "are not counted, need not be independent of each other, can overlap, and may be consulted in the description of several recording units," which "generally surround the recording units they help to identify … or be located elsewhere, such as in footnotes, indices, glossaries, headlines, or introductions" (Krippendorff [1]). Ha, and James [19] recommended the use of the home page of websites as a context unit, arguing that "it served as the front door of the entire website… [where] most visitors to a website decide whether they will continue to browse [the] site," as well as to "provide consistency across the sample since all units were a single page." This assertion was reinforced by McMillan [18], who concluded that "[t]he most common context unit used for [CA] studies was the 'Web site'." However, McMillan [18] also mentioned that "[m]any of the studies did not specify what was meant by the Web site," which could result in the home page, some pages, or all pages of the website being analyzed.

All these challenges and opportunities of the 'new media have had ramifications for how CA is being implemented as a research method. Based on an investigation of several DRs of Indonesian higher education (HE), this article discusses the challenges and opportunities regarding the application of CA in dynamic online content on the Internet.

The Investigation of Indonesian HE DRs

The previous study, which has served as a case study for this article, was the longitudinal study of Indonesian HE DRs that was conducted in two phases: (Liauw [4])

- Data Collection 1 (DC1): November 19th, 2014 to February 1st, 2015
- Data Collection 2 (DC2): December 1st, 2016 to January 20th, 2017

This previous study has adopted a quantitative-qualitative approach, in that although content analysis was utilized mainly to gather quantitative data, some additional qualitative assessments were also conducted. The quantitative component mostly involved "non-frequency" content indicators (coding units) as per George's [9] definition. The additional qualitative assessments, however, were neither quantified nor part of the coding schedule. They were only intended to be additional observations to inform the study regarding the local practices of Indonesian HE institutions in populating and managing their DRs.

Several limitations in this study need to be mentioned regarding the efficacy of content analysis when applied to DR websites. Firstly, the content analysis was applied to the metadata and documents contained in the DRs to gather information relevant to the characteristics and structure of the DRs. The content analysis was not applied to the individual documents or works to gather information pertaining to the topic or subject of each work. As an example, when analyzing a DR no attempt was made to gather information on the subject areas covered by the works contained therein. Instead, information was gathered on the various types of work represented (e.g. teaching materials, theses/dissertations, published materials, etc.).

Secondly, DR 'contents' reside behind a database, which means that they are not always available in the form of static web pages that can be analyzed as a whole representation of the website. They need to be retrieved using an interface that enables users to explore the 'contents' of the DR, either through the use of keywords/key-phrases in the search function or by browsing the DR's hierarchical structure. It is also the nature of DRs to contain records numbering from hundreds to hundreds of thousands, and these numbers can change (increasing or decreasing) as the DRs are investigated. In this circumstance it is not possible to analyze the whole 'contents' of a DR. A content analysis can only be completed by taking samples of the 'contents' (records), which can then be used to formulate indicative conclusion(s).

A further potential limitation is that data collection was undertaken by a sole coder, whereas Neuendorf [2] has recommended the use of "at least two coders, to establish intercoder reliability." Thus, the study did not fully satisfy Neuendorf's recommendation in this regard.

The previous study has resulted in several conclusions (Liauw & Genoni [20]). Firstly, the Indonesian HE institutions are still in a stage where they are still experimenting with their DRs. This conclusion was based on the fact that a number of DRs had less number of records when surveyed, compared to the data gathered from OpenDOAR. Some DRs had even changes their DR software. It was found, however, that Indonesian HE DRs have shown significant growth rate – in terms of number of DRs and deposit profile of the DRs – in DC2, compared to DC1. Secondly, Indonesian HE DRs were conceived more as corporate information management systems than as a response to the crisis in scholarly communication (Green Open Access/OA). This conclusion was based on results indicating that in Indonesian HE DRs, theses and dissertations were the most dominant type of work. The practices of populating and managing the DR contents has also lent a support to the conclusion. These local practices indicate that the availability and accessibility of the full-text documents in DRs served more as evidential records for administrative and management purposes than as a dissemination platform for scholarly content. Thirdly, it is "very likely that institutional prestige in terms of Webometrics ranking and the need to combat plagiarism have determined the growth and characteristics of OA in Indonesia more than the need to make Indonesian research visible and accessible."

Based on the previous study described above, this article will focus and discuss the efficacy, opportunities, and challenges of CA as a method when applied to the dynamic online content, such as in DRs. However, some background information on the design as well as the results of the previous study will be provided where appropriate. Details regarding the data collection process of this previous study can be found in an article titled "A Different Shade of Green: A Survey of Indonesian Higher Education Institutional Repositories" published in Journal of Librarianship and Scholarly Communication (Liauw & Genoni [20]) and a doctoral thesis titled "Institutional repositories in the Indonesian higher education sector: Current state and future prospect" (Liauw [4]). This article will not focus on the details of the data collection process from this previous study. This article, instead, will discuss some issues encountered in the data collection process that could potentially have some ramifications in the accuracy of the results, as well as the adaptations taken to mitigate them.

The sampling frame used in the previous study was the use of several online resources (directories and/or lists) relating to institutional repositories: Webometrics' Ranking Web of Repositories (July 2014 & July 2016 edition) at http://repositories.webometrics.info/en/Asia/Indonesia ; Open Directory of Open Access Repositories (OpenDOAR) at http://opendoar.org; and Registry of Open Access Repositories (ROAR) at http://roar.eprints.org. These online resources were used to compile a list of Indonesian HE DRs.

The sampling unit used in this study was the individual website of the DRs. Each URL of DRs resulted from the combined lists above was manually inspected to gather the relevant data. Table 1 presents the coding variables/units used in this study. DRs that were not accessible after three separate attempts (on three different dates) were excluded from the study.

Table 1. Coding variables for content analysis of Indonesian HE DRs (Liauw [4]).

Variables	Options	Type
	Demographics	
Acronym	N/A	Text
Institution or IR Name	N/A	Text
Year (of establishment)	N/A	Numeric
Status	State	Numeric (1 or empty)
	Private	Numeric (1 or empty)
Region	Java	Numeric (1 or empty)
	Bali-Nusa Tenggara	Numeric (1 or empty)
	Sumatra	Numeric (1 or empty)
	Kalimantan	Numeric (1 or empty)
	Sulawesi	Numeric (1 or empty)
	Maluku	Numeric (1 or empty)
	Papua	Numeric (1 or empty)
# Digital Objects	Manual	Numeric
	OpenDOAR	Numeric
	ROAR	Numeric
IR Software	DSpace	Numeric (1 or empty)
	Eprints	Numeric (1 or empty)
	GDL (Ganesha Digital Library)	Numeric (1 or empty)
	Other/In-house	Numeric (1 or empty)
Source/List Used	WEBO (Webometrics)	Numeric (1 or empty)
	OpenDOAR	Numeric (1 or empty)
	ROAR	Numeric (1 or empty)
Date of Inspection	N/A	Date
	Structural features	
Exploration Tools	B (Browse)	Numeric (1 or empty)
	S (Search)	Numeric (1 or empty)
Links	LI (Link to Institutional Website)	Numeric (1 or empty)
	LL (Link to Library Website)	Numeric (1 or empty)
	NL (No Link to Either)	Numeric (1 or empty)
Access Statistics	Y (Yes)	Numeric (1 or empty)
	N (No)	Numeric (1 or empty)
Collection Naming Practices	Good	Numeric (1 or empty)
	Fair	Numeric (1 or empty)
	Poor	Numeric (1 or empty)
	Content categories	
Types of Works	PUB (Published)	Numeric (1 or empty)
	UNPUB (Unpublished)	Numeric (1 or empty)
	THESES (Theses/Dissertations)	Numeric (1 or empty)
	TEACH (Teaching Materials)	Numeric (1 or empty)
	STDW (Student Works)	Numeric (1 or empty)
	UREC (University Records)	Numeric (1 or empty)
	SPEC (Special Collections)	Numeric (1 or empty)
	OTHER	Numeric (1 or empty)
Author Naming Convention	Y (Yes)	Numeric (1 or empty)
	N (No)	Numeric (1 or empty)
Standardized Access Points	Standardized Subject Headings	Numeric (1 or empty)
	Free-text Keywords	Numeric (1 or empty)
	Mix	Numeric (1 or empty)
	Not Available	Numeric (1 or empty)
Language of Access Points	English	Numeric (1 or empty)
	Indonesian	Numeric (1 or empty)
	Mix	Numeric (1 or empty)
Public Availability of Full-Text	All/Most (n > 90%)	Numeric (1 or empty)
	Some (25% <= n <= 90%)	Numeric (1 or empty)
	Minimal (0% < n < 25%)	Numeric (1 or empty)
	No Full-Text (0%)	Numeric (1 or empty)
Openness	OA (Open Access) - Public Availability of Full Text > 90%	Numeric (1 or empty)
	NOA (Not Open Access) - Public Availability of Full Text <= 90%	Numeric (1 or empty)

There were 52 DRs included in DC1 and 81 DRs in DC2.

Coding variables for content analysis of Indonesian HE DRs is presented in Table 1 below, divided into three categories to make it easier to understand: demographics, structural features, and content categories. The "Numeric (1 or empty)" in the "Type" column is meant to act as checking the "Options" box (with numeric "1") or leaving the box empty.

The context unit of this content analysis was the individual (metadata) records on the DR website. The individual (metadata) records might be sufficient as a 'deciding' factor but they might not be sufficient as a 'consistency' factor, as asserted by Ha and James [19], since different DR would have different number of (metadata) records.

The full list of DRs analyzed in this study (52 in DC1 and 81 in DC2) can be found as an external dataset (Liauw [4]). Subset or summary tables might be used in the discussions to highlight aspects of CA.

Results and Discussions

Challenges in Applying Content Analysis in Dynamic Online Contents

In general, data collection was straightforward since almost all, except one, of the coding variables were "non-frequency" variables. However, some complications were encountered when assessing and categorizing some variables that might potentially affect the validity of the study. The following section discusses these 'complications' and the steps taken to try to mitigate them.

Collection Naming Practices

Collection naming practices, to a certain extent, could influence discoverability of contents by enhancing or inhibiting navigability for users. Therefore, this coding variable was intended to assess discoverability and navigability. DRs that used single criterion to categorize contents at a certain level of collection hierarchy would enable users to easily navigate that DR to find content. The criterion could be based on Type of Work (published journal articles, teaching materials, university records, etc.); Type of Media (book, booklet, flyer, poster, etc.); Type of Content (text, image, video, etc.); or other aspects such as organizational structure. DRs that used multiple criteria to categorize contents at a certain level of collection hierarchy tended to confuse users. In this study the categories for collection naming practices were as follows:

- Good: collection naming used a single criterion at a certain level of collection hierarchy;
- Fair: collection naming used for more than one criterion at a certain level of collection hierarchy. The practice might cause some guesswork when navigating the collections, but in general the practice did not confuse users; and
- Poor: collection naming uses multiple criteria at a certain level of collection hierarchy. The practice definitely causes confusion for users in navigating the collections.

The initial criteria set in the planning stage (Good, Fair, and Poor) did not seem to be adequate to accommodate the various local practices in naming the collection—something that was not detected in the pilot phase of the CA. Firstly, the initial criteria were set to assess the usefulness of the collection names in assisting users to quickly grasp the scope of each collection in the respective DR by considering collection names. In other words, collection naming is a practical equivalent to the hierarchical structure of the collections, the naming of directories or folders per se in the respective DR. In the data collection phase, it was found that in some DRs the collection names or the hierarchical structure of the collections was sufficiently straightforward and easy to comprehend. However, these DRs also have some local practices that were deemed confusing for users. For example, some DRs placed documents related to (or manifestations of) the same work into separate records. This practice usually involved the main documents (Microsoft Word or Portable Document Format/PDF) and the presentation slides related to the main documents. Thus, although the collection naming practices were considered to be "Good," the overall collection management practices of the DRs could potentially create confusion for users. Two effects were apparent:

- the criteria set for the variable (Collection Naming Practice) were not adequate to quantify the assessment on the navigability and discoverability of contents in DRs or there was a mismatch between the two due to a condition that was not detected earlier in the design or pilot phase; and
- during the CA process in DC1, the quantification of "Collection Naming Practice" has been influenced by the bias from another judgment not explicitly set in the criteria for the variable in question, namely practices in how document(s) of the same work is/are recorded in the DR.

Since it was obvious then that the initial criteria were no longer useful for assessing the navigability of DRs, they needed to be expanded to include the local practices as mentioned above. However, expanding the initial criteria to include local practices may make them too broad and too time consuming to be executed. Especially since, as mentioned earlier, local practices of Indonesian HE institutions in populating and managing their DRs would only be gathered as additional qualitative observations. They were not intended to be part of the quantitative data collected for a CA study.

Secondly, based on further observations of the DR software used, an observable trend emerged that certain DR software (Eprints) has provided generic categorization or hierarchical structure for the collections based on organizational structure of the institution, while also providing the possibility to completely modify the structure based on another aspect. This generic structure has helped in providing an easy-to-navigate environment. Most DRs that used Eprints have kept this generic structure to manage their collections, with very few exceptions. It was observed that other DR software (DSpace, Ganesha Digital Library/GDL, or Other/In-house) have taken a different approach by allowing users or institutions to define the collections freely. Thus, DRs that have used software other than Eprints have tended to get more negative assessments in term of their ease-of-use compared to Eprints.

The fact that the situation described above was only fully realized after DC1 was completed has only exacerbated the problem. Thus, it was decided that the "Collection Naming Practice" coding variable would not be used in DC2 since the results would definitely be biased and skewed by the DR software used. Had there been more than one coder, this situation might have been avoidable since this issue would have come up during the early phase of the study when the coders would need to reconcile any differences in their coding.

A lesson could be learned from the case discussed above. There is always a risk in conducting additional qualitative assessments while doing the quantitative part of CA on dynamic online content. The same risk does present in an offline environment, but in an online environment the 'slippery slope' is steeper since the 'temptation' to investigate further any perceived anomaly is always just a click away. As in the case of the "Collection Naming Practice" discussed above, the temptation for further investigations had led to biased results. This 'temptation,' however, does not seem to be the only problem encountered in this study, as is discussed in the following section.

Types of Works and Public Availability of Full-Text

The "Types of Work" variable in this study consisted of several categories: published works (PUB); un-published works (UNPUB); theses and dissertations (THESES); teaching materials (TEACH); student works (STDW); university records (UREC); special collections (SPEC); and other (OTHER). Due to space consideration, reasoning and explanations on the categories will not be elaborated here but can be found in Table 3 of the article by Liauw and Genoni [20].

Information on the types of work was gathered by taking at least three sample records from each smallest unit of the collection in the repository and inspecting the metadata. It was only necessary to detect the existence of the different types of work in the IRs, and no attempt was made to calculate item (digital objects) counts for each type of work.

In terms of detecting the existence of certain type of work, it was not sufficient to only inspect the metadata, for two reasons. Firstly, the "Type of Work" variable in this study was categorized differently than those categories provided by each institution using different DR software ("Item Type" in Eprints and "dc.type" in DSpace). Secondly, there have been numerous cases identified where Indonesian HE DRs have used the default categories provided by the DR software incorrectly or have customized them entirely to fit local needs. As an example, in an DR the "Article" category has been used for journal articles, conference papers, and newspaper article. Thus, inspection(s) needed to be conducted to the file(s)/digital object(s) in each record to ascertain the type of work associated with each record.

Besides identifying the types of work, further inspection was conducted for each sampled record to collect data relating to "Public Availability of Full-Text" variable. This was done by examining all file(s) or digital object(s) attached to each record to ascertain several characteristics:

- whether the full-text document(s) reflected the type of work recorded in the metadata;
- whether the full-text document(s) represented the complete work as reflected in the metadata; and
- whether the full-text document(s) was/were publicly accessible (no protection or access-restriction).

The number of records with publicly accessible full-text document(s) was recorded in "Open Access" (OA) category while the ones with fully or partially protected or restricted access were recorded as "non-Open Access" (non-OA) category. The "Openness" variable was added for the sole purpose of making it easier for the researchers to count the DRs with (presumed) OA policies in place (publicly accessible full-text document found in >90% of the sampled records) and DRs without OA policies (publicly accessible full-text document found in <=90% of the sampled records).

The fact that this CA was conducted to gather both quantitative data and some qualitative observations has introduced another complication. As mentioned above, the data on the types of work were gathered by taking at least three sample records from each smallest unit of (digital object) collection in the repository. This meant that the sample records examined would have only been three records had no 'perceived anomaly' (unusual practices in managing or populating the DRs) been found. However, additional sample records were examined in cases where such an anomaly was found, in order to investigate further the nature of the 'perceived anomaly.' These additional records examined had been taken into account in the calculation of the percentage of the availability of full-text documents in the respective DR. Thus, the total number of records sampled in one DR might or might not be identical to other DRs. In a strictly quantitative CA this was not an ideal situation since it might introduce a bias; or at least the public availability of full-text documents could not be, strictly speaking, compared to one another. This condition had been identified as an additional limitation to the one mentioned earlier in this section.

In terms of the "Public Availability of Full-Text" variable, University of Indonesia's DR provided a good opportunity to test the credibility of the record sampling technique employed in this content analysis study of DRs. The University's DR has separated each collection into two categories and labelled them respectively with "Open" where the full-text document(s) of each work was publicly accessible, and "Membership" where full-text document(s) of each work was accessible only to members (internal/ campus members). Manual inspections of each category, by taking some random sample records, had proven that accessibility status of the full-text document(s) of each work was consistently enforced. Since the DR software has provided record counts for each collection, exact calculations – based on these categories – could be made to identify the number of records where the full-text document(s) of each work is publicly accessible. This number could then be divided by the total number of records in the DR to produce the percentage of records in the DR, where public access was granted. The figure obtained was 40%. This figure was very close to the figure obtained by taking random samples of at least three records from each of the lowest hierarchies of collection in the DR (43.9%).

There were, however, some complicating factors in terms of the data collection for the "Public Full-Text Availability" variable. Firstly, there were some legitimate reasons why certain works in DRs could not be accessed, which should be prevented from affecting the "Public Full-Text Availability" criterion. When examining the Satya Wacana Christian University's (UKSW) DR, it was found that some records prevented public access to the full-text document(s) due to incomplete administrative document(s), such as author consent page and author's no-plagiarism-statement page. These reason(s) is/are explicitly stated in the records. In this case, the study has excluded these records from being counted in determining the value for the "Public Full-Text Availability" variable.

Table 2. Distribution of collections in Computer Science College's (STIKOM) and Sunan Ampel State Islamic University's (UINAMPEL-DL) DRs (Liauw [21]).

Type of Work	STIKOM		UINAMPEL_DL	
	# of Records	%	# of Records	%
Article	20	1.15	332	2.94
Book	3	0.17	71	0.63
Book Section	11	0.63	8	0.07
Conference or Workshop Item	190	10.91	261	2.31
Thesis	1,517	**87.14**	10,615	**94.02**
Other	0	0	3	0.03
TOTAL	1,741	100	11,290	100

Secondly, the sampling method, which required a minimum of three records randomly sampled from each of the lowest hierarchies of collection in the DR, would introduce an unintended bias in DRs that had very 'skewed' distribution of collections. Two cases in DC2 could serve as examples: Computer Science College (STIKOM) and Sunan Ampel State Islamic University (UINAMPEL-DL). Table 2 shows the distribution of collections in both DRs.

Both DRs had heavily skewed distribution of collections, where Thesis collection consisted of 87.14% and 94.02% of the total DR contents respectively. In DRs with heavily skewed distribution of collection such as these, taking the same number of sample records from each collection could potentially misrepresent the characteristics of the DRs, especially those that relied on the number and composition of the sampled records, which in this study it related to the "Public Full-Text Availability" variable. Taking Computer Science College's (STIKOM) IR as an example, which was not the most extreme case of an DR with a skewed distribution of collections, we find the DR had five different collections. Assuming that the study took three sample records from each collection, there would be fifteen sample records in total; giving each sample record an equal weight of 6.67%. Two sampled records from the "Book" collection (with the least number of records) that didn't provide public access to the full-text document(s) of the respective work would have been enough to drop the DR's degree of openness; a drop of 13.33% to 86.67%. In reality, when we took into account the distribution of collections in this DR, these two records would have only been worth 0.11%; keeping the DR's degree of openness at 99.89%. On one hand, had the study stuck rigidly to the same number of sampled records in each collection, the result would not have been representative of the DR being surveyed. On the other hand, had the study strived to represent all the collections in an DR, it would have had to take too many sample records in the dominant collection to maintain equal weight among the sampled records; an impossible task to undertake manually.

As a compromise of these two difficult choices, the study adopted a 'middle ground' approach, which enabled it to represent – to certain extent – all collections proportionately while at the same time manually doable. This approach consisted of two steps. In the first step, a minimum of three sample records were selected as usual; resulting in a percentage for the degree of openness of the DR. In the second step, a number of additional records were selected randomly from the dominant collection(s) to determine the consistency of public accessibility status of the full-text document(s) associated with each record in the collection. In the case where all or most of the additionally-sampled records did have consistent status, then the percentage for the degree of openness for the DR was:

- determined only by the additionally-sampled records from the dominant collection(s) if the dominant collection(s) singly or collectively comprised of nearly or more than 90% of the whole DR contents; or
- determined by the sampled records and additionally-sampled records from the dominant collection(s) if the dominant collection(s) singly or collectively comprised of fairly less than 90% of the whole DR contents.

In the case where the additionally-sampled records did not have consistent status, then only the calculation from the first step was taken as the percentage for the degree of openness of the DR.

This proportionality versus do-ability problem can easily be alleviated in future similar studies by employing a more automated process of sampling the records in an DR proportionately based on the makeup of its collections and/or taking into account the distribution of records based on the Year. However, in terms of Indonesian HE DRs, it is also important to anticipate local practices where they:

- provided one or more file(s)/digital object(s) associated to the work in the respective record but not in its entirety, meaning that only some part(s) of the complete work was/were provided; and/or
- provided one or more file(s)/digital object(s) associated to the work in the respective record that represented the complete work but restricted public access to one or more of the file(s)/digital object(s).

The second problem might be easier to anticipate technically but the first problem will be much more difficult to tackle in an automated process.

Besides the challenges identified above, more have emerged in the context of longitudinal CA study.

Challenges in Longitudinal Study Context and Identifying Opportunities

That DC2 was a repeat of DC1 (longitudinal study) also posed some complications due to some changes that had happened in a number of institutions between the execution of DC1 and DC2. Some of the changes were merely 'cosmetic' in nature, such as changing the URL or switching URLs between the institutions' DR and another system (institutional website, blog, library's online catalog/OPAC, etc.). However other changes were more substantial, such as change of DR software, some features of the DR not functioning properly, broken links to the full-text document(s), etc. These changes could potentially introduce some kind of bias into the longitudinal study since they have rendered the comparisons – in the stricter sense of the word – between the results from DC1 and DC2 impossible to be undertaken. Speaking about their study on blogs, Herring *et al.* [22] had hinted at this challenge when they stated that "[p]erhaps the greatest challenge in analyzing blogs longitudinally lies in identifying comparable samples at different points in time." This is another further limitation of this study.

Also, according to Herring *et al.* [22] "[t]he results of the longitudinal content analysis can be grouped into three categories: change, stability and variability." They defined "change" as "a clear pattern of increasing or decreasing over time," and "stability" as "characteristics that do not change appreciably over time," while "variability" as "results that do not show a clear directional pattern, but rather fluctuate from sample to sample." As was evidenced from this study, coding variables that were "demographic" and "structural features" in nature tended to be stable, with some exceptions when the institutions had changed their DR software. On the other hand, coding variables that were "content categories" in nature showed observable changes between DC1 and DC2. This study did not observe any variability issue since the longitudinal study was done only in two different periods of time. Variability would only be possible to be observed in longitudinal studies that involves at least three different periods of time.

Conclusion

The discussions thus far have led to some realizations regarding the efficacy of CA when implemented to dynamic online content on the Internet, especially to contents in DRs. The dynamic nature of the Internet, as an infrastructure as well as in terms of its ever-changing content, has introduced challenges to and demanded improvisations in the application of CA.

The customizability of Internet and/or web-based application/software – in this context, the DR software

– has made it almost impossible to conduct pilot study that could anticipate all, or even most, variations of the way application/software is customized to fit local needs and practices. This condition might result in complications in any studies, where these complications were not foreseen during the pilot phase and were too difficult to mitigate while maintaining the do-ability of the study once the study has started.

Conducting additional qualitative assessments while doing the quantitative part of CA based on dynamic online contents will always carry with it some risk of introducing bias to the study. The dynamic online environment has made it far too easy for any researcher(s) to 'deviate' from the original design of the study since further investigation is always just a click away; a situation that would only be exacerbated in CA studies with sole coder.

In addition to the challenges identified above, the Internet and the web-based media have also provided opportunities for the application of CA. Firstly, the digital nature of the Internet and web-based media has enabled the availability of abundant sources of data that are ready to be assessed and processed using digital means. Although some researchers have argued that this is also a challenge in terms of sampling, this abundant availability of data has nevertheless opened so many new opportunities for CA studies. Secondly, despite the current technological limitations, the Internet and web-based media have created opportunities for automating the CA processes. In the case of this study, an automated CA would have been able to perform much better sampling of records in DRs by taking into account the proportion of individual collection in each DR. An automated CA might have even been able to conduct assessments for the whole contents or records in an DR in much less time compared to manual CA.

In the context of representativeness of CA results, human involvement and/or judgements will still be crucial since it's still important to take into account local practices and idiosyncrasies of the media being investigated. Thus, a combination of automated and manual CA is still the preferred choice.

References

1. Krippendorff, K., *Content Analysis: An Introduction to Its Methodology*, 3rd ed., Sage Los Angeles, CA,, 2013.
2. Neuendorf, K. A., *The Content Analysis Guidebook*, Thousand Oaks, CA, 2002.
3. Rice, R. E. and Rogers, E. M., New Methods and Data for the Study of New Media. In R. E. Rice (Ed.), *The New Media: Communication, Research, and Technology* (pp. 81-99), Sage, Beverly Hills, CA, 1984.
4. Liauw, T.T., *Institutional Repositories in the Indonesian Higher Education Sector: Current State and Future Prospect*, Thesis, Faculty of Humanities, Curtin University, Western Australian, 2018, https://espace.curtin.edu.au/handle/20.500.11937/73546.
5. Berelson, B., *Content Analysis in Communication Research*, Hafner Publishing Company, New York, NY, 1952.
6. Holsti, O. R., *Content Analysis for the Social Sciences and Humanities*, Addison-Wesley, Reading, MA, 1969.
7. Lasswell, H. D., Lerner, D. and Pool, I. d. S., *The Comparative Study of Symbols*, Stanford University Press, Standford, CA, 1952.
8. Lazarfeld, P. F. and Barton, A. H., Qualitative Measurement in the Social Sciences: Classification, Typologies, and Indices. In D. Lerner and H. D. Lasswell (Eds.), (*Vol. The Policy Sciences: Recent Developments in Scope and Method*, pp. 155-192), London, UK, 1951.
9. George, A., Quantitative and Qualitative Approaches to Content Analysis. In K. Krippendorff and M. A. Bock (Eds.), *The Content Analysis Reader* (pp. 144-155), Sage, Los Angeles, CA, 2009.
10. Mayring, P., Qualitative Content Analysis. *Forum: Qualitative Social Research*, 1(2), 2000.
11. Becker, J. and Lißmann, H.-J., *Inhaltsanalyse - Kritik einer Sozialwissenschaftlichen Methode. Arbeitspapiere zur Politischen Soziologie 5*, Olzog Verlag GmbH, Munich, 1973.
12. Ghose, S. and Dou, W., Interactive Functions and Their Impacts on the Appeal of Internet Presence Sites. *Journal of Advertising Research*, 38 (2), 1998, pp. 29–43.
13. Schneider, S. M. and Foot, K. A., The Web as an Object of Study, *New Media & Society*, 6(1), 2004, pp. 114-122, doi:10.1177/1461444804039912.
14. Herring, S. C., Web Content Analysis: Expanding the Paradigm. In J. Hunsinger, L. Klastrup, and M. Allen (Eds.), *The International Handbook of Internet Research* (pp. 233-249), Springer Verlag, 2013.
15. Newhagen, J. E. and Rafaeli, S., Why Communication Researchers Should Study the Internet: A Dialogue. *Journal of Computer-Mediated Communication*, 1(4), 1996, pp. 0-0, doi:10.1111/j.1083-6101.1996.tb00172.x.
16. Weare, C. and Lin, W.-Y., Content Analysis of the World Wide Web: Opportunities and Challenges, *Social Science Computer Review*, 18(3), 2000, pp. 272-292, doi:10.1177/089443930001800304.
17. Bates, M. J. and Lu, S., An Exploratory Profile of Personal Home Pages: Content, Design, Metaphors. *Online & CD Rom Review*, 21(6), 1997, pp. 331-340.
18. McMillan, S. J., The Microscope and the Moving Target: The Challenge of Applying Content Analysis to the World Wide Web. *Journalism and Mass Communication Quarterly*, 77(1), 2000, pp. 80-98, doi:10.1177/107769900007700107.

19. Ha, L. and James, E. L., Interactivity Reexamined: A Baseline Analysis of Early Business Web Sites. *Journal of Broadcasting & Electronic Media*, 42(4), 1998, pp. 457-474, doi:10.1080/08838159809364462.
20. Liauw, T. T. and Genoni, P., A Different Shade of Green: A Survey of Indonesian Higher Education Institutional Repositories, *Journal of Librarianship and Scholarly Communication*, 2017, doi:10.7710/2162-3309.2136.
21. Liauw, T. T., *Longitudinal Content Analysis of Indonesian Higher Education Institutional Repositories (Nov 19, 2014 - Feb 01, 2015 and Dec 1, 2016 – Jan 20, 2017)*, 2017, [Dataset], https://dx.doi.org/ 10.4225/06/55B074C6EC97E
22. Herring, S. C., Scheidt, L. A., Kouper, I. And Wright, E., Longitudinal Content Analysis of Blogs: 2003-2004. In M. Tremayne (Ed.), *Blogging, Citizenship, and the Future of Media* (pp. 3-20), Routledge, New York, NY, 2012.