CENSORSHIP GLOSSARCHIVE PROJECT PHASE ONE: DEVELOPING METADATA SCHEMA FOR CRYPTIC CIRCUMLOCUTIONS IN CHINESE SOCIAL MEDIA

A THESIS SUBMITTED TO THE GRADUATE DIVISION OF THE UNIVERSITY OF HAWAIʻI AT MĀNOA IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER

OF

LIBRARY AND INFORMATION SCIENCE

MAY 2014

By

Matthew C. da Silva

Thesis Committee:
Andrew Wertheimer, Chairperson
Noriko Asato
Hong Jiang

Keywords: Social Media, Metadata, Censorship, China, *Weibo*

6489

Dedicated to the victims of the Tiananmen Square massacre on the 25th anniversary of this tragic event.

ABSTRACT

Chinese censors constantly evaluate and sanitize the electronic speech of her citizens in the name of service to the greater good. Yet while the hidden Internet enforcement agencies with specialized tools seek to enforce vaguely defined and sometimes shifting standards of expression, China's people continue to offer their own resistance ranging from thoughtful critiques to wisecracks, through both overt speech and cryptic circumlocutions.

Little is known about what China's users know about the censors or what the censors know about their fellow cyber citizens. This study seeks to provide a basis by which we may begin to expand our knowledge of Internet censorship as practiced in the Peoples's Republic of China and, by extension, the rest of the world through a cyber-ethnographic examination of the politically sensitive posts on China's version of Twitter, *Weibo* with an emphasis on the deliberately deceptive practice of cryptic circumlocutions. An analysis of the known cryptic circumlocutions shows that while they may be arranged into a taxonomy, the metadata schema does not require strict adherence to the taxonomy in order to adequately catalog each discrete instance of Internet communication. The metadata schema proposed here is for the cataloging of various types of discrete Internet communication ranging from entire sites, articles on news sites, to blog comments or social media postings with an eye towards indexing the nature of sensitive expression as well as the type of cryptic expression.

TABLE OF CONTENTS

LIST OF FIGURES

PREFACE

This thesis represents the preliminary step in the establishment of a project dedicated to the study of internet censorship and efforts to resist it. The concept grew from a linguistic curiosity over the manner in which Internet users, particularly those in China and especially those users on the *Weibo* social media website, invent ways of making oblique references to risky topics in order to throw censors off their trail. I was fascinated by the creativity users demonstrated in innovating new ways to generate references meant to talk over the heads, as it were, of the machines and speak directly to their human compatriots. Linguistically, the practice is intriguing because, unlike typical exchanges, it employs deliberate obfuscation — a tactic which itself begs many questions regarding motivating logic and efficacy.

Originally, I intended to create an electronic glossary of cryptic Chinese terms to serve as a working aid for Sinologists and armchair China-watchers who might be stymied by the occasional cryptic reference. While that is still the goal, I realized it would also be necessary to catalog examples of where these came from in order to provide evidence for what will be a constantly evolving relation between word and meaning. In other words, the product is composed of a glossary for the linguistic elements as well as a digital archive to house digital specimens of their use. For lack of a better term, I have coined the word "glossarchive" to describe this hybrid tool.

The collection of these proofs require certain guidelines for indexing. These guidelines, known in library and information science as a "metadata schema," are presented in this thesis. One of the side benefits of this schema is, because it was developed to deal with an added layer of obfuscation, a simplified version of it can still be used in the collection and cataloging of more day-to-day social media samples that can be found in places such as Twitter where censorship is not an issue.

It became clear there were other potential benefits as this endeavor began to take on a life of its own. Other possible developments include, a censorship index for China and other countries, as well as tools and information that could be put to good use by organizations like Freedom House and the Free Weibo, to name a few. My hope now is the schema will be adopted, not only in the execution of this endeavor but in future cyber studies as well.

INTRODUCTION

The Internet can be seen as both a socially disruptive technology and an instrument for shaping the "collective" will. To lose sight of either half of this faustian equation denies the Internet's full impact. Anyone can wield the power of the Internet, be they activist or terrorist, regardless the nobility of intent.

This is true of many other tools of humankind: language, art, mathematics, writing and so forth. Where the Internet differs quite drastically from all these (even as it incorporates all these) is its intermingling of the subjective human will with an independently existing objective reality. It is a product of human desire and creativity which augments our shared reality even as it is informed by it. Today in many respects we live with the Internet as we live *through* it. This process where the subjective is objectivized and the objective is subjectivized is cyclical and may even be categorized as a type of feedback loop, but only if left unchecked.

An example of a controlled feedback cycle can be found in the Internet as governed and utilized in the PRC today. This is where social media can be both disruptive to those who are both for and against the status quo. On the one hand, you have activists yearning for rights and freedoms most people could not live without — the right to speech, to worship and assembly, to control one's own body, or have a say in the size of one's own family. On the other hand, you have the single ruling party and its desire to maintain social stability, and preserve the social contract through the curtailing of acts such as gambling, pornography, drugs, and criminal cyber activity. Though various views are represented on weblogs, the news sites, and social media, the government reserves the right to pull the plug.

Electronic speech censorship and information control are not uniquely Chinese phenomena. The debates rage on in the United States and abroad about speech, freedom, and privacy on the Internet. Article 19 of the United Nations' Universal Declaration of Human Rights underscores the right of all people to the freedom of expression, "Everyone has the right to freedom of opinion and expression; this right includes freedom to hold opinions without interference and to seek, receive and impart information and ideas through any media and regardless of frontiers." According to the Freedom House "Freedom on the Net 2013" report, China ranks above only Cuba and Iran overall but

would be worse if in the metric measuring violations of user rights, scoring 38 out of a possible 40 points. More frightening is the increasing impunity with which China's censors convince non-Chinese news media outlets to drop stories even going so far as to resort to threats or actual physical violence (Cook 7). Already we are influenced by the Chinese information paradigm.

The American Library Association (ALA) has been a participant and progenitor of this debate having long fought at the forefront with cultural and governmental entities in an effort to protect the rights of all people to unfettered access to libraries and subsequent modes and media of information. Article 9 of the ALA's 1939 *Code of Ethics for Librarians* states the librarian is charged to "study the present and future needs of the library, and should acquire materials on the basis of those needs" while also exhorting provisions to a "wide range of publications" and "varied viewpoints" while article 19 prohibits librarians from turning "the library's resources to personal use, to the detriment of services which the library renders to its patrons."

Librarians have always wrestled with the task of balancing the needs of their patrons with the dictates of their governing bodies be they schools, churches, universities, or local governments. Over the past century, the ALA has stood for patron rights during dark periods of the republic by drafting the Library Bill of Rights in response to Cold War McCarthyism in 1951 and again several times thereafter to address other issues such as self-censorship through book selection and making public spaces available to all. The introduction to the 1981 ALA *Statement on Professional Ethics* expresses a belief in an informed public as necessary part of a strong democratic union. "In a political system grounded in an informed citizenry, librarians are members of a profession explicitly committed to intellectual freedom and the freedom of access to information. We have a special obligation to ensure the free flow of information and ideas to the present and future generations."

In the information age, the definitions of words such as "publishing" and "intellectual content" are being stretched beyond former recognition. Almost anyone can broadcast their thoughts on the internet and some users do so without fully realizing how many others or exactly who is receiving these opinions. Yet just because the project of

intellectual content creation and publishing is more easily entered into and more broadly practiced does not necessarily call for stricter limits on what can and cannot be expressed.

In the PRC, however, it is much easier to transgress the law when engaging in expression. From politically sensitive topics like Taiwanese independence to specific yet innocuous words such as "river crab" the censors are continually at work evaluating and sanitizing the electronic speech of her citizens all for the sake of *shehuizhi'an* (社会治安) or public stability.[1] Yet while a shadowy world of Internet enforcement agencies and specialized tools seek to enforce vaguely defined and sometimes shifting standards of expression, China's people continue to offer their own resistance in the form of thoughtful critiques and wisecracks, through both overt and surreptitious speech.

Little is known about what these users know about the censors, what the censors know about their fellow cyber citizens, and who is the more effective. This study seeks to provide a basis by which we may begin to expand our knowledge of Internet censorship as practiced in China and, by extension, the rest of the world. The metadata schema proposed here is for the cataloging of discrete Internet expressions ranging from entire sites, articles on news sites, to blog comments or social media postings with an eye towards indexing the nature of sensitive expression as well as the type of cryptic expression, if one exists.

It is hoped this schema or one like it may be used by computer scientists, cyber-ethnographers, political scientists, activists, intelligence analysts, and advocates to track the changes in online speech and see further behind the curtain of government secrecy and discern more clearly the types of tactics, training, organization, and techniques they use as well as the developing responses employed by their targeted users. One of the

---

[1] This term *shehuizhi'an* is also translated "public security" and is related to *gong'an* (公安). Where the *gong* (公) character is commonly translated as "public" and the *an* (安) character encompasses separate yet overlapping concepts in English such as peace, stability, or security. In the current Chinese constitution, three out of five uses of the term *gong'an* refers to state organs or forces charged with maintaining order. *Shehuizhi'an* on the other hand, is used thrice in the same constitution and every time as a restrictive legal concept. There are other terms, such as *shehuizhixu* (社会秩序) which refers to social order and is also invoked as a restriction on citizens' right in articles on speech and religion (articles 28 and 36, respectively). There are other types of orders *zhixu* and stabilities *zhi'an* referenced in the Chinese constitution as well.

reasons why China is an important entity in this game of cyber cat and mouse is because of the unique visual qualities of written Chinese which gave rise to a number of tactics which allow for some characters to act as seemingly unrelated stand-ins for known trigger words or phrases (such as proper nouns or government euphemisms) which might call too much attention to the user's posting. By using words which invoke other terms either by visual phonetic cues, the clever user can say one thing, but mean another. These techniques seem primarily designed to fool mechanical text collection and detection tools which will be discussed in more detail in the following sections.

The cryptic circumlocutions (literally "talk arounds") present a unique challenge in the collection and cataloging of politically sensitive and therefore endangered speech due to their deliberately deceptive semantic and linguistic traits. It is for this reason they are an appropriate starting point for the study of this complex matter. It is hoped that by working through the issue of social media cataloging at this level of complexity will provide a type of template for handling other issues which will inevitably arise as the study spreads to other linguistic and cultural areas such as Turkish, Persian Farsi, Russian, and English.

This paper is organized into sections beginning with the basic known and assumed mechanics of Internet data collection and aggregation. From there, we will see an overview of the research collection process which can be described as cyber-ethnography of politically sensitive posts on *Weibo*. This is followed by an analysis of known types of cryptic circumlocutions. The section following this deals with the issue of circumlocutions in a more abstract way, approaching the phenomenon as a type of metaphorical device. The final two sections detail the proposed cataloging schema as well as some proposed areas and avenues for future research. At the end the reader will find a glossary of the main terms used in this work and an appendix with suggested metadata tags.

TEHCNOLOGICAL AND LEGAL LANDSCAPE

Although the Internet is a tool used by broad segments of modern society, details regarding its daily operations are not understood even among some information experts. We do not need to perform an exhaustive review for the purposes of this research, but for the sake of clarity it is beneficial to review key background areas. The following section gives some detail concerning the Internet in general and its utilization in China in particular.

How the Internet is Organized

Who organizes the Internet and how? At this stage of development, one may well argue the Internet organizes itself. In the early days, however, the project of Internet organization was anything but automatic. Now, while web masters acquiring web hosting involve a transaction between two people, the standards upon which they rely, those dictating URL assignments and Internet Protocol addresses for example, were created by governing bodies such as the International Standards Organization, the National Information Standards Organization, the Unicode Consortium or Worldwide Web Consortium (W3C), as well as universities and corporate entities like Microsoft.

Most people do not think about the role these organizations play in their daily lives. By the time we get around to accessing information through our electronic resource of choice the influence of these organizations is relegated deep within the background of a smoothly running information universe. It is this topmost and most visible layer where the user searches for and accesses their information. This is similar to the train passenger who is less likely to dwell on the hard labor that went into laying down miles of tracks while focusing on their immediate environment: the conductor, the engineer, and the engine itself.

The Internet user also admires and appreciates that which is most apparent. On the Internet, however, the networks, browser, and search engines are soulless and mechanical in contrast to train conductors and engineers. The Internet is too vast, too fast-moving, and too diverse for a reasonably sized team of human beings to organize. The bulk of this labor is taken up by programs bouncing from site to site, clicking links and analyzing the textual and image content of each page before going on. These programs are a type of "bot" (ie, an automated script) called a "crawler". Crawlers are so named because they

"crawl" over the "web" of the Internet, taking data and metadata from the pages they visit and sending it to databases for indexing. These crawlers are why search engines know what sites are most likely to satisfy user queries.

It is through this automatic eye that most of us receive our information from the Internet but though this method is economical in terms of the speed, it has known weaknesses inherent to their machine nature. One of these shortcomings is the crawler's inability to fill in information when it comes across a form, meaning that any site which asks for a name, address, or other such information requiring human cognition and informed response is a dead end for crawlers. One way websites protect themselves from being overrun by non-human bots and spammers is through the deployment of Captcha text input requirement. Captcha is a rough acronym which breaks out to "Completely Automated Public Turing test for telling Computers and Humans Apart" and serves as a test of the user's humanity by asking them to recognize alphanumerics which are distorted beyond a bot's capability to machine read them (Engber). Bots are not only used by search engines to index websites, but they can also infiltrate forums or blog comment strings to either collect information or pester users with spam ads and unsolicited advertising. The Captcha program weeds bots both with distorted text and input fields. Bots have difficulty understanding what should go into some input fields and therefore fail to fully index sites which utilize fields. For example, a simple Google search for "sex offenders near 123 My Street" will not return the number and addresses of all the registered sex offenders in one's area. This information can be accessed by going to the local government website, finding the sex offender registry, and then inputing one's own home address for the most up to date results. Even though the information exists on the Internet, search engines cannot provide the answer to the query because bots are stymied by input fields.

According to Sherman and Price (2003) this indexed portion of the web is "vast" with 20 percent of sites failing to register with search engines and others avoiding detection, by accident or by design, through the use of forms, interfaces, or password protection (296-298). One estimation says this leaves fully almost 90 percent of the Internet unexamined by the very programs which we are most reliant on to access

information content on the web (Gill). This uncharted territory is known as the "invisible" or "dark" web.

Despite the fact crawlers are such blunt instruments, they are still undoubtedly useful tools in searching the Internet for all manner of content, even questionable content. While the King, Pan, and Roberts study demonstrates the ultimate goal of censorship on *Weibo* is to block horizontal communication, especially when it comes to organizing protests or other such action, the process of finding horizontal communication and tagging it for content may still be a task best left to crawlers and other automated programs.

It is reasonable to assume that on some level Chinese users are cognizant of these tools, even if only by intuition. The very existence of cryptic homophones reveals a mindset pitting human rationality against unthinking automated processes. The human mind cannot only read what is on the screen, but what is said in between the lines. As anyone who has used automatic translation can attest, programing has not yet reached that level of sophistication. Though the process of detection and weeding out is not entirely automated, to a certain extent "users' intuition" attests to an understanding that the successful transmission of a controversial message requires the ability to outsmart machine as well as man.

The Chinese Cyber Reality

There are many ways in which using the Internet in China differs from the average users' experience in other parts of the free world. For one, the organization of the information and communications infrastructure and its history tell a story of government control through corporate exploitation. A second major difference is the existence of a government-promulgated broad list of topics barred from mention and activities which cannot be conducted. A third difference is the Great Firewall, commonly referred to as the GFW for short, which blocks local users from accessing non-Chinese sites.

One way to test this is to log onto the Internet after changing your browser settings to a Chinese proxy server. Doing this tricks other servers into believing your connection is originating from within China. You will have access to a number of sites which would normally be blocked to non-Chinese users for copyright reasons but on the other hand, typing in the web address for Google (even the dot cn Google China) will

result in an error message. Other Western sites such as Twitter and facebook as well as some news sites will come up with a 503 error (server not available), 404, or some other error. This is what the Internet is like for PRC users. Even if you know the web address, access is blocked. It is not that the search engine (the most popular is Baidu) will not return the official site if you search for it (it does "find" the sites many times even if they are banned), nor is it merely the case that you will not be able to access the site through the search. Access is blocked even if the site address is input directly into the address bar. This is significant because that method is how one would access a site that has been created but not published (the so-called "dark" web). The sites and their domains are blacklisted and cannot be accessed via normal connections.

According to Harwitt and Clark in their article "Government Policy and Control Over China's Internet" development and control of both the Chinese Internet infrastructure as well as content has been the purview of government agencies or corporations whose leadership is staffed with former military and government officials. What is not controlled directly, is suppressed through painful regulation meant to put Internet Content Providers on the hook for individual user behavior.

> The goal of these measures seems to be to intimidate users into censoring their own web content. As it is technically impossible for the Chinese government to screen all domestic web sites at all times, the tactic of 'killing the chicken to scare the monkeys' (publicizing punishment to intimidate the masses) is one of the few tools the authorities can use to prevent [Internet Content Providers] from crossing politically acceptable boundaries. (25)

The article reveals a system where unofficial policy has historically been set by personnel, and providers have no choice but to collude with government or go out of business.

In addition to these controls, there are new laws which also place greater burdens on the users themselves. A 2013 law for social media places a limit on the popularity of some *Weibo* posts, threatening prosecution for defamation, fines, and up to three years jail time if a post is re-posted (retweeted) over five hundred times with inaccurate information (Kaiman). Even in the early years of the Chinese Internet, the government

has not hesitated to prosecute individual users who have transgressed content regulations. One of the earliest examples was Huang Qi from Sichuan whose website on the atrocities of Tiananmen earned him a fine and five years in prison for subverting state power (Harwitt and Clark 25).

Today's social media users in China are merely learning to live by the same rules that Chinese reporters have had to abide by before the Internet. The Center for International Media Assistance (CIMA) reports the Chinese government is growing increasingly belligerent not only with her own reporters but with foreign news outlets and reporters, resorting to tactics from physical violence to denying visas to more subtle pressures like co-opting editors into self-censoring. As a result, fewer local and international outlets will even touch topics "that one former Chinese diplomat said were internally called 'the five poisonous groups.' There are Tibetans, Uighurs, practitioners of the Falun Gong spiritual group, Chinese democracy activists, and proponents of Taiwanese independence" (Cook 11). Whereas once the Chinese press fed highly distilled to limited information regarding the Five Poisonous Groups, now Internet users are faced with having their own blogs, comments, or social media posts distilled, or "harmonized", by an unseen hand using unknown tools executing an opaque mission of thought control.

The tools and methods being employed by the Chinese government are only slowly coming to light. One of the best articles covering the operational aspects of censorship is the King, Pan, and Roberts article "How Censorship in China Allows Government Criticism" by analyzing "volume bursts," or surges in activity focused on a particular topic or keyword, they concluded that it is not necessarily the topic matter itself which leads censors to act, but the "horizontal" nature of high volume activity for paternalistic reasons.

> [C]ensorship is primarily aimed at restricting the spread of information that may lead to collection action, regardless of whether or not the expression is in direct opposition to the state … [suggesting] the Chinese regime believes suppressing social media posts with collective action potential, rather than suppression of criticism, is crucial to maintaining power. (328)

Perhaps Winston, the protagonist in George Orwell's *1984* would have felt less suppressed had Big Brother allowed him as much personal space to express negative thoughts about the government where he could mutter such caustic comments to himself. Orwell's world was less forgiving if no less omnipresent, technologically speaking. Big Brother bugged every home and brutally policed every real or imagined criticism. King, Pan, and Roberts present a PRC strategy that is somewhat counter-intuitive but more practical. The idea of prioritizing any speech trait over criticism when trying to control media does not fit our traditional understanding of thought control. This is why many users may not anticipate the approach.

User belief regarding what the government does and how it does it will shape their attempts to avoid detection or censorship, and as much as beliefs differ from user to user, so too will their tactics. We cannot possibly anticipate all conceivable permutations of their understanding of Internet censorship, but it helps to know about the infrastructure, legal, and information frameworks at work in the Chinese Internet.

RESEARCH

The collection phase of this research was conducted as an abbreviated cyber-ethnography. Cyber-ethnography, or netnography as it is also known, applies some ethnographic methods and tools to a "virtual" or cyber community. Ward cautions against the "artificially polarized" pitfalls where the researcher inappropriately maps physical constructs to the virtual world or idealizes the virtual as an electronic utopia that cannot be subjected to the limitations of the physical world. Ward recognizes cyber-based social sciences must recognize the dual nature of virtual communities.

> When traditional ethnographic research methods are applied to study virtual aggregations, they often have the effect to reinforce the dichotomous relationship that has emerged between the physical and the virtual. That is the virtual community becomes depicted as an entity that exists entirely within the virtual realm, as if in isolation from the physical; having played no part in the construction of our hybrid culture …. The virtual community can never fully escape the confines of the physical. (96)

Ward also cautions against the researcher playing a role in defining the online community being researched because it denies the more fluid, less rule-bound nature of virtual communities and how they organically self-identify. In the case of this research, however, the community in question is identified as PRC *Weibo* users who access and interact with the other users through a service infrastructure located within the PRC's boundaries meaning user and service provider alike are subject to the same laws, regulations, and requirements. This is a suitable working definition for community in this study as I am primarily interested in the large virtual community defined by a characteristic nationality.

It is important to note I was not physically present in China at any time during the course of the research. Physical presence is not a necessary requirement if the community in question is virtual and accessible from the outside. By definition, the cyber-ethnographer cannot be physically present in a non-corporeal community, but he can exploit those characteristics of the physical network to "visit," virtually speaking, those same communities.

Aside from this necessity, one of the great advantages of choosing to examine the social media environment is the fact that it is entirely public. Regardless the probable preconceptions of the user concerning possible government reaction to their own speech — the laws of the land, user guidelines, as well as the *jubao* (举报) "report this" option that exists on every *Weibo* post next to the "like" and "re-post" functions all point to a reasonable expectation that users are making their pronouncements public with informed consent.[2] When a user submits a post which could be socially or politically sensitive, it is likely the activity is motivated by a desire to communicate (vertically or horizontally).

While social media sites are properly understood as a public space the general rules of interaction depart from traditional social gatherings in physical space or telephonic conversations. They are public due to the freedom of users to explore the networked entities, but the written interaction and self-created network of followers encourage a more intimate and personally resonant approach. An example of this difference would be to compare how a controversial statement may be received in the two different contexts. Imagine two employees of the same company meeting over drinks where one employee expresses criticism of the company president saying, "Bilk Buckmore is taking the company in the wrong direction." In that intimate environment, the employees might feel safe to engage in criticism, deflection, or apologetics regarding their boss.

If the same employee posted the same words on his Twitter feed, however, the repercussions are much more likely to be dramatic. A known employee has gone "on record" in the public sphere, expressing his lack of confidence in the company president. Depending on the employee's reputation, stocks may fall and dissatisfaction could be

---

[2] The "report content" feature is a fixture in US social media sites as well where its use (and abuse) is related more to maintaining a semblance of civility (and possibly sterility) in discourse than as an official channel for informing on others. Many official websites, such as Sina.com, contain reporting links to various municipal or national government authorities at the bottom of their main pages. Exactly who is informed of what when a *Weibo* user reports on another's post is unknown. As stated above, the line between government and commercial enterprise is blurred and the responsibility for content places a good portion of the onus on the providers themselves. In such a cultural environment, it is safe to assume not all reported posts are handled at the lowest level of responsibility.

sowed among the rank and file. Ultimately, the offending company man could be reprimanded or fired in response to his indiscretion.

The same words have remarkably different effects based solely on the forum or medium where they are expressed. In the United States, users, especially public figures, can face intense backlash for what they write online. This happens when the user's personal sense of propriety fails to jibe with what is considered socially acceptable, or responsible, speech within the forum of social media.

As a cyber-ethnographer of social media posts I need to take into account both the public nature of the posts as well as the potential for volatility which may be associated with their disclosure. Even though the space is public, the nature of personal interactions on social media still call for some degree of discretion on the part of the researcher. It would be unconscionable to do the censors' job for them or offer any sort of assistance intentionally or otherwise in singling out those who transgress Chinese speech laws. The goal of this work is to explore methods for cataloging socially or politically objectionable speech collected from Chinese social media, but the collection process should be conducted in a manner that is respectful of individual rights. It is therefore recommended that this part of the process be compartmentalized to prevent leakage of personally identifiable information such as the user names. While these specific details need to be retained in the primary collection process, they do not have to be made available to all researchers and should generally be kept from any online or other widely-accessible manifestation of data analysis or cataloging. Access to identifiable information should be granted on a case by case basis depending on the credentials and information needs of the researcher requesting access. Even though the processes of observation and collection are occurring online, the stored items will not be databased in a web-accessible application.

This is not to say that all identifiable information is concealed, however. As previously stated the expectation of privacy on social media, especially horizontally potent forums such as *Weibo*, is nonexistent and not all speech constitutes socially or politically scandalous fare, though it may be safe to say the vast majority of it is not. As a case in point, one of the perennial trending topics on *Weibo* is #晚安# or "good night." Collection of such mundane content where there is no identifiable cause for concern need not be carried out with the same required protections for user identity.

Similarly, posts which have reached a widespread or "viral" trending status such that a large number of users have already liked, reported, re-posted, or responded to them also far exceed any reasonable expectation for privacy and may likewise be considered general public knowledge. The proper threshold for making this determination, whether it is one hundred "likes" or one thousand, is a matter for debate. The practice of not revealing identifiable information can be seen as equal parts paranoia and polite consideration.

Finally, posts which are collected outside of China but sent by Chinese dissident and activist expats as users of the US Twitter or other non-Chinese social media may also be cataloged with open identities because these users are more informed, not subject to Chinese law, and want their posts to be widely circulated in order to further foment their cause.

Overall, this desire to protect the identity of users who engage in sensitive speech is a suggested additional bulwark so the project and its findings do not become inadvertent tools of the very regimes it seeks to study. It is an extension of the desire to err on the side of an abundance of caution even if the user in question might disagree with the need for such caution.

This collection research was conducted in two phases due to the technical and cyber-social learning curve which required careful evaluation at each step in order to minimize external threats such as being blocked. The first phase was conducted from October of 2013 to the present on Twitter from the United States and the second collection phase started on the 22nd of January 2014 and is also ongoing. Each phase had distinct challenges and rewards as well as several points in common. Here I will present a brief overview of the important points pertaining to both phases.

In terms of overall approach, my goal was to remain as hands off as possible. In Twitter, this meant minimal engagement. If I saw something politically engaging, I "liked" the post and/or re-tweeted it sometimes doing my best to provide a short English translation in the process. Some of the users I followed also followed back, but not all and one in particular would not allow anyone s/he did not trust to see their postings to the exclusion of myself.

On *Weibo* I was even more cautious and used my access only to "lurk." That is, I looked around but never posted, re-posted, or liked. My motivation was not just to play impartial observer as a researcher, but to keep my user profile as free as possible of self-incriminating reasons which could lead to being banned or blocked.

One of the great challenges throughout the process was dealing with the language barrier. It is one thing to use the Chinese language in professional or academic contexts and quite another to be confronted with more obscure references and colloquial expressions used in social media. It will not surprise those familiar with the effects of social media to discover the language used on *Weibo* was at times coarse and riddled with obscenities and insults. Nonetheless the rampant employment of such vulgarity came as a bit of a shock. Additionally, I believe the Chinese language to have a decided advantage over English in the context of short-comment social media like *Weibo*. One hundred and forty characters in English is barely enough room to form a cogent thought while the same limitation in Chinese allows for much more information in fewer characters. My evidence is anecdotal stemming from the fact the English translations of Chinese posts regularly exceeded the character limit. Chinese does not require as many articles as English and most words are only two characters with the exception of most past tense verbs and four character sayings which unpack to entire idioms in English. It is perhaps admitting too much as a researcher, but the process of reading Chinese tweets quickly led to mental exhaustion, especially for longer tweets or those requiring background research in order to fully understand the context. When the process of translation resulted in a confirmed politically sensitive post, it was very rewarding, but many posts were more mundane.

Throughout the course of the research I encountered three general types of social media users who utilized the Chinese language: the expat abroad, Chinese users in China, and other Chinese speakers in China or elsewhere.

The most common user type on Twitter was the Chinese expat. These users were of a more political bent and tended to be dissidents living abroad in cities such as New York, Paris, or Austin. These users were the most critical of the regime and served an essential function of passing news from *Weibo* users to the outside world.

The most common user type on *Weibo* was, not surprisingly, the typical PRC netizen. These were the most likely to experience pressure from the Chinese government and also the least likely to engage in behavior that would warrant suspicion.

The third group would be Taiwanese and other Chinese-speakers such as Singaporeans, Malay, and westerners. These users lived both inside and outside the Chinese Internet. Some were reporters or academic specialists, some lived as teachers or business professionals in China. Many of the non-PRC citizens could be easily identified by their preference for using the traditional character set while those living in China used simplified characters exclusively. Politically this group tended to maintain a practiced impartial neutrality.

Phase One: Twitter

Twitter provided a few unexpected surprises as an unofficial conduit into the world of the Chinese *Weibo* and supplied a number of critical posts, including many pictures. In this preliminary phase my primary goal was to get in touch with persons known for their critical stances against the Chinese government as a means of establishing contacts to serve as unofficial and uncensored information sources to compliment official news sources such as the China News Agency *Xinhuashe,* sina.com, and dissident news outlets like NTDTV. The information presented by these users was not vetted by editors or censors nor was it translated. This meant the information was more timely and "raw" in terms of distance between the user reporting and the event or circumstance in question. This unofficial news feed became powerful tools for contextualizing content found on the *Weibo* side.

I was thankful to find a rather large community of Chinese users who were, as stated above, comprised largely of expats, refugees, and political dissidents. Among those who I found and followed were Zhou Fengsuo, one of the student leaders during the Tiananmen democracy movement and Chen Guangcheng, the blind lawyer who was outspoken in his criticism of the one-child policy and, thanks to the efforts of Secretary of State Hillary Clinton, was able to escape jail and come to the United States. Both of these users were well-established Twitter entities with extensive networks of friends and followers. From there, I made several other connections and viewed their prodigious tweets as an alternative type of news feed for unreported or underreported events.

I did not expect to find cryptic references in their messages mainly because, operating in an American social media environment, there was no need for the Chinese users to temper their criticism of China. Yet there were many images with arresting content which were worthy of attention from mainland China censors, many images contained a watermark indicating their source of origin as *Weibo*. One image in particular depicted a group of men wearing Guy Fawkes masks and holding signs to protest police maltreatment of the daughter of activist Zhang Lin (see Figure 1.) The Guy Fawkes mask sends an unambiguous antigovernment message. Traditionally, the mask has been used to mock a failed Catholic terrorist attempt to overthrow the British monarchy. Catholicism has always been illegal under the Communist regime in China. In modern visual orthography, the mask represents various anti-government or anti-establishment movements. No matter which ideology you believe the masks reference, contextually speaking, the message is clearly anti-government. In addition to this, the signs being held by the men offer a direct protest mentioning Lin's daughter Anni: "National Security disgraces the nation if Anni is unsafe." Like all pictures that originated on *Weibo*, this one too bears the site's watermark and unique identifier, even though I first experienced it through my Twitter feed. This is evidence Chinese Twitter users pass politically sensitive *Weibo* posts around the Great Firewall to the "other side" of the Internet.

Figure 1. Upper half: activist Lin and his daughter Anni. The lower half shows four men in Guy Fawkes masks holding signs that read "National Security disgraces the nation if Anni is unsafe." Note the Weibo logo and image id watermark in the lower right corner indicating image origin.

Even though there was no need to engage in censor-dodging activities on Twitter, it was clear from these and other posts that Chinese users were using Twitter as both an alternative method of communication as well as a digital holding pen for threatened speech. In addition to the above picture were many others including an image of angry rioters protesting government incompetence in relief efforts after a storm and a snapshot of a racist and threatening bumper sticker on a cop car in Tibet (see Figures 2 and 3.)

Figure 2. Scene of unrest at 余姚 *Yuyao.*

Figure 3. Bumper sticker "Don't mess with me or I'll let you die without missing a beat." The license plate indicates the vehicle in question is a police truck in Tibet.

Though expectations for collection were low in the early days of the primary phase, the social media website has not only proven itself as a reliable source of unfiltered information on current and ongoing events, I discovered its use by dissidents as a source for a great number of images as well as postings directly from *Weibo*. Though I spent almost no time using Twitter during the second phase, I believe any future collection endeavors need to include Twitter as part of the collection plan.

Phase Two: *Weibo*

Aside from the linguistic challenges already mentioned, there were a number of factors threatening to derail my goal of accessing *Weibo* for this phase of research. In the end the threat of having to overcome these challenges influenced the aggressiveness of my research. I feared having to start over should my activity be tagged as suspicious. Being banned or blocked could also have resulted in the sanctioning of my IP address meaning that even if I could connect to the Chinese Internet again, I could possibly be blocked from creating a new *Weibo* account as well. Once on the "other side" of the Great Firewall, I wanted to tread as lightly as possible.

This research began with a basic working knowledge of information systems and several working concepts on the policing of unapproved speech. Unlike the King study, this research was not conducted with guidance from in-country contacts or even specialized equipment or software. One of the concepts guiding this research was to explore if the process could be completed without a great deal of investment and technological support. Thus far, the required hardware and software meant one laptop, an Internet connection, and a browser (in this case, Firefox). It is hoped others who have interest in monitoring or collecting sensitive speech will likewise be encouraged to follow suit.

Here I will give an account of the issues with creating an account on *Weibo*. One of the first and most lasting issues was slow load times, especially when trying to log into the service. Even after account creation, this was a process that seemed to work only some of the time. There were many instances when the login would cycle as if unable to continue. After prolonged periods of up to five minutes, I would stop the page from loading and reload again. Doing this either resulted in access to my *Weibo* home page or

took me back to the login screen. At no time in the process did I allow the site to keep my account logged in or to save the password information in my browser. It remains to seen if using these options would speed up the process.

In general, the Firefox add-ons I used for security and anonymity purposes became great impediments when trying to log into *Weibo*. One of my favorite tools is a script-blocker which allows only user-approved scripts to run while keeping a log of all active or inactive scripts associated with the site. Throughout my attempts to create a username and password and beyond, the script blocker proved to be a deal-breaker. None of the essential pages would load as long as it was engaged. I believe the site either did not allow the script blocker by design or there was some other unintended consequence associated with its use that prevented proper account creation.

Once the blocker was switched off, however, it was easier to move from page to page. One of the functions blocked by the script blocker was the Captcha verification. While the script blocker was engaged, the browser would not display the images containing distorted lettering. This verification device was commonly employed and was required for each login. This is unusual. Even on sites which utilize the Captcha verification, the feature is not required every time a user logs into his or her account. In theory, it may be a matter of customer service to utilize a verification method that requires users to establish their humanity when logging onto a social media account. It may also be a matter of expedience in a legal environment where the both the service providers and the users can be made to account for the content of their posts.

The fact that the Captcha is always required when manually logging into *Weibo* then may be attributed to a number of reasons which are not necessarily mutually exclusive: 1) *Weibo* experiences a high incidence of user accounts being surreptitiously co-opted and this process involves bots. 2) *Weibo* has elected to use the Captcha because users themselves complained of their accounts being accessed by others. 3) Captcha is commonly used in many services in the Chinese Internet either because it is culturally part of common practice or it is required in some respect and *Weibo* therefore is not an outlier. 4) There was something about my own browser settings and IP address which triggered the Captcha to appear every time I logged into the site.

The confusion that arises from this simple fact illustrates one of the more difficult aspects of cyber-ethnography. Although I frequently experienced annoyance at having to fill in the Captcha it was not until I reviewed my notes that I realized there was something about its employment that was unusual. It is easy in the moment to dismiss such instances as a glitch or somehow arising out of user error, but in the quiet of careful analysis certain details begin to emerge. The Captcha is not the only element which I dismissed in the moment but found curious in retrospect. In the ever-changing landscape of social media, it is very important to keep good records along with the original text, web archives, and screen grabs to facilitate revisiting should the researcher hit on some interesting fact.

Another obstacle to account creation were the requirements for an email account and a telephone number. The email needed to be a local Chinese email address and all the services I came to required at least a local Chinese phone number to associate with the new account and often also asked for a local mailing address as well. It was also possible to create a *Weibo* account by using a smartphone but that ended in the same in-country address conundrum as the email account. Having neither a local Chinese phone number or mailing address, I was effectively blocked from creating a Chinese user account for *Weibo*.

At this point I contemplated creating a mailing address based on online maps, but dismissed the idea as unethical. I also contemplated using an in-country contact to create the account for me, but for similar reasons did not want to put someone else in the position of responsibility for my actions and online activities.

After spending days failing to get an email account through various service providers I eventually found one with verification requirements lax enough to allow for the account creation. Once that was established I went back to *Weibo* and created my user account.

By this time, I had spent a great deal of time trying to find a way around the email/phone/address chain of requirements and became concerned what would happen if the process had to be repeated with a different service provider. Moreover, I worried any activity leading to either my *Weibo* or email accounts being closed or forbidden might also associate my terminal's IP address. If that was the case, not only would I be back at

square one in terms of account creation, but I might also have to deal with more daunting issues requiring greater technological assistance in the form of specialized software, hardware, and other network-specific workarounds. This fear greatly colored how I conducted the research. As discussed above, I tried to keep my searches to more innocuous terms but this I did only after clicking around a few of the mainstream trending topics from the righthand side of the screen. Once I felt my user history had built up enough instances of typical use, I started searching for more controversial terms such as the names of known activists in the news, "police," "good police," and "censors." With the exception perhaps of the police-related terms, these searches turned up insignificant results. The police-related terms turned up posts that were largely laudatory or politically neutral. Within these results, however, there were a number of images which were posted as visual proof of abuse or mistreatment at the hands of authorities.

Based on these searches, I was able to draw a number of conclusions chief of which is the difficulty in searching for new homophones. For example, during the disappearance of Malaysian flight 370, I tried a number of alternative characters based on the possible pronunciations of 370, "*sanqiling"* or "*sangoudong,*" but these searches did not produce any results from the period of the plane disappearance. The topic of the plane was trending on *Weibo* at the time, but I hoped to find evidence of cryptic circumlocutions in case of occasional blackout due to probable critical activity spikes. More information would have been necessary to see if the topic (or related topic) trended to the degree that censors would have been obliged to filter posts. My theory is cryptic references are not resorted to until users discover that certain posts end up blocked. This would have been easy to test for if I wanted to do more than lurk and send out some test posts with problematic content. I was still gun shy and feared losing my account, so the only test available was to search *Weibo* posts for clues. This technique ran into a wall of statistical improbability.

The other possibility is my linguistic skills are too poor to develop effective homophones. As a non-native speaker I lack a well-developed sense for word play such as puns making it difficult to create a sufficiently clever in-joke like many of the known homophones. Another possibility is I was focusing on the wrong words. Perhaps the homophones exist, but were a play on any number of related terms, phrases, or numbers.

At this point, the work felt more like archaeology than ethnography. Like a treasure hunter without a map, I dug around the site, trying various permutations without any sense whether or not they would yield results. Certainly one way to curtail this frustration would be to scan for activity spikes followed by telltale lulls indicating censorship activity. This would better focus search efforts where the users would be most likely to act out against speech control and employ improvised countermeasures.

Although it was difficult to find new or newly coined homophones, I discovered the established ones are still in use. There were numerous examples of 河蟹"river crab" being employed as a reference to censorship. One user lamented being "river crabbed" while another used the term in response to reported new guidelines threatening to cut down on online (mainly Western) content distribution. Although the King article listed censors and censorship as the most commonly curtailed expression, both cryptic and direct references to censors and censorship seemed to exist in a fair number with more critical posts demonstrating a preference for the cryptic term (a fact speaking to how such circumlocutions are used). Yet this finding still jibes with the King article's determination that mere use of a sensitive term does not in itself warrant censorship unless that usage is spiking in reaction to some real world event.

Images were plentiful however and there were many which cast the government, police, or officials in a negative light. It is debatable whether or not such posts constitute attempts at cryptic communication, though I believe they should be included. The argument can be made that images and their contents are also a type of reverse Turing test for the censors and can only be understood by human scanning, putting them out reach of the crawlers, thereby rendering them labor-intensive and time-consuming for censors to process.[3] It is also plausible images are used not so much to avoid censors as to deflect user culpability. Unless it has been somehow altered, it is difficult to prove an image was posted with the intent of spreading rumors or falsehoods. The new laws penalizing users for creating popular posts which contain untrue remarks are in theory not

[3] The Turing test was developed by Alan Mathison Turing, a mathematician and inventor of artificial intelligence, as a series of questions used to determine the degree to which a computer's AI resembles human intelligence.The more human-like the answer, the higher the score. Here, the Captcha is being used to weed out non-human bots by offering a task beyond the capacity of an Internet bot program.

as easily applied to images, provided care has been taken to state only factual details in the text accompanying the post. The fact that many of these images are posted by lawyers may support this view.

Images are powerfully persuasive tools as well and are used by propagandist and activist alike, but images also have a certain degree of immediacy which make them less tempting for censors because they deal with particulars, meaning those persons within the image itself or those relating to the person in the image. A single photo can say much but does not always provide a sense of problems in general or convey a sweeping statement about the power of the state or the prevalence of abuses and corruption on a national scale. Images thus tend to be about the proximate and feature local events and local people. All of these are reasons why someone who fears censorship might resort to images, yet at the same time it must be admitted that avoiding censorship may not necessarily be one of the motivating factors.

It is quite hard to ascertain in general if these techniques are utilized to avoid censorship or if there is some other motivating factor. It is not always easy to determine what motivates every user, especially if what we hope to find is a deliberate form of subterfuge. Any search on *Weibo* reveals a wide variety of opinions and attitudes as well as varying degrees of awareness. In one post a picture meant to show police brutality elicited an unsympathetic response accusing the original poster of going to lengths to make the imprisoned man look worse off than he was, demonstrating a degree of psychological resistance to being manipulated by online images (see Figure 4).

Figure 4. Images taken of client as proof of physical abuse at the hands of police posted by his lawyer online. Listed abuses included being chained, beaten, and not allowed sleep or go to the bathroom, resulting in a severe hernia. The poster was most moved when the prisoner recited the full name and badge number of the cop whose sole duty it was to rouse him and prevent him from sleeping. Though the images are disturbing, reactions included some cynical responses.

In another thread where users were whipping themselves into an anti-Taiwanese frenzy, a user boasted about PRC users heaping abuse on Taiwanese (who she described as extremely stupid) by logging onto facebook Taiwan and taunting them. The post included a rather large screen grab of an exchange between numerous users from both sides of the straits on Taiwanese facebook (see Figure 5). The *Weibo* community of users expressed approbation by liking and reposting the original message over one thousand times in the span of less than a day. The post's appeal was likely due to its nationalistic enthusiasm but it ironically boasted circumventing its own country's laws as well as the cyber gatekeeper, the Great Firewall which is supposed to prevent Chinese users from accessing facebook.

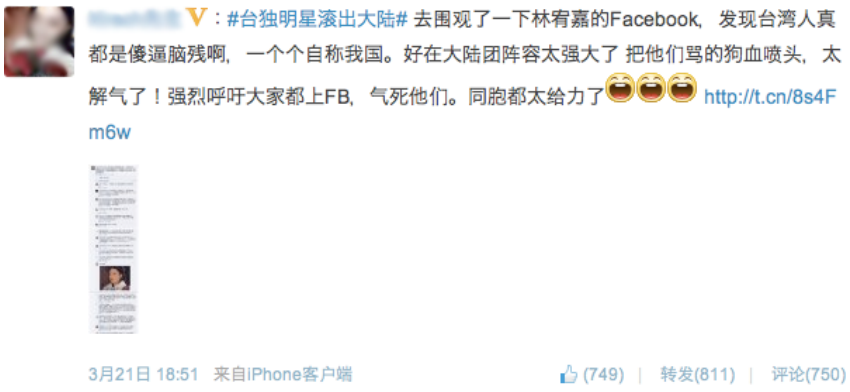

Figure 5. A post expressing pro-establishment sentiment regarding the sensitive topic of Taiwanese independence. Imbedded in the post is a screen grab from Taiwanese facebook.

While we cannot always know why someone choses to use an image or a homophone or straight talk over any other manner of communicating online, that does not mean the researcher has no access to the meaning of the post itself. Even when the method of communication is self-contradictory as in the previous example, it is generally clear what they intended to say. The lawyer wanted to elicit sympathy and perhaps outrage over the treatment of his client. The Taiwan-basher sought ridicule and condescension. This much is obvious in their choice of words and images. The ultimate goal was communication, even when tactics to throw web crawlers off the scent are deliberately employed. Communication is the common end goal that unites and defines all online social network activity. We may even employ a term from philosophy by saying horizontal communication is the teleological purpose of this activity regardless of whether or not secondary or tertiary goals and motives played a part in the overall form.

This ambiguity represents a central challenge to the validity of scanning social media for cryptic references. Even if a user choses to be coy with their wording (or images) it can be difficult to say with certainty *why* they choose the message form unless they are asked directly or a broad survey of attitudes backs up our assumptions. Thus, for the time being, extrapolating intent from format requires more hypothesizing than deducing. Obviously, not all social media communication is likely to contain cryptic references. In terms of observable message content and form, there are indicators which can help researchers narrow their focus and eventually improve the quality of the overall

research. The following sections address likely forms and traits and propose an underlying philosophy for an indexing schema.

ANALYSIS OF KNOWN TYPES

If our goal is the collection of endangered speech with an eye towards seeing how users attempt to prevent, delay, or circumvent perceived normal modes of message detection and elimination, then we must also create schema for the organization of our collection. Before a schema for organizing scrolls or books was created , there had to be scrolls or books to organize. Before a database can be created for personnel, there must at least be some personnel to organize. We turn now to examining some examples of cryptic Internet communications in order to examine their characteristics.

Homophones

Perhaps the most popular example of cryptic Internet communication in Chinese cyberspace are those employing characters with innocuous meaning to imply something altogether different based on the simple fact that in written Chinese, there is not always a one to one ratio between a phoneme (the "sound") and a given character. For example, there are 67 simplified characters which can be pronounced *shi* (pronounced like the English word "sure"); 13 in the first tone, 11 in the second tone, seven in the third tone, 34 in the fourth tone, and two in fifth, or neutral, tone. While *shi* is the most common sound in Chinese, there are a number of characters which are unique to their phonemes such as *fo* (佛) and *ri* (日). In general, however, the majority of phonemes can correspond to more than one character, especially if you are willing to ignore tonality.

As a result, it is possible to substitute one character for another. When a crawler reads a character, it primarily recognizes the widely known uses and meanings. Unless it has been directed to tag for example, "river crab" (河蟹) as a possible substitute word for the more volatile "harmony" (和谐) then it will not tag the text as relating to any of the banned or sensitive topics.

For the uninitiated, the stigma attached to the Chinese word for "harmony" can be confusing. The reason for its ignoble place in the eyes of government thought control is itself a story which outlines the evolution of a cryptic circumlocution.

The word originally earned a bad reputation, not among censors but among Chinese users when it became part of the standard phrasing explaining why a site had been taken down on the basis of questionable content. Visitors to such sites would often see a message indicating the site was not available for reasons relating to social harmony.

Before long, the passive construction “be harmonized” had become part of a larger linguistic phenomenon where the passive construction of a verb became shorthand for the specious reasoning employed by officials or police.

People began talking about cases of dubious suicide with the passive “was suicided” to indicate the possibility the official explanation was too convenient. In like manner “was harmonized” quickly became synonymous with being censored. “Harmony” became the code word for censorship and an easy way to reference a heavy-handed tactic as well as the contradictory logic behind its employment. Eventually censors caught on, and began to cull the reference. The word became so popular even its homophone was eventually marked by censors as worthy of deletion. Leading to the point where many see the online mention of “river crabs” as a sidelong reference to “harmony” which is itself an official euphemism.

Recall at this point the revelation of the King study which demonstrates Chinese Internet communication analysis does not merely search for sensitive terms, but also flags postings based on repeated horizontal action or activity bursts. My thesis assumes that certain words or phrases are flagged for monitoring by crawlers. I believe even unremarkable posts earn some degree of scrutiny before judgments are made to censor or block them. I imagine a system where the post is given a degree of criticality specifying a place along a range like a temperature between hot and cold. This is purely hypothetical but like the King study which jibes with Chinese defamation law, this hypothesis is supported by Chinese legal attitudes regarding restricted speech.

Along these same assumptions, I also assume most Chinese users will avoid openly using words relating to sensitive topics either out of an instinctual desire to keep a low profile or because they have a technical understanding of how tools like crawlers work. A more detailed understanding of how the average netizen acts and what he or she knows or believes to be happening “behind the scenes” is an important aspect of the overall picture and a critical area for future research.

“Harmony” is not the only watchword that has gone through lexical or logical evolutions caused by government censors. Another well-known homophone is the alpaca or “grass mud horse” (草泥马) — a word which bears startling phonetic similarity to a very unsavory action towards one’s mother. At one time, the term meant nothing at all. (It

is not even the actual word for alpaca in Chinese.) Over time, however, censors caught on to the joke and started deleting references because of their obscene nature. Once again, anti-censorship sentiment resurfaced in the form of online postings of stuffed alpacas. These images were posted, not because there was a growing contingent of users who wanted to offend mothers, but because users believed the government's zeal to quash vulgarity had been too heavy-handed.

An advantage of the homophone may lie in a certain perceived level of plausible deniability. Indeed, it may even be inferred by the backlash created when censors choose to target words like "river crab" or "alpaca" that users are upset over the sanctioning of otherwise perfectly legal words and want to fight back over the watchword list's encroachment on even mundane language. The common imagination is seized with panic at the thought of an increasing number of words subject to scrutiny. How far could it possibly go before every word is suspect? This bristling against censorship activity by popularizing cryptic speech is as much a plea to allow mundane speech as it is a cry for greater freedom to express controversial ideas. People should be allowed to speak about river crabs if they so choose. It may be that those who are caught using homophones will argue their case in a way which also meant to shame a government that chooses to so intensely monitor the words of her citizens. They may feel as if they have recourse to feign ignorance if caught posting questionable content.

A disadvantage to using homophones is likewise the threat of implicating harmless little words (or other users who believe they are talking about something harmless) into a wider counter-censorship conspiracy. The threat is not that the words may not be understood and dismissed, but that they may be misunderstood and quoted, replied to, re-tweeted, or picked up into a larger conversation by earnest yet oblivious users.

Altergraphs

Perhaps related to homophones is the tactic of creatively employing characters to convey a message visually instead of phonetically. "Atlergraphs" is a suggested term which can be used to classify visual variants suggestive of glyphs in all languages. One example of altergraphic English from the early days of the Internet is the "leet speak"

employed by hackers on forums.[4] In Chinese, altergraphs use the radicals in the characters to convey a message. And example from the blog ifgogo.com is the rendering of "software" from the simplified 软件 to a series of four characters by exploding the radicals to 车欠人牛.[5] Careful examination will reveal the same elements are in both the two character and the four character rendering, and because the four character rendering is meaningless (literally "car owes man and cow"), the reader is cued to decipher the true intent.

The advantage of this method lies in its subtlety as both the meaning and the pronunciation of the word may be rendered undetectable by crawlers once altered. The disadvantage lies in the fact that the simplified character set does not employ as many radicals, nor are the radicals as detailed. Because of this, there are fewer characters overall that one may use to camouflage their message. Additionally, this type of message requires a certain level of familiarity with characters on the part of the reader by turning reading from a passive linguistic skill into a quasi-active one. There is also the possibility the altergraphs will be dismissed as typos or other errors.

Acrostics

Written Chinese was originally read vertically and while the practice is still evident today the vast majority of Internet outlets use horizontally aligned characters. In January of 2013, however, there was a rash of incidents where Chinese news sites arranged their layouts to deliver a two-fold message (See Figure 6).[6] Read horizontally, these sites were delivering random news stories presented as a series of hyperlinked headlines, but when the first character of each was read vertically, users discovered the news pages had inserted hidden messages of encouragement to the Shanghai paper *Southern Weekly* which had become the center of controversy after one of its editorials

---

[4] A good article on the origin and development of leet speak can be found on Wikipedia at http://en.wikipedia.org/wiki/Leet

[5] The blog post contains several other tricks for altering characters to circumvent censors including "Martian" and vertical aligning much like acrostics. The full address is http://www.ifgogo.com/95/how-do-bloggers-avoid-censorship-in-china/

[6] Chin and Spegle wrote an article for the *Wall Street Journal* website on the development of events leading to the surreptitious protest by online news sources. The Chinese language *Epoch Times* online story contained screen captures of various examples.

led many to believe it was rewritten for propagandistic reasons. Once the trend was discovered, however, these other news outlets experienced government scrutiny on their own sites, resulting in reprisals for some of the editors.

Figure 6. Examples of acrostics hidden in unrelated headlines. Source: epochtimes.com

The advantage of the acrostic is it hides messages by reorganizing the sequence of characters. Normal crawlers would not recognize the stacked characters as being associated because the way they normally scan text.

The disadvantage is it would not take much to allow a crawler to "read" the text in a given page vertically or even diagonally, though it would add to the indexing process. Also the examples in the *Southern Weekly* case are all from establishment online news

outlets where acrostics are more noticeable because of their attention to layout and story organization, unlike the average blog or microblog where making such layout arrangements would be either tedious or impossible. This is the second disadvantage to acrostics — that they require a bit of website layout and design savvy to execute. Browsers can have their font settings changed, windows can be expanded to various widths, and screen resolution settings can also affect how text displays on a screen. The acrostic must be designed into a layout such that the first letter of each line is always the first letter, even if the acrostic were to be hidden elsewhere in the text. Such rigid layouts are more common on well-managed sites like news sites and some blogs. Acrostics are limited to the types of sites where they would be widely noticed by user and censor alike.

Cued Messages

One of the postings discovered over the course of this research was a post where the characters of two different sentences were interlaced like cards shuffled in a deck. It was obvious the message was not meant to be read straight through because it would not have made sense. The characters of one sentence were set inside parentheses while those belonging to the other sentence were not (see Figure 7). In this particular case, the usage is a contrivance meant to convey the hurried thought of feeling one thing about the police officer while her body was telling another: say hello to breakfast, again. Aside from the unsavory aspect of relating what it feels like to vomit inside a patrol car, the post content is quite benign. Yet its discovery made me realize there are more ways to arrange characters in a message than just through acrostics. Although to my knowledge there has not been a case where this method has been employed with the specific goal of circumventing censors, it nonetheless struck me as having the potential to do so through its rearranging of characters in a way that is not depending on site formatting like acrostics.

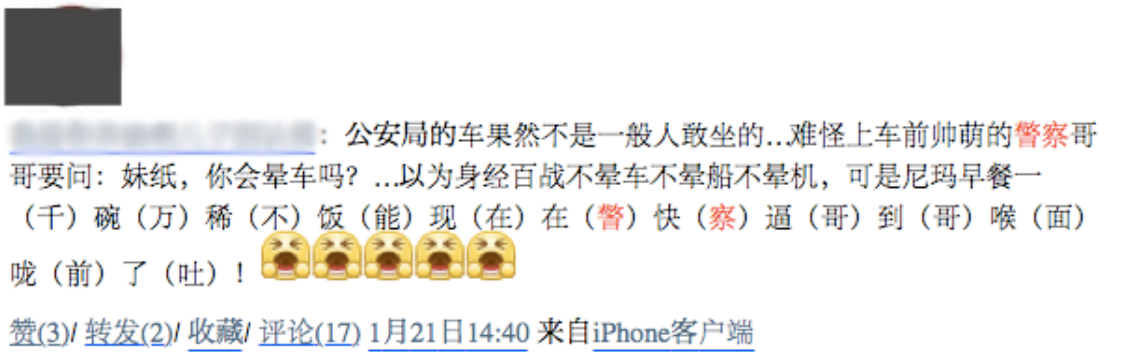

Figure 7. User relating the embarrassment of vomiting in the backseat of a police car driven by a cute young cop. The final two sentences, once de-interlaced, read: "...but that damn breakfast gruel is pushing its way to my throat/ by all means I can't hurl in front of that young cop!"

I suggest the term "Cued Message" to describe this technique because it requires a signal, be it subtle or overt, that the message cannot be properly understood unless some operation is first performed. This need not always be as simple as interleaving characters of two sentences and setting them apart with parentheses. A similar cue can be communicated by simply foregoing the parentheses altogether and trusting the reader would be clever enough not to try to read the mishmash of characters straight through. The cue could also be a given code word somewhere in the message that informs the reader the message is covert. There is endless room to apply one's own homegrown cryptographic techniques including using simple character jumbles or preset algorithms applied to the old standard telegraph codes or unicode. With proper preparation, pre-information age cipher techniques such as a one-time pad could also be used to send messages back and forth over public Internet forums with nearly unbreakable encryption.

This is likely one of the strong advantages of this technique — that it can be used with various degrees of "encryption" depending on whether the user wants to broadcast a message to the masses or quietly communicate with a set number of parties. It also provides a higher degree of protection from crawlers because watchwords can be obscured in a variety of ways.

The drawback of this technique is that it may require some degree of sophistication and coordination for more complicated versions and, the greater the complexity of the cipher, the smaller the group of people who will understand it. In general, cryptic circumlocutions are meant to take advantage of the horizontal breadth of

social media. If there are secret or "underground" organizations communicating through the Internet, there are less overt ways of sending encrypted messages and posting them on sites for the common user to stumble upon.

Images

That which is said to be worth a thousand words may also contain thousands more; it in fact also may reveal faces and other secrets. Image files containing text are more common on Chinese social media than in the United States owing largely to the fact that Chinese users embed the text of entire articles as graphic images in their posts, whereas English-speaking users prefer to insert links to the article's original address. This practice appears to be the norm in China as it employed by the very news outlets which generate the stories to begin with. As stated in the previous section, the why I consider this to be a circumvention technique is because the text of an article cannot be scanned as text, for it is a digital graphic where the text is merely imbedded visual information. In addition to images of text from an online resource, text may also appear in a graphic image in ways less recognizable to character recognition programs, such as signs held up by protesters in the example of Lin in the previous section. Other images may also use calligraphy to evade automatic detection or hide text somewhere in the picture, a tactic which could also serve to offer plausible deniability.

It is not only text in the graphic that can earn the attention of censors. As noted in the previous section on research, concerned citizens, lawyers, and other activists commonly post images to reveal abuse by officials or police or to document ongoing events such as natural disasters, riots, or protests.

The advantage to using pictures is the fact that images require a great degree of human discernment in order to be understood. Even with facial and character recognition programs, much can be conveyed in a simple image that cannot be automatically tagged by crawlers. Pictures can also serve as a type of evidential proof given the human tendency to believe what we see.

This capability to present evidence is also one of the greatest drawbacks of pictures. Once tagged, they can be analyzed for digital "fingerprints" which could lead back to the source. Faces of those either deliberately or inadvertently appearing in the

images could also be problematic and could lead to the implication of those not related to the events or the person who posted the image.

General Taxonomy

The two broadest categories of cryptic circumlocution are textual and graphic, where these terms define the media of the message. Each of these broad categories can contain either text or visual characteristics regardless of the media type (see Figure 8). For example, the Lin image or an image of the text of the news article could serve as an example of graphic medium (conveyed via a jpeg or other image format) with textual elements. Conversely, an altergraph may be employed which conveys an image. A very simple example of this might be a winking emoticon to alert the reader to an ulterior meaning.

In theory, an image could contain just about any form of text cryptic circumlocution such as homophones, altergraphs, and acrostics as well as few of its own such as text hidden in the background and calligraphy. On the other hand, the types of images which can be conveyed by text are much more limited. It would take a long time to use characters to create a more detailed image, like the famous scene with tanks from Tiananmen, and the overall impact would be questionable.

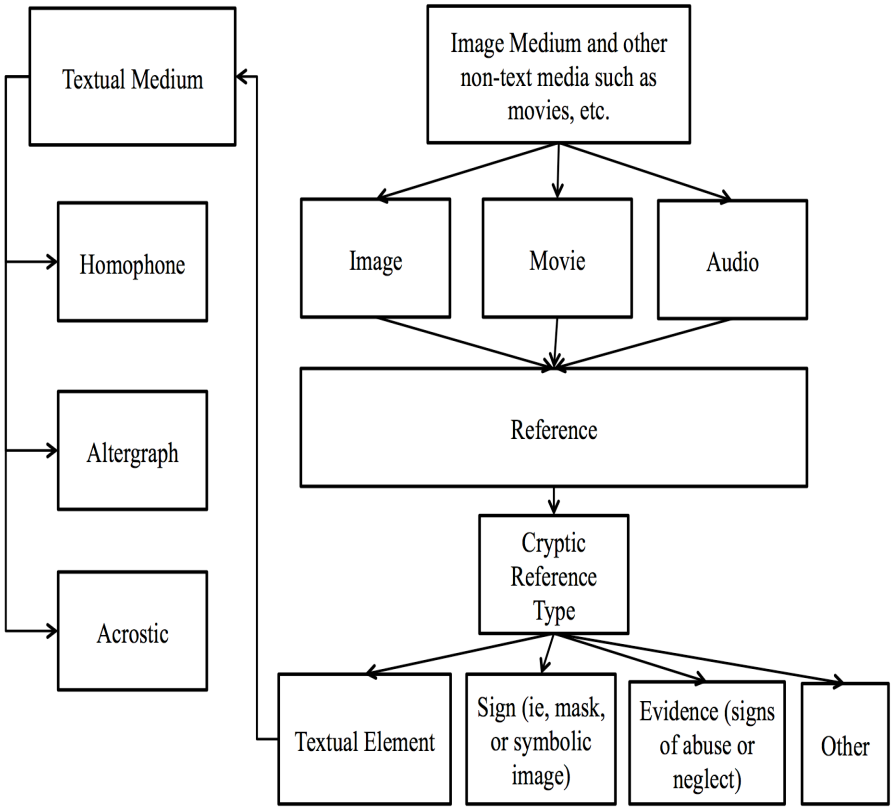

Figure 8. Suggested general taxonomy for cryptic circumlocutions.

BUILDING THE CATALOGING SCHEMA

In the research section, we examined methods for accessing the Chinese Internet as well as collection methods for archiving posts with probable controversial or problematic content. In the conceptual section, we analyzed those known, collected, and anticipated forms of expression as variant extensions of language in general and as the modal expression of metaphor in particular. In this section we will synthesize these concepts into a proposed cataloging schema. In order to begin, we need to review the requirements for this schema:

- It must allow for polyvalent interpretation of the content (gloss the relevant ways the posts can be understood with mind towards what is *meant*)
- It must allow for multiple ways of describing the storage of a single post (with the goal of preserving as much of the *information* as possible contained in all cataloged items.)
- It must facilitate a digital archive of digital objects (i.e. describe "the things themselves")

In addition to these conceptual requirements, the schema must also overcome a number of challenges:

- The multi-lingual nature of the collection
- The specialty of the topic matter which is both broad and obscure (Chinese government, national and local politics, social media, popular culture, etc.)
- Be broad enough to allow for future creativity and innovation.

In looking for examples, I researched a number of different metadata schemes and how they were developed. Each of the information products discussed in the articles dealt with different aspects relating to the creation of the censorship glossarchive. One of the main points of agreement among four of the five articles was an emphasis on the importance of the Dublin Core in the organizing of metadata and cataloging digital objects.[7]

---

[7] Section 1.2 of the home page for the Dublin Core Metadata Initiative, dublincore.org, describes itself as "small language for making a particular class of statements about resources." Designed by computer scientists, librarians, museum experts

The Given and Olsen article "Knowledge Organization" discusses the importance of proper organization of data in quantitative, qualitative, and textual methodologies for research. When creating a useful model for knowledge organization, Given and Olsen outline some issues to bear in mind when defining a controlled vocabulary.[8] A controlled vocabulary needs to address the specific research needs with a proper schema. Given and Olsen present the example of a data scheme where the researcher is interested in pets but the data hierarchy uses traditional zoological schema which obviously is not as useful as a customized hierarchy where animals are divided into wild or domesticated (162-163). This is the concept of "coextensiveness" and is defined by the area being covered as well as the number of categories it is to be divided into. Two other factors greatly influencing the creation of a controlled vocabulary are specificity and exhaustivity which are generally thought to be two forces that are in conflict with one another. Specificity, which corresponds to the number of hierarchical levels, is related to precision and exhaustivity; the term corresponds to the number of facets and is related to recall. Unfortunately

> [h]igh exhaustivity tends to lower precision, because the addition of more and more codes in the retrieval of irrelevant data alongside the relevant. Conversely, high specificity results in low recall. Since high specificity uses narrower categories, it produces fewer data in each category than does low specificity. (166)

Care must be taken, then, in the organization of data to find a balance between these conflicting qualities.

---

and other such professionals, its goals include simplicity of maintenance, commonly understood semantics, international scope, and extensibility. Though originally designed for "document-like objects," it is also flexible by design and can be used in cataloging other objects as we shall see.

[8] Controlled Vocabulary: Standard terms, usually in specialized language used in indexing, cataloging, and databasing in order to eliminate confusion and make searching more accurate. For example, a medical database might make the fine distinction between those conditions classified as a "disease" such as measles, and a "disorder" such as cri du chat which is genetic. Controlled vocabulary are often organized into an overall hierarchy such that a search for "marine mammals" will return more results than searches for "whales" or "dolphins" which are both types of marine mammals and would be recognized as subcategories by the database.

Greenberg's "Understanding Metadata and Metadata Schemes" provides background on the purpose of metadata. In addition to contextualizing and describing objects or data characteristics, the functional purpose of metadata can also include format documentation which is important for this archive because there are a number of ways to create a record of a social media posting (27). Greenberg covers a number of metadata schemes. At the simple end of the scale are Dublin Core, the Metadata Object Description Schema (MODS), and the Visual Resource Association's Core Categories (VRA Core). Of these, I feel the Dublin Core is the best fit for the project because of its simplicity and the fact that it is designed for digital-like objects, "which are defined as [a] textual object, [and] is also applicable to physical objects in many different formats" (30). The extreme simplicity of the Dublin Core is also designed to facilitate interoperability between systems and become more complex as needs arise. These will be important factors because censorship of online expression is not a phenomenon limited strictly to one country which means the project has the capacity to grow into a tool used worldwide by other watchdog organizations. According to Greenberg, the VRA Core is unlike the Dublin Core in that it is based on the cataloging of actual physical objects such as photographs and paintings. Because we are dealing primarily with digital objects which will commonly be accompanied by some form of text as part of the post (few will post images without adding some sort of note), the Dublin Core's digital and textual emphasis seem most appropriate.

One of this project's main goals is to create a glossary of cryptic neologisms to aid in future research in speech-oppressed online environments. I found a project with similar goals in the "Formosan Language Archive" by Zeitoun, Yu, and Weng. The archive is part of a larger attempt to preserve examples of over 24 Formosan languages which are already extinct or fast becoming extinct. As luck would have it, the archive's two main languages are Mandarin Chinese and English and it has therefore posed a number of poignant questions which also apply to this work.

One question encountered was the degree to which the censorship archive will accommodate interlingual cross-operability. That is, will every English entry also have a corresponding Chinese entry such that the archive can be used by native speakers in both languages? As an information scientist who is trying to address a human rights violation

through research and collection, I strongly believe it would be unconscionable to collect examples of this kind of oppression while leaving a linguistic barrier to those who it is aimed to assist. Unfortunately, this simple idea can effectively double the size and complexity of the database. When I consider also the possibility of expanding the database's reach to social media in other parts of the world such as the Middle East, the levels of complexity unfold again into exponentially larger degrees. While the ideal situation accommodates Chinese-speakers studying Chinese social media and Turkish-speakers studying Turkish social media, such an undertaking is well-beyond the current scope of this project and is worth being addressed as a major project in its own right.

This is not to say that the censorship database will not have language tags for both the collected data and the descriptive and explanatory notes, however. It will be essential for all information entered into the database to be tagged with the proper language. Doing so will facilitate future growth and linguistic expansion of the database such that it may expand functionality for users around the world.

The Zeitoun, Yu, and Weng article also brought up the necessity to include tags for the various project contributors along the way such as the fieldworkers who "collected" the language as well as transcribers and translators. Many of the correlating positions in this project can be entered automatically based on the indexer or translator's user log in identity. I believe this is especially important for the translation and explanation portions of the entry, as both of these entries themselves are created objects and should bear some marker, be it a name or independently generated identity number which would establish the creator identity. Having identity markers for the interpreting and cataloging operation also ensures a more uniform execution of input standards and facilitates efficient quality control.

Zeitoun, Yu, and Weng also underscored the importance of tags for morphemes. The Formosa language archive is meant to be a tool for recording as much data as possible for an array of rapidly disappearing languages and employs China Knowledge Information Processing tags to mark parts of speech in accordance with Chinese grammar concepts. The censorship glossarchive will not require such detailed descriptors as labeling every word with a part of speech, and therefore we have the luxury of simplification here. However, in order to meet our goal of finding and cataloging cryptic

references, we will need something similar that can isolate the word, phrase, and/or technique in question from the rest of the post and explain it. Because our approach takes the common metaphor as the basis for understanding the cryptic reference, we could label it a device (as in literary or artistic device) of a broader category, and include "cryptic reference" as one of the options next to "idiom" or "allusion."

The Lourdi, Papatheodoru, and Nikolaidou article "A Multilayer Metadata Scheme for Digital Folklore Collections" addresses the issue of dealing with heterogeneity as a factor in cultural preservation. One of the main problems with creating a digital archive of folklore in Greece was that much of the collection fieldwork was done in notebooks which were then stored as-is in the physical archive. These notebooks not only contained handwritten accounts, but also pictures such as drawings and photographs as well as recordings and actual physical objects such as small mementos and pieces of cloth. The solution to this problem was to create categories of metadata such as descriptive, technical, rights, and educational character and to arrange these in "nested" categories. A nested category is different than a hierarchy such as those used in the creation of controlled vocabulary as the categories help define how the various objects (notebooks, music recording) and not terms are related to each other. According to Lourdi, Papatheodoru, and Nikolaidou, by defining levels and how they relate to corresponding digital objects, the process of information discovery is enhanced

> either by (a) browsing the notebooks one by one or (b) searching their content keywords or combinations of several search criteria like time, place, usage etc. The proposed metadata model correlates all collection levels and defines the adequate elements for rich documentation and simultaneously making easier the task of the cataloguer, who is obliged to fill in the specific metadata fields for each level. (201)

In contrast with a controlled vocabulary, where hierarchy defines a term as either broad or narrow depending on the schema, nested categories allow different types of objects to share a common association with a given tale or event. In a similar way, the collected social media postings and their contents need to be related to a phenomenon or happening (such as an ongoing news story) or incident (such as a riot or activist arrest) which exists outside the post as related parts of a cohesive "story" of netizen reaction.

Another relevant article on creating a metadata schema for complex digital objects was the Lee, Tennis, Clarke, Carpenter effort to catalog video games for the Seattle Interactive Media Museum. The authors understood there is a great deal of ambiguity and subjectivity when describing video games and hit on two ways to address these issues. One way was to account for different types of users who would approach the games from different perspectives. These user types were classified as "personas" and were defined as player, parent, collector, academic, designer, and curator (107). Each of these came to the database with different information needs and criteria. Taking into account these needs helped to select the proper metadata for the database.

As previously mentioned, the proposed censorship database takes into account the possibility of users from other countries in allowing for future multilingual collection and cataloging. In addition to this, we need also to take into account security and the privacy rights of the users whose posts we collect. To this end, I decided that instead of defining metadata schema (with the exception of language) the different users or personas will be defined based on levels of security which will correspond to different levels of information access or what are called "views" in the field of database creation. These user-level are collector, researcher, random user, and the self-reporter.

The article also dealt with subjectivity in the area of game genre by recommending increased exhaustivity by increasing the number of facets (Lee, Tennis, Clarke, and Carpenter 113). As Given and Olsen caution, a greater number of facets will enhance recall but may have detrimental effects on precision. I decided to go with facets where type can be ascertained but depart when faced with something new that defies type. For example, the tradition of using homophones and homographs is well-known with some examples seeing frequent, even daily, use. Other techniques are certain to surface as the community of users mature and become more sophisticated (just as changes in technology will also lead to unexpected effects and unpredicted consequences). When this happens, we will need to tag the cryptic reference as "other" which will also prompt for a description that can later be used to justify the creation of a new class of cryptic reference.

Proposed Schema

The proposed schema places the cryptic reference as a type qualifier under a the Dublin Core element "Relation." The cryptic reference is dependent on the existence of the post content first and the existence of a reference to the tertiary phenomenon second. It is a descriptor of technique as manifested in the content of either the text or media and thus always follows the hierarchical order of "Description," "Relation," and "Type" (see Figure 9). The actual content of the posts is relegated to the Dublin Core "Description" element to relate the entire textual content of the post in the case of character-limited shorter social media posts or more generalized summaries for media such as images, longer posts such as blogs or entire websites, comments, or articles.

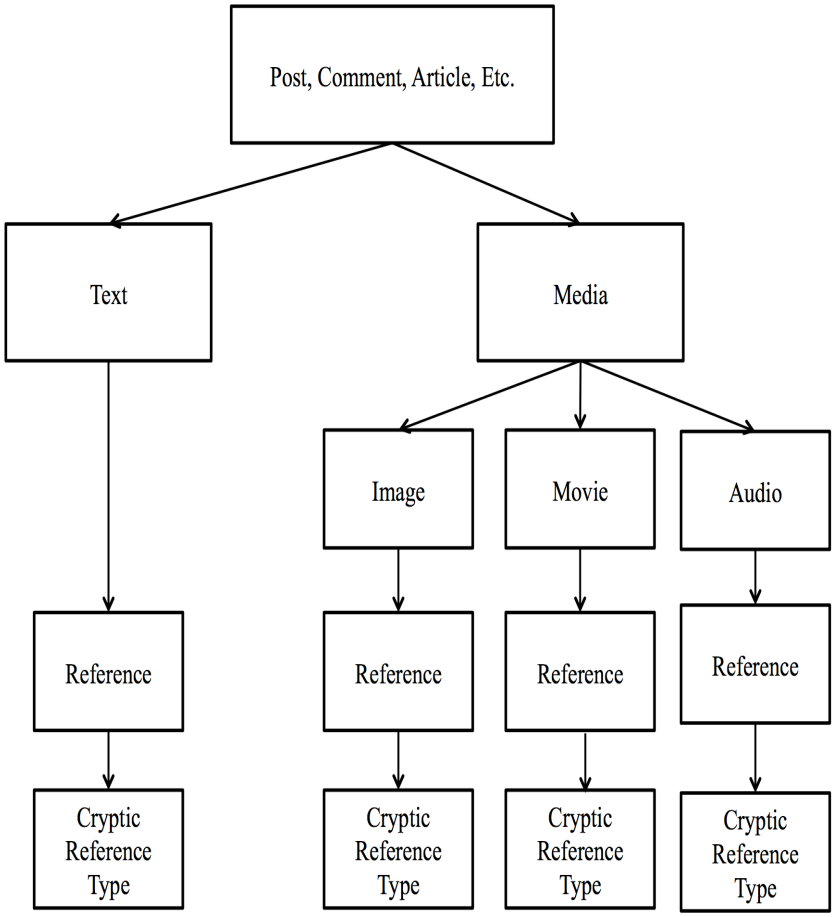

Figure 9. The Cryptic Reference serves as type amplifier for reference.

In an example application for a textual case, we take a March 21st instance where the homophone "river crab" is used to reference censorship activity. The archival process begins with the processing of the original post's representation (Figure 10).

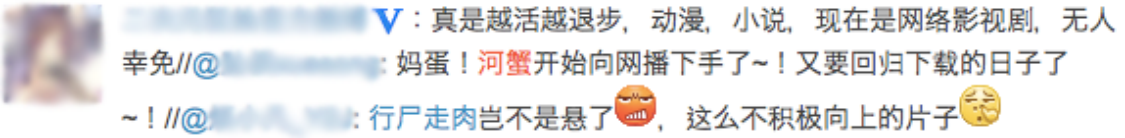

Figure 10. Screen capture of the River Crab post to be processed.

The post contents are input under "Description" as:

*Description="*妈蛋！河蟹开始向网播下手了~！又要回归下载的日子了~！//@XXXXXX: 行尸走肉岂不是悬了，这么不积极向上的片子*"*

This Description is also labeled as Chinese Mandarin. Two more descriptions are then added to this same original description, each with its own language tag. One will be marked pinyin:

Description="mādàn héxiè kāishǐxiàng wǎngbō xiàshǒule yòuyāo huíguīxià zǎi de rìzile @XXXX: xíngshī zǒuròu qǐbùshì xuánle nù zhèmebùjījí xiàngshàngdepiān zi wābíshǐ"

And finally a third description as the English translation:

Description="Dammit! The censors are starting to clamp down on the net! Once more I'll have to return days of downloads! //@XXXX: Aren't these zombies unsettling [Angry Face]. Such positive and uplifting films [Crying Face]."

All three versions are then given the same one-to-many relationship:

Relationship="References censorship"

The relationship element is then given the type element:

Type="homophone rivercrab for censorship"

The process is similar for images where the description of the image contents is labeled for any overt (such as a photographed message) or covert (such as wearing Guy Fawkes masks) messages. The type terminology used here does not jibe with recommended terminology in the Dublin Core Metadata Initiative, as it is being appropriated for a use which strays from its original design. The vocabulary utilizes my suggested innovation but it is not the only way to resolve the issue. The type could be listed as, for example, "homophone" which is then fleshed out in more detail with either a relation or description element such as *Description=river crab for censorship*. I also entertained the idea of suggesting a distinct element of relation type [RelationType] to both indicate a relationship as well as provide its nature.

The cryptic reference taxonomy discussed in the previous section, with its broader classes has been abandoned for this phase of the process. Valid superclasses of cryptic references can most certainly be said to exist; however, the concept will not be especially useful until a greater variety of examples can be collected. For the time being, the superclasses are an additional level of specificity that does little to aid precision.

The suggested cryptic reference schema is a requirement which arises from one of the primary research goals and may not find much use beyond this project. There are still a number of cataloging issues relating to Internet communication in general which should have much broader applicability, such as the suggested schema for cataloging social media posts.

Post collection and cataloging uses the same DC elements but in slightly different ways. For one, the "creator" element will be filled with the user name associated with the message while the "publisher" element will be the site domain (ie, weibo.com) and the "source" element will be be filled with the full link to the individual post (usually found by clicking on the date within the post). There are two dates that can be associated with the post itself, one for the original date as indicated at the bottom of the message, and the other should be the date it was processed for cataloging. The "format" element indicates how the post has been stored. In addition to various image formats for screen captures, a post can also be saved as a web archive. This format is ideal as it preserves as much as possible from the original post to the point where links can be clicked to explore the connections, likes, and re-posts. Unfortunately, this format results in larger files and in some cases, the web archive takes a while to generate. At the other end of the spectrum are text captures of the screen content. This is nothing more than simply highlighting the text of the post in question and pasting it into a document program such as Word. Depending on the sophistication of the document program, the copy may carry over certain graphic additions from the original such as font color or animated emoticons, but many important formatting cues will be lost, and the end result could be a confusing mash of content, buttons, or functions all displayed as the same type of textual information.

Posts have a mandatory "accrual method" element to differentiate how it was collected. There are three total probable values for this field. The first will be self-report

or second-hand report which will be automatically assigned when Chinese users access the site to relate personal experiences with censorship. The “creator” entry for these will default to “anonymous.” The second possible value for “accrual method” is crawler or other automatic method. It goes without saying the “creator” field for these will be N/A and it will serve primarily as a place-holder entry until it can be processed by a cataloger or researcher, in which case the method will switch to cataloger and be associated with a “creator” field entry of user identity designator. All “accrual method” entries will also include a date field to indicate the date when the post was processed for collection.

The views enabled by the three different personas is an essential function for ensuring the censorship glossarchive project is not exploited in a malicious manner. The three levels planned for are the collector role (to include both the cataloger and collection management personnel), the outside researcher role, and the self-reporter role. The first has full access while the last has only temporary access to their own entered data until the “submit” button is hit. The second role will have variable access depending on the information needs of the researcher. Figure 11 breaks down the various data fields and the possible views for each persona.

ROLES	Collector, Curator	Researcher	Self-Reporter
DATA FIELDS			
POST DATA			
Desc/Text Contents	F	L	T
Has/Image, etc.	F	F	T
Post Date	F	F	T
Source/ Post Link	F	L	T
Creator/User name	F	L	T
Publisher/Domain	F	F	T
Format/Archive, etc.	F	F	T
Accrual Method/How entered	F	F	T
Accrual Date	F	F	T
DESCRIPTION DATA			
Desc/Text Chinese	F	F	T
Desc/Text Pinyin	F	F	X
Desc/Text Pinyin/Creator	F	X	X
Desc/Text English	F	F	X
Desc/text English/Creator	F	X	X
Relation/Reference Sensitive Issue	F	L	X
Type of Relation/Homophone, etc.	F	L	X

Key	Meaning
L =	Limited or Provisional Access
T =	Temporary Access, Own data only
F =	Full Access
X =	No Access

Figure 11. List of major data points and degree of access allowed for each user type.

In addition to the departure from the Dublin Core Metadata Initiative recommendation mentioned in the above explanation of the "type" field, the database makes another departure from the recommended standards by using the standard United States government and Department of Defense language trigraphs from the Common Human Resources Information Standards (CHRIS) document.[9] The Dublin Core recommended ISO 639 standard is woefully lacking. It contains only one code for "Chinese" and lacks codes for the Cantonese and Wu dialects which are the two largest Chinese dialects outside of Mandarin. The CHRIS language list, on the other hand, accounts for 13 dialects of Chinese alone in addition to hundreds of other languages and

[9] The Common Human Resources Information Standards document contains trigraphs for over 500 languages and dialects and is used by the Peace Corps and Department of Defense for designating language ability for personnel management. The version promulgated 16 August 2011 is available at URL: http://discovhr.hrm.osd.mil/DT2/discovhr.html?CHRIS:&155

dialects to include many of the languages of Asia, the Middle East, six dialects of Spanish (including Chavacano), and even Latin and Old English.

The cataloging elements discussed above represent a bare minimum for processing individual Internet communication events. In fact, there are many more elements which can be added in order to increase entry granularity. For example, a suggested rudimentary controlled vocabulary and system of entry standards are in the appendices.

There are several ways to input text when the "relation" element indicates a reference to a controversial topic. One would be to use *relation="references censorship"* as above, but for more detail, the words *criticism of*, *intimation regarding*, and *accusation regarding* could be used. These three words can be seen as logically nested terms, at least as their usage is envisioned here, with each communicating a stronger sense of protest than the last. For this project, *criticism* is used to convey a negative opinion but without any claim to harm. The *"references criticism of censorship"* therefore communicates the post relayed a certain level of displeasure or dissatisfaction but did not go so far as to claim any verifiable negative effect. In the case where *intimation regarding* is used, the post is communicating some harm was done or wrongdoing committed but the evidence is either anecdotal or secondhand. For entries where an *accusation regarding* is used, the post itself claims to offer evidence of harm or wrongdoing such as a photograph of the aftermath of police brutality or a picture of an arrest warrant.

The Dublin Core element *Subject* can also be used to associate any number of keywords or subject tags describing not only the content of the message but the types of parties involved. The recommended tags are available in Appendix A.

FUTURE RESEARCH AND CONCLUSION

Probably the greatest indicator of the potential of this project to lead to a further understanding is the many new avenues of research that can be conducted to gain a clearer picture of modern man's place in this sub-created landscape that is the Internet. The new research ranges from the quantitative to the qualitative and requires a variety of fields of expertise including but not limited to political science, languages and linguistics, psychology, computer science and programming, and philosophy.

One of the largest missing pieces in the puzzle is a survey of Chinese Internet users' knowledge of and attitudes towards the state's censorship apparatus. It makes a huge difference if users believe the immediate threat comes from an Internet bot or a human being capable of quickly deciphering polyvalent references or cues. It makes a difference if users believe their posts are in danger of being cut for the content or the popularity of their topic. It makes a difference if the average user knows what the Great Firewall is and still a bigger difference if their understanding of how to get around it is common knowledge or expert advice. Filling in these gaps would bring the cyber-ethnographer closer to understanding motivations which would then lead to clearer mapping of inferences and intentions.

Other research could examine more closely censorship protocols. Are jingoistic statements supporting the mainland PRC views on Taiwanese independence more tolerable than those which demonize Japan even if they are equally popular? Does censorship reveal a weakness for humor? These are questions that can be answered only by thorough analysis and comparison of trends, and for that we need more data and better indexing and cataloging systems. For example, collection and analysis may enable us to see which posts are organic and sincere expressions of sentiment and which are "sock puppet" users trying to influence the conversation.

Better anonymity software and more secure equipment can empower researcher and expat activist alike to conduct more interactive forays into the social media networks of various countries. Now, several companies are envisioning ways in which the Internet's reach can expand and its presence become more pervasive. What technological advancements would facilitate this growth? What defenses will these innovations deploy? These are questions for engineers, programmers, cyber security professionals,

communications experts, and computer scientists to grapple with, but at this time we cannot fully understand the impact of these technological advances.

This project casts light on an obscure subject. We must not forget the place where censorship is committed is the scene of a crime. In light of the Universal Declaration of Human Rights, it is a violation to silence dissent voicing valid concerns and criticisms. We virtually examine the crime scene as detectives when the crime itself is the removal of all signs of commission. A regime so weak it fears the opinions of its own people acts more out of self-interest than out of service and is less committed to the communal good and more committed to delaying its own decline.

This endeavor is as much about compassion for one's fellow man as it is about capturing history in the moment. Equal parts advocacy and data collection, it is an attempt to express solidarity even as it collects others' expressions. We have both the means and the rationale to peer beyond the filter of official organs such as *Xinhuashe* or the Sina website. By working to raise awareness of human rights and doing what is ethical in the face of grave abuses, we can learn more about about information control while pioneering technologies and methods promoting information freedom.

APPENDIX A
SAMPLE METADATA TAGS WITH DUBLIN CORE CATEGORY AND TAG DEFINITION

Metadata Tags	**Recommended DC element, usage/definition**
acrostic	Relation> Description, characters conveying cryptic meaning through their arrangement on the web page
activist	Subject, person or persons known to be strong advocates of a position.
altergraph	Relation> Description, characters conveying cryptic meaning through their visual appearance or arrangement
audio	Description or Type, audio file Format mandatory.
business	Subject, an enterprise or corporation organized around the sale of goods or services for profit
city	Subject, municipality
city government	Subject, the institution bearing legal authority over its citizens at the municipal level
civilian	Subject, non-military citizen
criticism of	Relation> Description, post conveys disagreement or dissatisfaction concerning some sensitive topic
cued message	Relation> Description, characters conveying cryptic meaning which can only be understood after a certain operation is performed
family	Subject, a group of people, usually related and usually living in the same household
government	Subject, the institution bearing legal authority over its citizens

hint	Relation or Subject, obscurely conveyed
homophone	Relation> Description, characters conveying cryptic meaning through their pronunciation
image	Description or Type, picture file Format mandatory
incrimination regarding	Relation> Description, post purports to offer some degree of evidence that harm has resulted or wrongdoing committed
individual	Subject, a person
intimation regarding	Relation> Description, post conveys belief in some harm has been done or wrongdoing committed
jumble	Relation> Description, characters of a message rearranged such that the true message cannot be read by reading in a normal fashion
lawyer	Subject, a lawyer
military	Subject, institution charged with the defense of the nation
movement	Subject, a group dedicated to a certain political end
national government	Subject, the institution bearing legal authority over its citizens at the national level
neighborhood	Subject, select grouping of houses and streets
network	Subject, arranged communication along preset channels, or persons utilizing same
officials	Subject, members of national or local government bodies or their agencies
PAP	Subject, the People's Armed Police
PLA	Subject, People's Liberation Army (i.e. army)
PLAAF	Subject, People's Liberation Army Air Force (i.e. air force)
PLAN	Subject, People's Liberation Army Navy (i.e. navy)

PLANAF	Subject, People's Liberation Army Navy Air Force (i.e. naval air force)
police	Subject, the police
political organization	Subject, organization with ties to a political party
political party	Subject, politically organized group which may or may not be officially recognized as a party by the government
protestors	Subject, person or group publicly demonstrating
provincial government	Subject, the institution bearing legal authority over its citizens at the level of province
religious group	Subject, person belonging to a religious organization such as a church or cult
report	Relation or Subject, relaying information, story or conversation
resistors	Subject, members of organized anti-government movement
sarcasm	Relation or Subject, sarcastic turn of phrase
security	Subject, private employees paid to provide security
site	Subject, website
threat	Relation or Subject, conveying imminent harm or danger if certain conditions are not met
town	Subject, town
town government	Subject, the institution bearing legal authority over its citizens at the town level
video	Description or Type, movie file Format mandatory

APPENDIX B
GLOSSARY

Altergraph	(n.) Coined term referring text that has been visually changed in such a way as to obscure the meaning from machine reading.
Captcha	(n.) Short for "Completely Automated Public Turing test to tell Computers and Humans Apart". A reverse Turing test used to verify a user's humanity.
Crawler	(n.) Type of program which scours web pages for content for databasing or statistical analysis by searching for and collecting key terms in the text.
Cryptic Circumlocution	(n.) A technique of "talking around" banned or sensitive topics in such a way that decreases or eliminates the likelihood of one's online speech of being censored or blocked. Although the term broadly refers to textual manipulations such as homophones or altergraphs, for the sake this speech research images are included as well because they are difficult for crawlers to automatically recognize and tag.
Cued Message	(n.) Coined term referring to a message hidden in such a way that it cannot be understood unless a type of operation is performed on the text such as rearranging the characters.
Five Poisonous Groups	(idiom) An association of people whose held beliefs strongly contrast with government ideology in China. These groups are

	Tibetans, the Uighur, Pro-democracy Movement, Falun Gong practitioners, and proponents of Taiwanese independence.
Great Firewall	(n.) GFW for short, the firewall which prevents free exchange of information between China and the global Internet, effectively creating a national intranet for Chinese users. This can be circumvented through the use of proxy servers.
Harmonize	(vb.) Government censorship is conducted in the name of the greater social good, or "harmony". When a site or other type of page is censored, users are told it has been deleted for the sake of harmony. Thus, the passive construction “was harmonized” became a cynical reference to the process of being censored, and its homophone, which means "river crab" became another sardonic reference to censorship in general.
Homophone	(n.) A type of cryptic circumlocution where the true meaning of the word is hidden in the pronunciation beneath irrelevant characters.
Horizontal Communication	(adj. n.) According to King, this is communication with or among the general population. Vertical communication can be seen as expressing one's feelings to their government, while horizontal communication is more grassroots. It is the type of communication which can spread ideas to others and thus must be monitored in order to maintain social control.
Leet speak	(n.) Short for "elite speak,” this is a type of altergraph utilized by hackers to disguise written English so that it could not be

understood by non-humans.

Metadata	(n.) Data about data. Descriptors for fields used to enter information so digital objects can be cataloged.
Metadata schema	(n.) The manner in which metadata requirements and fields are arranged in order to catalog or index a certain type of object be it digital, intellectual, or physical.
Netizen	(n.) Indicates the user of the Internet within the specific context of a given site or borders of a given nation with rules, regulations, customs, and laws to which the users and service or content providers must adhere.
Re-post	(vb.) Generic for "re-tweet". Used on Weibo.
Social Media	(adj. n.) Internet sites which allow for social interaction. In the United States popular examples include facebook and Twitter.
User	(n.) Generally, one who accesses the Internet and specific sites. Specifically, one who creates a site account to access online services, such as banking or social media.
Weibo	(n.) 微薄 Chinese version of Twitter.

WORKS CITED

American Library Association. *Code of Ethics for Librarians*. 1939. Web. 1 Apr. 2014. <http://www.ala.org/advocacy/proethics/history/index5>.

American Library Association. *Statement on Professional Ethics, 1981 Introduction*. 1981. Web. <http://www.ala.org/advocacy/proethics/history/index4>.

Chin, Josh, and Brian Spegele. "Censorship Protest Gains Support in China." *Wall Street Journal*, 7 Jan. 2013. Web. 3 Mar. 2013. <http://online.wsj.com/news/articles/SB10001424127887323706704578227502841925808>.

Cook. Sarah. "The Long Shadow of Chinese Censorship: How the Communist Party's Media Restrictions Affect News Outlets Around the World." Center for Internet Media Assistance, 22 Oct. 2013. Web. 10 Nov. 2013. <cima.ned.org/sites/default/files/CIMA-China_Sarah Cook.pdf>.

Department of Defense. *Common Human Resources Information Standards: Foreign Language (Reference Element)*. 16 Aug. 2011. Web. 29 Mar. 2014. <http://discovhr.hrm.osd.mil/DT2/discovhr.html?CHRIS:&155>.

Dublin Core Metadata Initiative. 2005. Web. 14 Mar. 2014. <http://dublincore.org>.

Engber, Daniel. "Who Made That Captcha?" *New York Times*, 17 Jan. 2014. Web. 18 Feb. 2014. <http://www.nytimes.com/2014/01/19/magazine/who-made-that-captcha.html?_r=0>.

"Freedom on the Net 2013." Freedom House, 3 Oct. 2013. Web. 10 Nov. 2013. <http://www.freedomhouse.org/report/freedom-net/freedom-net-2013>.

Gil, Paul. "What is the 'Invisible Web'?" About.com, Aug. 2013. Web. 22 Sept. 2013. <http://netforbeginners.about.com/cs/secondaryweb1/a/secondaryweb.htm>.

Given, Lisa, and Hope Olson. "Knowledge Organization in Research: A Conceptual Model for Organizing Data." *Library and Information Science Research.* 25 (2003): 157-176. *ProQuest*. Web. 12 Mar. 2014.

Guo, Aw. "So, How do the Chinese Bloggers Avoid the 'Censorship'?" IfGoGo.com, 1 July 2008. Web. 3 Jul. 2013. <http://www.ifgogo.com/95/how-do-bloggers-avoid-censorship-in-china/>.

Greenberg, Jane. "Understanding Metadata and Metadata Schemes." *Cataloging & Classification Quarterly*, 40:3-4 (2005): 17-36. *ProQuest.* Web. 12 Mar. 2014.

Harwit, Eric, and Duncan Clark. "Government Policy and Control over China's Internet." *Chinese Cyberspaces: Technological Changes and Political Effects.* Eds. Jens Damm and Simona Thomas. New York: Routledge, 2006. 12-41.

Kaiman, Jonathan. "China Cracks Down on Social Media with Threat of Jail for 'Online Rumors.'" *The Guardian*, 10 Sept. 2013. Web. 22 Sept. 2013. <http://www.theguardian.com/world/2013/sep/10/china-social-media-jail-rumours>.

King, Gary, Jennifer Pan, and Margaret E. Roberts. "How Censorship in China Allows Government Criticism but Silences Collective Expression." *American Political Science Review,* 107.2 (2013): 326-343. *ProQuest*. Web. 11 Nov. 2013.

Lee, Jin Ha, Joseph T. Tennis, Rachel Ivy Clarke, and Michael Carpenter. "Developing a video game metadata schema for the Seattle Interactive Media Museum."

International Journal on Digital Libraries, 13 (2013): 105-117. *ProQuest.* Web. 10 Mar. 2014.

"Leet." Wikipedia. Web. 15 Mar. 2014. <http://en.wikipedia.org/wiki/Leet>.

Lourdi, Irene, Christos Papatheodoru, and Mara Nikolaidou. "A Multilayer Metadata Scheme for Digital Folklore Collections." *Journal of Information Science*. 33.2 (2007): 197-213. *ProQuest*. Web. 11 Mar. 2014.

Orwell, George. *1984*. New York: Signet Classics, 1949.

People's Republic of China. *Constitution of the People's Republic of China*. 5th National People's Congress. 4 Dec. 1982. Web. 20 Feb. 2014. <http://english.gov.cn/2005-08/05/content_20813.htm>.

Sherman, Chris, and Gary Price. "The Invisible Web: Uncovering Sources Search Engines Can't See." *Library Trends*. 52.2 (2003): 282-298. *EbscoHost*. Web. 29 Apr. 2014.

United Nations. *Universal Declaration of Human Rights.*10 December 1948. Web. <http://www.un.org/en/documents/udhr/>.

Ward, Katie. "Cyber-ethnography and the Emergence of the Virtually New Community." *Journal of Information Technology,* 14 (1999): 95-105. *ProQuest*. Web. 5 Jan. 2014.

Zeitoun, Elizabeth, Ching-hua Yu, and Cui-xia Weng. "The Formosan Language Archive: Development of a Multimedia Tool to Salvage the Languages and Oral Traditions of the Indigenous Tribes of Taiwan." *Oceanic Linguistics,* 42.1 (2003). *ProQuest*. Web. 12 Mar. 2014.

“各大网站罕见以藏头诗声援南周抗议中宣部 [Rarely Seen Hidden Wordplay on Major Websites Support Southern Weekly and Protest Central Propaganda Section].” *Epoch Times*, 8 Jan. 2013. Web. 2 Feb. 2013. <http://www.epochtimes.com/gb/13/1/8/n3771136.htm各大网站罕见以藏头诗声援南周抗议中宣部.html>.