TECHNOLOGY SCAN:

PROFILES OF INFORMATION AND COMMUNICATIONS TECHNOLOGY (ICT) PROJECTS WITH POTENTIAL APPLICATION FOR ADDRESSING DANGEROUS SPEECH

Prepared For

The Sudikoff Annual Interdisciplinary Seminar on Genocide Prevention
February 20-21, 2014

Center for the Prevention of Genocide
US Holocaust Memorial Museum, Washington, DC

I. **Overview**
II. **Headline findings**
III. **Profiles**

A. Detecting, analyzing and/or visualizing information

1. Umati
2. CrisisTracker
3. Humanitarian Tracker
4. Geo Listening
5. EMM NewsBrief

B. Projects that use ICT to effect engagement or action

1. Uchaguzi
2. Sisi Ni Amani
3. The Early Warning Radio Network and The LRA Crisis Tracker
4. UN ICT Emergency Humanitarian Platform and emergency.lu

C. Relevant technologies

1. Predictive analytics / Crush
2. Social network analysis / Netlytic and Orgnet
3. Sentiment analysis / Lexalytics and Crimson Hexagon

D. Sample project: Violent extremism online.

TECHNOLOGY SCAN:
PROFILES OF ICT PROJECTS WITH POTENTIAL APPLICATION FOR ADDRESSING DANGEROUS SPEECH

I. Overview

Information and communications technologies (ICT) provide the means to disseminate hate speech through multiple channels, instantaneously and on a mass scale. Such dissemination can be as quick and easy as clicking “send,” and examples of how technology has been used in this way are common and well known.

Countering such speech requires a great deal more effort and planning. While there’s no simple solution, and the uses of ICT to counter dangerous speech are mostly still embryonic, an array of projects are showing ICT’S potential for detecting, analyzing, and countering dangerous speech – without infringing on freedom of expression.

The focus of the 2014 Sudikoff Seminar is “Countering Dangerous Speech, Protecting Free Speech: Practical Strategies to Prevent Genocide.” Its goal is to engage participants in developing a toolkit of specific policies and approaches that can prevent and/or counter dangerous speech – the subset of hate speech with the potential to catalyze violence -- while preserving the right to free expression.

As part of the seminar, participants will discuss potential uses of ICT for this purpose. As a way to stimulate ideas and thinking, the Center for the Prevention of Genocide has compiled short profiles of a range of projects, companies and technology fields that are working to address hate speech or undertaking analogous work in the human rights field. We present these profiles to stimulate discussion about how ICT can be used to address dangerous speech. We are particularly interested in exploring the capacity of ICT to:

1. Detect when dangerous speech is happening, its intensity, and where it is on the rise.

2. Identify and analyze the reach and exposure – and methods of dissemination -- of such messages.

3. Counter dangerous speech without restricting free speech.

The field of individuals and organizations using ICT to address dangerous speech is new and emerging. The field of those using ICT to address hate speech more generally or on a wider range of human rights issues is more developed. While our primary interest is

This paper was prepared by Jill Savitt, Senior Advisor to the Center for the Prevention of Genocide. NB: All underlined terms are hyperlinks.

dangerous speech, we have catalogued projects that also use ICT to address a wider range of issues because they may generate new thinking on how to address dangerous speech.

The projects profiled on the following pages are not the only initiatives that exist, and they have been selected solely to serve as points of reference. Their inclusion here is not an endorsement of their goals or efforts. Many of these projects have not been evaluated (either because they are too new or because the originators of the projects have either not yet evaluated them or made the evaluations public). Therefore, this paper does not make a judgment on the effectiveness of these projects. Rather, this paper has been prepared to identify a range of ICT efforts with potential relevance for using innovative technological approaches to detect, map, address or counter hate speech and dangerous speech.

It deserves mention that there are scores of web-based media literacy programs designed to educate individuals, especially young people, about hate on the web, as well as a range of organizations that monitor hate on the web. This paper does not attempt to profile such efforts.

II. Headline findings

The process of compiling these profiles has generated some findings about this emerging field:

- **Detecting the problem is not necessarily the problem.** The field of *detecting* dangerous speech is much more evolved than the field of *addressing or countering* it, and those undertaking hate speech detection efforts often do not marry their analysis or visualizations with projects to counter hate speech.

- **Addressing the problem is still an issue.** A great deal of experimentation is happening in the use of ICT to address or counter dangerous speech, but there are still no "best practices" in this arena. The Umati project is the most advanced in the field and has the potential to generate best practices and replication.

- **Human input.** Because of the sensitivity of the issues addressed and nuances in language, ICT projects about dangerous speech cannot be fully automated. In descriptions of their efforts, project creators note the importance of human intervention – analysis, input, judgment – in ICT projects to address or counter hate/dangerous speech.

- **Language barriers**. Many of the projects noted the need to have language experts involved in translating not only literal speech, but in interpreting coded language and tone.

- **Detecting rising temperatures**. Detecting hate speech as a snapshot does not provide a clear picture of potential danger or violence. Many of these projects indicate that charting the change in sentiment over time – in volume or negativity/tone – presents a clearer picture of threats of violence based on dangerous speech.

- **Detecting reach.** Newer technologies are able to analyze the reach of information, which is increasingly important for this work. Understanding how dangerous speech is being disseminated – and the reach of its exposure – is critical to assessing both its capacity to catalyze violence and how best to counter it.

III. Profiles[*]

The following profiles of ICT efforts fall into three basic categories:

a. Projects that detect, analyze and/or visualize information on particular subjects or types of speech

b. Projects that connect detection, analysis and/or visualization of information to engagement or action.

c. Examples of relevant technology fields.

Under each of these headings, we have identified one or two projects to serve as examples; these are provided merely as illustrations. NB: All underlined terms are hyperlinks.

A. DETECTING, ANALYZING AND/OR VISUALIZING INFORMATION

A range of organizations have devoted resources to detecting and cataloguing examples of hate speech or other aspects of a human rights or humanitarian crisis, and then mapping or visualizing this data so it can be: used to tell a powerful story, viewed geographically, utilized as a tool to connect people with information or resources, and analyzed over time.

1. Umati

During the 2007 general election in Kenya, mobile text messages were used to incite violence among the population. Leading up to the 2013 election, Umati monitored, recorded and analyzed incidents of hate and dangerous speech in social media (Facebook and Twitter), online blogs/forums, websites and the comment sections of online newspapers in six languages. The examples of speech detected were then collected and categorized according to their “dangerousness,” using the criteria for dangerous speech developed by Susan Benesch. A main goal of Umati was to define – and widely publicize

[*] Much of the language for each of these profiles was taken directly from text written by the creators of the projects or technologies – paraphrased or quoted directly -- as a way to preserve their meaning and intention.

-- the type of speech that was “harmful to Kenyan society and use this early detection as a way to prevent the violence such speech has the potential to catalyze.”

While heavily reliant on human input, Umati has developed a significant database of inflammatory speech – the largest to date with more than 5,000 examples -- and has been moving to use this database to automate the collection process, using machine learning and natural language processing. The project now tracks trends and fluctuations in “dangerous speech” in the Kenyan online space, identifies its most common sources and targets, and correlates trends with triggering events. Space prevents a full accounting of Umati here, but the project has released several reports about its efforts, which merit review since this project is by far the most advanced in the field of using ICT to respond to dangerous speech.

2. CrisisTracker

CrisisTracker is a web platform that extracts “situation awareness reports” from Twitter messages during human rights or humanitarian disasters. The platform was founded based on the fact that during human rights and humanitarian crises, online social media – and particularly Twitter -- have emerged as a means for affected populations to communicate to the world what is happening on the ground. CrisisTracker filters and organizes tweets and then geo-locates them according to location, with a special emphasis on highlighting actionable information.

CrisisTracker’s technology combines *automated processing* with *crowdsourcing techniques.* Breaking this down: CrisisTracker’s automated algorithm identifies patterns of repetition in messages from multiple people who are independently reporting on the same event (such as an attack or a natural disaster in a particular place) and then clusters similar messages as an index about an event, geo-locates the index on a map and then enhances the index with images, video and news articles to create a “situation awareness report” about the event. The project is intended to make “distributed knowledge (shared primarily through Twitter) more accessible to emergency response professionals, victims and others, as well as to simplify direct communication between members of different groups.” As an example: CrisisTracker manages a project on Syria that creates “situational awareness” reports from various cities affected by the violence by combining clustered tweets with news stories, photos and video.

4. Humanitarian Tracker

Humanitarian Tracker “offers tools, methods, and training by which citizen journalists can share reports of what they witness on the ground, worldwide, about human rights violations, the spread of disease, rape, conflicts, or disasters.” From their vantage point in local communities, users share text, photo or video reports, which are then aggregated on a live map. The platform also uses data mining tools that scan sources on the web (official news reports, Twitter, Facebook and blogs) and plots that information as well. The goal of Humanitarian Tracker is to “expose injustice and abuse, creating narratives to challenge the status quo and mobilize for action.”

By way of example of the platform in action, the Humanitarian Tracker has been applied to two projects in Syria: 1) Epidemic Tracker to help clinicians track tuberculosis and viral hepatitis epidemics in Syria as a result of the unrest and 2) Syria Tracker, a crowd-sourcing effort to collect reports of human rights violations and causalities. Since it was launched in April 2011, Syria Tracker has published more than 4,000 geo-tagged eyewitness reports from citizen journalists and more than 160,000 official news reports (from a filtered stream of media news, Twitter messages, and Facebook postings).

3. Geo Listening

The previous examples in this section outlined data collection and visualization across countries and regions. Data collection can also be done in discrete targeted communities. Concerned about bullying online among teens, several companies have emerged to help schools monitor the online activity of their students. Geo Listening, for example, "monitors, analyzes and reports on social media posted by students from publicly available forums and provides a report on the frequency and severity of students' posts related to bullying, cyber bullying, despair, hate, harm, crime, vandalism, substance abuse, and truancy." Geo Listening searches keywords and sentiments on posts that can be viewed publicly and provides schools with a customized report. The service cannot read Facebook posts that are designated for "friends" or "friends of friends" and does not monitor email, SMS, MMS, phone calls, voicemails or tamper with any privacy setting of a social network user. According to the chief executive of Geo Listening, the company's reports are "a sprinkling of technology and a whole lot of human capital."

4. EMM NewsBrief

Many of the projects listed above combine crowdsourcing and automated processing of information about a particular event or subject from mainstream news sources. Over the past two decades mining media has been revolutionized. From literal "clipping services" (scissors, tape, photocopying) to Lexis/Nexis to Google searches and now Google Alerts, the task of compiling news on an issue or event can now be entirely automated.

Newer services that compile news coverage now provide analyses of the content, which can be especially useful in detecting hate or dangerous speech. For example, **EMM NewsBrief**, "gathers reports from news portals world-wide in 60 languages, classifies the articles, analyses the news texts by extracting information from them, aggregates the information, issues alerts and produces intuitive visual presentations of the information." The software that powers EMM was created by the Joint Research Centre to mine and cluster mainstream news automatically. All news articles referencing the same event or subject are grouped into clusters and displayed by cluster size. Summaries are produced every ten minutes, thus providing an analysis of the extent of coverage on given topics in real time. This software does not search social media.

B. PROJECTS THAT USE ICT TO EFFECT ENGAGEMENT OR ACTION

In the previous section, the featured projects focused primarily on detecting, collecting, analyzing and visually representing data. Many of the following projects collect such data but, significantly, also use ICT tools to stimulate responses to the phenomena the data record.

1. **Uchaguzi**

Uchaguzi (which means "elections" in Kiswahili) is an application of the Ushahidi crowdsourcing platform, which allowed Kenyans to report on election-related events on the ground In Kenya, including inflammatory speech and violence, via SMS.

During the 2013 elections in Kenya, Uchaguzi was able to monitor elections in real time and act as a central repository of data. This enabled Uchaguzi staff to issue warnings to citizens and law enforcement in areas where violence was breaking out or imminent. The goal of the effort was to extend the common practice of traditional election observation by engaging citizens in election monitoring. Citizens are a valuable source of information for election observers since they can verify and further explain examples of violence or dangerous speech to electoral authorities or security personnel, therefore considerably expanding "eyes and ears" on the ground.

2. **Sisi Ni Amani**

Sisi ni Amani works in three main areas in Kenya: "SMS-based programming for civic-engagement and peace; mitigating land conflict through dialogue and education; and civic engagement through forums and debates." The Sisi Ni Amani SMS project was created to use mobile technology as a tool for peace in the lead up to the 2013 Kenyan elections, recognizing the role that texts and other forms of communication technology played in contributing to violence in the previous election. To use this platform, "community peace ambassadors" send out, via SMS, "peace messages targeted at specific incidents at a micro level with the aim of preventing, reducing or stopping election-based violence. Members of the public can also use the platform to report incidents of violence within their communities free of charge." In addition, Sisi Ni Amani uses SMS to issue violence prevention messages and to combat rumors on a range of issues, particularly grievances over land-based issues. The organization relies on local leadership to identify emerging issues and to serve as the messengers for educational and positive messaging.

3. **The Early Warning Radio Network and The LRA Crisis Tracker**

The weak technology infrastructure in the Democratic Republic of Congo and the Central African Republic makes it difficult to report, in real time, atrocities committed by the Lord's Resistance Army and prevents remote communities from learning about potential LRA attacks. Invisible Children's Early Warning Radio Network is a system of high-frequency, two-way, long-range radios that allow communities in DRC and CAR to report LRA movement to one another on regular security calls. The radio network also alerts security and humanitarian groups to LRA activities. Information from the radio network is vetted by Invisible Children and then fed into **The LRA Crisis Tracker**, a crisis-mapping website that provides near-real-time information on current LRA activity. The LRA Crisis Tracker also includes a digital map, a breaking news feed, regular data-analysis reports, and a mobile application. The cloud-computing platform Salesforce, a content management system, delivers data to the mapping system and application.

4. UN ICT Emergency Humanitarian Platform and emergency.lu

The humanitarian response arm of the United Nations has created large-scale ICT programs for emergency response operations. These projects provide for the immediate deployment of ICT staff to humanitarian crisis situations, for instance deploying satellite infrastructure and capacity; communication and coordination services; satellite ground terminals for regular and transportable antennas and the transportation of equipment to a disaster area. Two of the major partnerships for the UN team are with Vodaphone and emergency.lu, a private-public-partnership. Relevant to the field of hate speech, these projects have allowed humanitarian responders to send mass broadcast text messages to affected populations using mobile devices -- equipment that affected individuals already have available to them.

C. RELEVANT TECHNOLOGIES

The tools and capacities of a range of specialized technology fields hold great promise for addressing dangerous speech. The following are brief descriptions of some of these fields and examples of projects or companies using them.

1. Predictive analytics.

Predictive analytics combines statistics, modeling, machine learning, and data mining to analyze current and historical data as a way to make predictions about future or potential events. The sectors that most commonly use predictive analytics are the insurance and actuarial sectors, marketing, financial institutions, retail and pharmaceuticals, among others. For example, predictive analytics is a centerpiece of credit scoring to predict a customer's ability to take on debt based on credit history, previous loan applications, current income and the like.

- **Crush (Criminal Reduction Utilizing Statistical History).** One example of where predictive analytics could intersect with dangerous speech is its use in a software package created by IBM called Crush (Criminal Reduction Utilizing Statistical History). Crush seeks to enhance police force effectiveness by predicting where and when future crimes might take place, based on an evaluation of patterns of past and current criminal incidents. The software combines such data with a range of other datasets including crime reports, intelligence briefings, offender behavior profiles and even weather forecasts. Once aggregated, Crush then identifies potential hot spots and flashpoints, so police forces can allocate resources to areas where particular crimes are most likely to occur. Applying this technology to hate speech might include compiling data about previous incidents of hate speech and related violence with information about the locations of known purveyors of such speech and a calendar of future potential triggering events (such as elections or anniversaries of significant occasions).

2. Social network analysis

Social network analysis is an area of computer science that views social relationships in a network – based on how individual nodes are related by contact, instances of communication, organizational affiliations and other connections. These networks are then depicted visually to show degrees of connectedness. This technology has been used, for instance, to chart the spread of contagious diseases. It can also be used to map the flow of information, power dynamics within communities, or criminal behavior, among other issues. Social network analysis focuses, in particular, on how connections influence choices and behaviors -- and the consequent effects of these actions.

Related to the field of hate speech, social network analysis can be used to map the reach of messages (how many people are exposed to it) and how it was disseminated, as well as the source of the message and the sphere of influence of the message's purveyor. There is a great deal of activity in this field which has been growing rapidly. The following examples – by no means unique -- are provided to illustrate how companies promote the capabilities of their software.

- **Netlytic**. Netlytic is a "cloud-based text and social networks analysis company that can summarize large volumes of text to reveal social networks based on online conversations on social media sites such as Twitter, YouTube, blogs, online forums and chats." The developers of Netlytic say it can be used to capture tweets, blog comments, forum postings and text messages; find and explore emerging themes of discussions among individuals within a data set; and build and visualize communication networks to discover and explore emerging social connections among individuals within online communities and is "ideally suited for analyzing online interactions within any large online communities." Interesting for the intersection with dangerous speech, Netlytic's creators say it can be used to "automatically discover what people within an online community are talking about, who is talking to whom, how often they are communicating, the nature of their relationships or interactions (are community members happy, friendly and supportive; or are they angry, hostile and disrespectful to each other) and relatively how strong their relationships are."

- **Orgnet**. Orgnet is much like Netlytic, but also links the reach and exposure of messages to action. Orgnet works on "visualizing, modeling, diagnosing and improving networks and structures in organizations, communities and ecosystems." Orgnet has used social network analysis to uncover a criminal conspiracy among slumlords, map the outbreak of a disease, chart the spread of HIV in a prison system, map linking patterns among blogs, expose business ties and money flows in corruption cases, and lobbying patterns in healthcare reform around key senators.

3. Sentiment analysis

Sentiment analysis – also called opinion mining – gauges the feelings and opinions of authors of online content. The technology combines natural language processing, computational linguistics, and text analytics to identify and extract subjective information in source materials to measure the positivity or negativity surrounding a topic. There are scores – if not hundreds – of companies in this field. Two examples:

- **Lexalytics.** Lexalytics claims to be one of the largest companies conducting text analytics and sentiment analysis for "horizon scanning and risk management applications as well as in the more traditional areas of reputation management." Their software, Salience Engine, is a multi-lingual text analysis engine that is currently used for business intelligence, social- media monitoring, reputation management, survey analysis, and customer satisfaction. The company claims to be able to measure how attitudes, feelings and opinions change over time (i.e. whether they are becoming more positive or negative).

- **Crimson Hexagon.** Crimson Hexagon, based in Boston, delivers "relevant data, information, and insights from social media" using a mix of social analysis technologies. Its core technology is a language classification and opinion analysis algorithm that measures the proportions of large, dynamic social media conversations and reports on who is participating in the conversation, including the social influence of authors. The company then can represent digital conversations and opinions visually to show the conversational reach of opinions as well as the relationships among authors.

D. SAMPLE PROJECT: VIOLENT EXTREMISM ONLINE.

The following is an example of a project directly related to dangerous speech that utilized many of the techniques and technologies describe above. We provide it as a way to show how a range of technologies and approaches can be utilized in a targeted campaign to detect and respond to dangerous speech without restricting free speech.

The US government retained the strategic communications consulting firm Creativa (formerly Constrat) to undertake a project to detect, analyze and to respond to violent extremism online. This profile was compiled from an interview with Creativa's CEO, Jason HInton.[1]

Creativa first catalogued the universe of radical Islamic extremist websites, blogs and online message forums focused on promoting violence, used data mining techniques to analyze the content and documented the volume of conversation on such sites and how much time users spent on the sites. The analysis also identified the highest profile, influential, and prolific "speakers" in the conversations and their spheres of influence. Using network analysis technology, Creativa mined the social media environment to explore how messages from these sites were disseminated and charted the networks of communication.

From this analysis, Creativa was able to describe the volume of extremist and inciting language, as well as break it down by topic. Using a statistical algorithm, Creativa then undertook a sentiment analysis and plotted the speech and the tone of opinions (by "speaker" and/or outlet) on a spectrum from moderate to militant. This provided a picture of the universe of the dangerous speech, the most influential speaker and the reach of the content. That was the detection and analysis phase.

The second part of the project was to implement a pro-active campaign to reach out to those involved in the conversation, with customized messages – threads in discussion forums, comments on blogs, re-tweets and posts -- that could counter inflammatory messages with positive ones. The focus of the strategy was to utilize moderate religious leaders as messengers – individuals with appropriate stature or credentials who could provide interpretations of religious law or information about why violent extremism was counter to Islam.

Because the project was classified, an evaluation of the success of the effort is not available. However, in assessing the campaign, Creativa noted that the following were the most crucial components:

- **Retaining personnel for the effort who had cultural and language capabilities.** In the research phase, this meant drawing from the fields of social scientists, social statisticians and sociology and psychology professionals. "Personnel was the single most important factor in developing the project," said Jason Hinton, the CEO of Creativa. "You can't merely be collecting data against queries. You need people with language and cultural background to bring out the cognitive component, to recognize characters and foreign words and phrases, and sarcasm. Software can't do that. People were the core of the project."

- **Developing a sound quantitative analysis (scoring speakers and outlets from moderate to extremist).** In order to figure out how to respond to

[1] Interview conducted by Jill Savitt, January 21, 2014. A similar project available online, Facebook Fatwa, which Creativa undertook fro the Foundation for Defense of Democracies, documents what radical Saudi Wahhabists preach to their followers about the United States and non-Muslims on social media sites.

negative posts, Creativa needed to create a scoring algorithm – and plot speech along a spectrum -- so outgoing messages could be customized to the particular sentiments of the speakers.

- **Output and the ability to visualize the information.** Because the campaign involved such massive volumes of data, Creativa needed to generate visual versions of the data, so it could be digested and analyzed and a campaign outreach strategy could be developed.