Analyzing Trends and Topics in Internet Governance and Cybersecurity Debates Found in Twelve Years of IGF Transcripts

Derrick L. Cogburn
American University
dcogburn@american.edu

Abstract

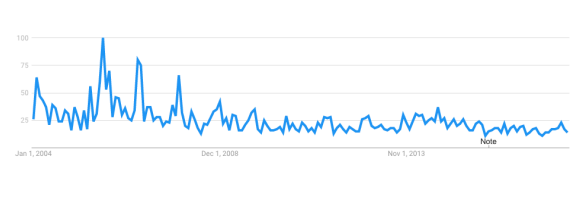
Internet Governance research generates substantial and innovative, interdisciplinary global scholarship. What are key topics and themes in this research area, and how do they relate to cybersecurity? This paper answers these questions by analyzing transcripts from twelve years of the UN Internet Governance Forum (IGF), asking: (1) What key themes, topics, and entities are discussed at IGF? (2) Which issues have remained consistent at IGF, and which have changed? And (3) to what extent is the NIST Cybersecurity Framework represented at IGF? Using the CRISP-DM approach to text mining, we find human rights as the most dominant IGF theme, followed by freedom of expression, with disability being a persistent issue. During entity extraction cybersecurity emerges prominently, as does blockchain and IoT. Topic Modeling illustrates the resilience of human rights, but also identifies the IANA transition, accessibility, and "fake news." Finally, the NIST cybersecurity framework is represented clearly in the data.

1. Introduction

Internet Governance has been an active and rich domain for interdisciplinary academic research since the late 1990s [1][2][3][4][5][6][7][8][9][10]. It has addressed a wide range of technical and policy challenges around the multistakeholder development and allocation of scarce resources related to the functioning and stability of the TCP/IP (Transmission Control Protocol/ Internet Protocol) that lies at the heart of the Internet and the policy implications.

As one example of the interest in this subject, Figure 1 below illustrates the Google Trends searches for Internet Governance since 2004 and indicates November 2005 as the height of popularity. Aside from this spike, searches for Internet Governance have remained steady. Interestingly, 2005 marked the final months leading up to the second phase of the World Summit on the Information Society (WSIS), held from 16-18 November 2005 in Tunis.

Figure 1. Google Trends Searches for Internet Governance, 2004-2017

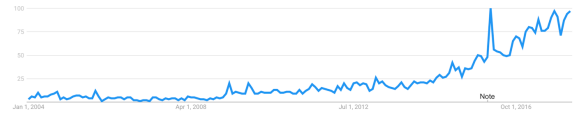


Debates and issues related to Internet Governance dominated the first phase of WSIS held in Geneva from 10-12 December 2003, along with development issues such as financing, infrastructure, and human capacity building [6][12]. At the conclusion of WSIS Tunis, the United Nations launched the Internet Governance Forum (IGF), giving it an initial five-year mandate, and renewing it for another ten years in 2010.

Internet Governance was certainly an important research topic before WSIS, but this global meeting helped to accelerate a broad, multistakeholder focus beyond the narrow technical and academic approach that had dominated the field since the mid 1990s. It spawned an interdisciplinary grouping of scholars known as GigaNet, the Global Internet Governance Academic Network [12][13].

During the same time period, 2004-2017, while Internet Governance searches held steady after WSIS, there was a noticeable increase in searches for the term "cybersecurity", with a sharp increase starting in 2013. Figure 2. Illustrates this trend.

Figure 2. Google Trends Searches for Cybersecurity, 2004-2017

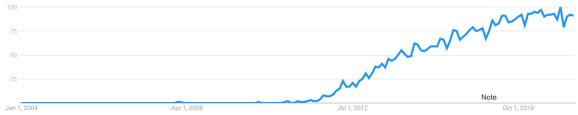


URI: https://hdl.handle.net/10125/59530
ISBN: 978-0-9981331-2-6

On one hand, this rise in searches for cybersecurity is correlated temporally with a steady increase in actual cybersecurity attacks. These attacks include the 2007 *TJX* apparel and home goods company hack, the 2010 *Stuxnet* attack against Iran's nuclear centrifuges, the 2013 *Target*, 2014 *Home Depot,* and 2015 *Office of Personnel Management* data breaches, and the 2013 Snowden revelations of US Government widespread surveillance of the Internet [11][12]. On the other hand, this increase also corresponded with the launch of the 2013 Obama Administration Executive Order on Cybersecurity and the subsequent launch of the 2014 NIST Framework for Improving Critical Infrastructure Cybersecurity[14]

During the same period, the terms "big data" and "analytics" have also become much more widespread. A number of factors are contributing to these developments, including the rise in processing power and speed, and the availability of large-scale data. Between 2004 and 2012, Google Trends indicates search for the term "Big Data" was flat, with an average popularity score of 5. But over the past four years, it has exploded, jumping to 15 in January 2012 to a high of 100 in March 2017. Figure 3. below illustrates this trend, which is even more exacerbated when adding the term analytics.

Figure 3. Google Trends Searches for Big Data Analytics 2004 - 2017



While the terms big data and analytics have grown in popularity, it is important to note there is not just one accepted definition of big data. There is in fact, a relative nature to the term, with its meaning changing based on the domain. For example, "big data" for earthquake engineering and upper atmospheric research is not the same for economics or many social sciences. The comment element of big data frequently uses the "3 Vs" model, referring to the Volume, Velocity, and Variety of data that is available today [15]. Some scholars add Veracity, Variability, and Value, as three additional "V" characteristics to consider when trying to understand the concept of big data.

Within this context, there is a particularly interesting type of data, that frequently goes underutilized. This data is unstructured textual data. Some sources estimate up to 70-80% of the world's available data is unstructured text [21]. This trend is accelerated by the growing digital production and digitization of text. Significant amounts of textual data are available for analysis, and growing, including: websites, blog posts, speeches, meeting transcripts, email archives, reports, published articles, and especially social media (e.g. Twitter, Facebook, RSS feeds). Take for example, "In the past 50 years, the New York Times produced 3 billion words" and "Twitter users produce 8 billion words – every single day" [16][17]. Each of these *genres* of data, have their own characteristics, that can be harnessed to augment analysis [22]. However, these very large-scale text-based datasets are not very large in terms of file size. One gigabyte of storage, an extremely small and common storage medium carried around by most students and faculty these days, can contain over 894,784 pages of plain text. A terabyte of data can contain 916,259,689 pages of plain text. It would take most humans an inordinately long time to read this level of data, but it could be easily carried around on the average thumb drive. This conundrum leads us to refer to the *relative nature* of big data. The big data of earthquake engineering and upper atmospheric research, is not at the scale of big data for most social scientists, especially those exploiting text as their data source.

This increasing availability of data is coupled with a corresponding increase in computational tools, storage, and big data analytics and text mining software, including both commercial and open source options. The combination of this infrastructure and the available data allows us to combine insights from both quantitative and qualitative big data, but especially from textual data sources.

2. Purpose

The purpose of this study is two-fold. First, we want to focus on substantive issues related to Internet Governance, and specifically to the IGF. We want to take an inductive approach to identifying core themes and key issues discussed over its twelve-year history, and understanding what issues have remained constant, which have changed, and when did they emerge/change. Then, we want to proceed deductively to understand the extent to which specific concepts, such as cybersecurity, are present in Internet Governance debates.

The second purpose of this study is to highlight the importance and potential impact of big data analytics and text mining as a technique and tool for Internet governance research in general, and cybersecurity research specifically.

3. Conceptual Framework for Text Mining

Text can contain substantial meaning and value to researchers. For decades, qualitative researchers have analyzed texts, doing deep and careful reading of relevant documents. There are two important dimensions to text: (1) semantics and; (2) syntax. *Syntax* is the "structure" of language, how individual words are composed to make well-formed sentences and paragraphs. Semantics refers to the meaning of words within their surrounding framework. As these qualitative research projects grew in size and complexity, Computer Assisted Qualitative Data Analysis Software (CAQDAS) was developed to help facilitate this process. While extremely helpful, these CAQDAS tools still require researchers to conduct close reading of all documents, adding codes to the text, developed *a priori* or while reading the documents.

The field of text mining is highly interdisciplinary, and encompasses multiple theoretical approaches and methods, with one common element – text as input information. This process has been aided by the widespread availability of machine-readable text. However, advances in the field of text mining, aided by concurrent increases in computational power and storage, have now accelerated the potential to uses these techniques across a range of fields. With these tools, we can take unstructured text and transform it into a structured numerical format, based on term frequencies, and subsequently apply standard data mining techniques. This approach allows us to finally unlock the vast amounts of valuable information locked away in texts.

There are many techniques available to exploit the power and potential of big data analytics and text mining in specific research projects, including text classification, text clustering, ontology and taxonomy creation, document summarization, and latent corpus analysis. In general, there are two philosophical approaches to text mining: (1) statistical and: (2) natural language processing. The statistical approach to text mining is based on the "bag of words" assumption. This approach assumes there is value in the words themselves and does not require the analysts to understand the syntax of the words. In contrast, the natural language processing (NLP) approach, focuses on first conducting part of speech (POS) tagging, and then pursues the analysis taking into consideration word and sentence structure. In this study, we take primarily a statistical approach, but we recognize the value of NLP, and use this structure to contribute to answering one of our research questions.

Within the statistical approach to text mining, there are two broad divisions – *Inductive* and *Deductive*, each with their own methodologies and techniques. Inductive techniques allow us to ask broad exploratory questions about a large-scale text-based dataset, without specific *a priori* goals. For example, we can ask what key words and phrases characterize a dataset, and determine what topics, themes, and trends exist. We can identify named entities within the dataset, including countries, people, organizations, and acronyms. For each of these elements, we can use cross-tabulation techniques to determine how these findings may change in relation to other key variables, such as date, region, organizational type, etc.

In contrast, deductive techniques are confirmatory, and allow us to ask specific research questions of the data and to even test hypotheses. We can build, adopt, or adapt dictionaries or categorization models to help us explore specific topics in the dataset, to determine the degree of their presence or absence [23][24]. Specific variants of these models allow us to conduct sentiment analysis, to characterize positive and negative sentiment or polarity within the dataset [25]. Further, we can use supervised machine learning to develop classification models that allow us to predict text with a high degree of accuracy [26]. Through the use of these inductive and deductive techniques, we can begin to illustrate the tremendous potential of computational text mining for Internet governance and cybersecurity research.

4. Case Study: Internet Governance Forum

The WSIS Action Lines, adopted at the end of the 2003 World Summit on the Information Society, included the continued development of the Internet with its potential impact on all aspects of the world [27][12]. The original WSIS Action Lines included four key references to Internet and Internet Governance, and the 2005 WSIS Tunis Agenda mentions the Internet 80 times and Internet Governance 30 times. After the conclusion of WSIS Tunis, participants adopted the Tunis Agenda, which in addition to focusing on the coordinated implementation of the WSIS Action Lines, included a commitment to establish and support the UN Internet Governance Forum (IGF). The IGF was given a five-year mandate and was subsequently approved for another 10 years.

The first IGF was held in Athens, Greece in 2006, immediately after the conclusion of WSIS 2005 in Tunisia. Twelve IGFs have now been held, the most recent in December 2017 in Geneva, and one planned for November 2018 in Paris. Annually, thousands of multistakeholder actors, including: government, private sector, civil society, international organizations; participate in each IGF. For example, at the Inaugural

IGF in Athens 2006, there were more than 1,200 multistakeholder participants. At the most recent 2017 IGF in Geneva there were 2,219 from 142 countries.

These participants engage in a multi-day program (usually four days) of plenary sessions and workshops organized in large part by the Multistakeholder Advisory Group (MAG) appointed to advise the United Nations Secretary General on the forum. The program and the MAG are supported by Dynamic Coalitions (DCs), which are multistakeholder thematic networks.

Due in large part to the advocacy of groups promoting accessibility to the IGF for persons with disabilities, all IGFs since its inception have included captioning for at least the main sessions. The IGF Secretariat has thankfully made these captioning transcripts available to the public. Initially these transcripts did not cover all sessions, but over time, the coverage has become more comprehensive [18]. For example, in 2006, there were only 11 transcripts made available to the public, while in 2017, 215 transcripts were made available. Table 1. below illustrates this growth in the availability of IGF transcripts.

Table 1. Availability of IGF Transcripts

Year and Venue	No. of Transcripts
IGF 2006 - Athens	11
IGF 2007 - Rio de Jeneiro	14
IGF 2008 - Hyderabad	14
IGF 2009 - Sharm El Sheikh	15
IGF 2010 - Vilnius	114
IGF 2011 - Nairobi	61
IGF 2012 - Baku	8
IGF 2013 - Bali	63
IGF 2014 - Istanbul	138
IGF 2015 - Joao Pessoa	162
IGF 2016 - Guadalajara	205
IGF 2017 – Geneva	215
Total	1,020

5. Research Questions

With this brief case study in mind, and as a demonstration of several text mining techniques, we ask three key research questions in this paper, two inductive questions and one deductive question.

RQ1. What are the key themes, topics, and entities discussed at IGF over its lifetime?
RQ2. Which key issues have remained consistent at IGF, and which ones have changed?
RQ3. To what extent is the 2014 NIST Cybersecurity Framework represented at IGF?

6. Methodology

For this chapter, we use a methodological approach called the Cross-Industry Standard Process for Data Mining (CRISP-DM) for Text Mining [28]. Since text mining is still a relatively new and somewhat unstandardized field, the CRISP-DM approach can provide a well-understood, documented, and somewhat standardized process for executing and managing complex text mining projects.

The CRISP-DM for text mining has six stages, through which each text mining project must proceed. In Stage 1, the researcher is focused on determining the purpose of the text mining study. Focusing on what the researcher wants to accomplish, and the problem or opportunity identified by the researcher.

Stage 2 the researcher explores the availability and nature of the unstructured textual data you would like to exploit. The researcher has to determine if the data is available, in what format it is stored, and in what quantity.

Stage 3 focuses on preparing your data, which could include the steps of data cleaning, pre-processing, applying stopwords or exclusion lists to remove words that are too common without sending any signal to the researcher as to the substance of the dataset, and further data reduction techniques of stemming and lemmatization.

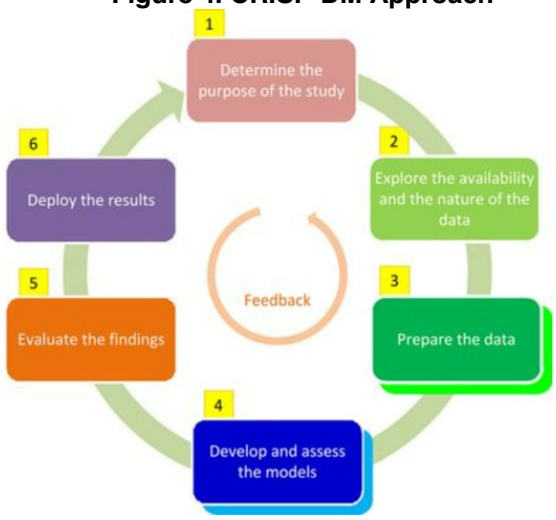
In Stage 4, the researcher develops the models and specific techniques they will use to analyze the data.

Stage 5 allows the researcher to evaluate the results of the analysis.

In Stage 6 the researcher focuses on deploying the results, in the form of recommendations and presentations.

At any point along the way, the researcher may decide to go back to a previous stage, or all the way to the beginning. Figure 4 below illustrates the six steps of the CRISP-DM methodology.

Figure 4. CRISP-DM Approach



Using an automated commercial tool called SiteSucker [30], we were able to collect all of the publicly available IGF transcripts. This data collection yielded 1,020 documents (made up of the following formats: .txt, .html, .doc, .pdf), in a file of 109.9 MB. The data collection represents the available transcripts of main sessions, and workshops where available.

Once the data was collected, we used a commercial software tool called the Provalis ProSuite to organize the project and conduct the text mining [19]. There are options for similar tools that are open source, such as R, using RStudio, and packages such as tm and Rvest, for most of this analysis. The first step is to build the corpus, which includes converting the textual data into numerical data, based on the word frequencies across documents. Upon import, we used the file structure (organized by date of the IGF), to automatically create a "Date" variable, which will allow us to filter the dataset by date and conduct a longitudinal analysis. We were also able to use this Date variable for cross-tabulation analysis. Then, we preprocessed the data, applying a typical English exclusion list, which removes words that appear frequently, but are deemed to send little "signal" related to the analysis, and as such may confuse the results (in general, words such as: a, and, the, etc.) The exclusion list may be modified to represent the content in the specific dataset (e.g. words deemed important to include in the analysis may be removed at any time, and new words added). We did not apply any stemming or lemmatization, which are techniques to pre-process a textual dataset, and reduce its overall size.

To answer the first two of our four research questions and explore the key topics, themes, and entities that have been important to the IGF, we used the commercial tool, Provalis ProSuite to establish the corpus [19]. Also, to identify changes over the twelve years of the IGF, we added a "Year" variable, so documents in the corpus could be separated by each individual IGF (other variables could be added, for example to compare workshops with plenaries, etc.).

We began with an inductive approach that focuses on term and document frequency, followed by phrase frequency. This "count-based evaluation" is one of the simplest approaches to text mining, similar to basic descriptive analysis of project variables in a statistical study. This approach is simple but is used frequently. Overall, a word or phrase (n gram) frequencies are quite popular in text mining. In this approach, if a word or phrase is used frequently in a dataset, with some important limitations discussed below, it is deemed to be important. In this analysis, we employ the Term Frequency by Inverse Document Frequency (TFxIDF) technique. TFxIDF is based on the basic idea that if a word appears frequently in a document, it is important; but if it appears in too many documents, it is less important. This is a common text mining heuristic to identify "important" words and phrases in a corpus.

Next, we used an inductive technique called "topic modeling", which essentially uses exploratory factor analysis on the underlying numerical representation of the corpus to identify "factors" which are interpreted as topics. However, unlike factor analysis, since the dataset is based on text, the software provides a textual suggestion of what the topic seems to represent. We employed topic modeling on the entire dataset, and separately for each of the twelve years. In addition, we used "Named Entity Extraction" to identify key organizations, countries, acronyms, and people across the entire dataset, and again for each year.

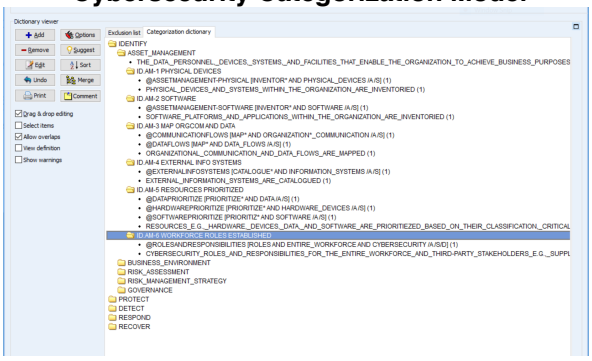
In order to answer our third and fourth research questions, we took a deductive approach. There are a number of deductive techniques we could use. A recent special issue of *IEEE Intelligent Systems* focused on Data Mining for Cybersecurity, and highlighted numerous interesting techniques [20]. To answer RQ 3 and understand how the Internet of Things was represented at IGF, we focused on hierarchical cluster analysis and categorization modeling (also known as dictionary development); and for RQ4 on the extent to which the 2014 NIST Cybersecurity Framework is included in IGF discussions, we also used categorization modeling.

Hierarchical cluster analysis allows us to examine the entire dataset for co-occurrences, and assess the themes or topics represented by specific clusters. Here, we are able to explore the cluster analysis to see if there is a cluster that appears to represent the Internet of Things.

Categorization modeling (known more generically as dictionary development), is an explicitly deductive technique [23][24]. Essentially, what dictionary development requires is to develop a semantic structure

that represents the concept one wants to explore within the dataset. It generally starts with the broadest categories within the concept (for example in the very popular use of dictionaries, sentiment analysis, these top categories tend to be the binary categories of "Positive" and "Negative"). Then, once those broadest categories are named, you may divide them further into broad sub-categories. Once the lowest levels of categories and sub-categories have been determined, specific words, phrases, and rules (which allow you to formulate criteria for inclusion of text which includes negations, and specifications for proximity of words and phrases) can be developed. When there is the occurrence of any of these elements, they accrue to the sub-category, which in turn aggregates up to the categories. This is a very powerful technique to identify to the extent to which a specific concept the researcher is interested in exploring is either present or absent in the dataset. In this study, we developed the categorization model from the 2014 NIST Cybersecurity Framework Core spreadsheet [14]. This framework has five primary categories (Identify, Protect, Detect, Respond, Recover) and within each category are multiple sub-categories, and sub-sub-categories. All of these elements are captured in our categorization model. We deployed these categorization models across the entire twelve-year period. Figure 5 below illustrates the categorization model.

Figure 5. Overview of 2014 NIST Cybersecurity Categorization Model



We could have similarly built another categorization model representing the EU Cybersecurity Framework and compared the degree to which each framework was represented in the dataset. Or, we could have explored the dataset to assess the degree to which the priorities of one stakeholder – say the Private Sector, represented by BASIS (Business Action in Support of the Information Society), supported by the International Chamber of Commerce (ICC) was represented in the dataset, relative to say the statements of the Civil Society Internet Governance Caucus (IGC). One final technique we could have employed would have been to use supervised machine learning to build a classifier to distinguish between the content of each stakeholder group, and then deployed that classifier to assess which stakeholder group had the most influence in the IGF processes. It would be a little tricky to do this in the IGF context, because there are no concrete "outcome documents" of each IGF, but this technique was used to great effect in an analysis we did of the various stakeholder contributions to the NetMundial conference (Cogburn, 2014).

7. Limitations

To our knowledge, our corpus of 1,020 IGF transcripts makes this is the largest study to-date of this important data. However, there are limitations to the study. First, even with the large number of transcripts made available in the later years, they do not cover all of the workshops and side-events associated with an IGF meeting. Further, this dataset does not differentiate between the main sessions and workshops. By adding a "session type" variable, we could further tease out any differences between the more "formal" main sessions and the more dynamic workshops.

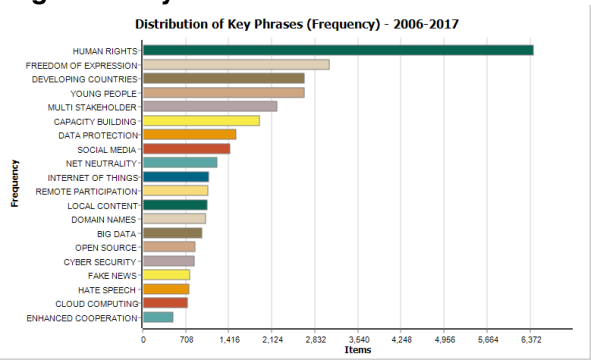
Also, of course, much of the work of the IGF is accomplished outside of the formal conference structure. As Goffman would say, this "backstage" behavior occurs during the coffee breaks, lunches, dinners, and the many receptions and parties associated with an IGF.

Finally, one of the great potential aspects of this dataset is being able to identify who (or at least from which sector) a specific contribution. Those elements are not coded in this dataset, so we are unable to conduct that kind of analysis at the moment.

8. Findings

To answer our first research question, "What are the key themes, topics, and entities discussed at WSIS over its lifetime?", we use the TFxIDF to first explore keywords frequency and then phrase frequency across all 12 years. Remember, TFxIDF is a recognized technique that helps us to identify those words and phrases that are deemed "important" in the dataset, and not just those that occur most frequently. Figure 6 below represents the top ten themes of 12 years of IGF represented by Phrases.

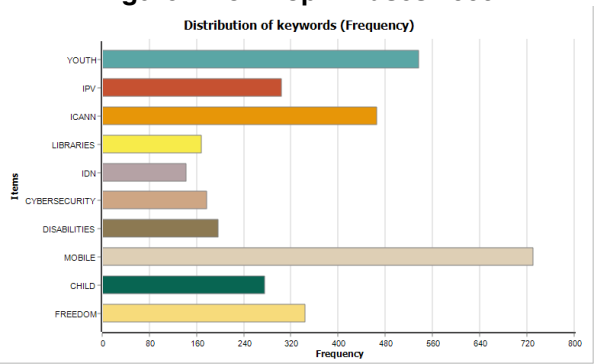
Figure 6. Key Phrases Across 12 Years of IGF



Here, we see "human rights" as the most frequently occurring phrase across all twelve years of the IGF, and by a substantial amount. Other, more liberal and less technical terms are seen occurring frequently, including "freedom of expression", "developing countries", "young people", "multistakeholder" and "capacity building." We have to go all the way to the 7th ranked phrase to find our find our first technical term, "data protection," followed by "social media", "net neutrality" and "Internet of Things."

Next, in order to answer our second research question, "Which key issues have remained consistent at IGF, and which ones have changed?", we explore the changes in key themes over the twelve years of IGF by identifying the top ten themes at the beginning (2006), middle (2011), and most recently (2017). Figures 7, 8, and 9 illustrate these beginning, middle, and recent top themes.

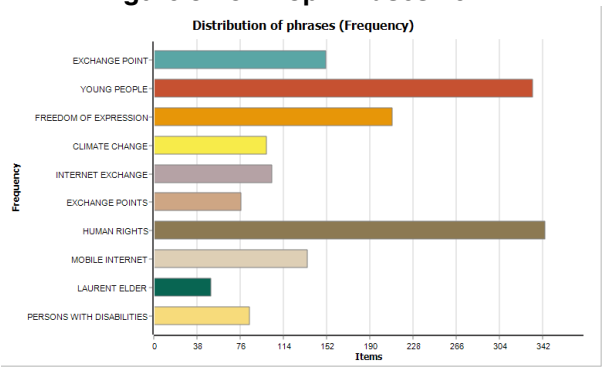
Figure 7. IGF Top Phrases 2006



Here, we see a mixture of technical and non-technical phases early in the IGF lifespan. At the top are "youth", "IPV" (Internet Protocol Version), "ICANN" (Internet Corporation for Assigned Names and Numbers), "libraries", "IDN" (Internationalized Domain Names), and "cybersecurity." We also see at this early stage "disabilities", which is both reflective of the advocacy efforts for persons with disabilities at IGF and of the earlier WSIS processes.

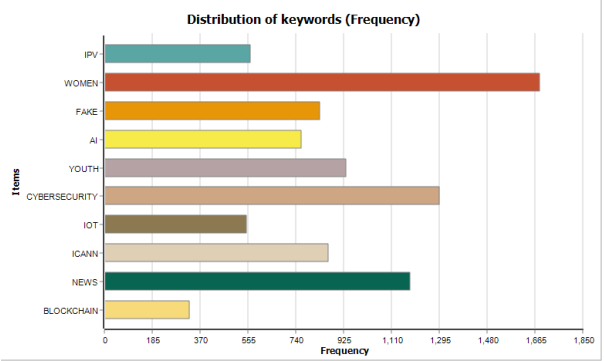
Then, towards the middle of the IGF lifespan, in 2011, we see a dominance of technical terms, such as "exchange point", and "mobile internet" but we also continue to see a strong focus on developmental issues, such as "young people", "freedom of expression", "climate change", "human rights", and again "persons with disabilities."

Figure 8. IGF Top Phrases 2011



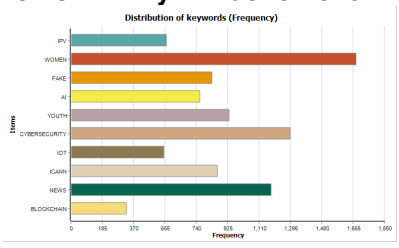
In the most recent IGF in 2017, we see many more technical terms at the top, including "cybersecurity", "AI" (Artificial Intelligence), "IOT" (Internet of Things), and "blockchain". With "women" and "youth" being the most development oriented phrases.

Figure 9. IGF Top Phrases 2017



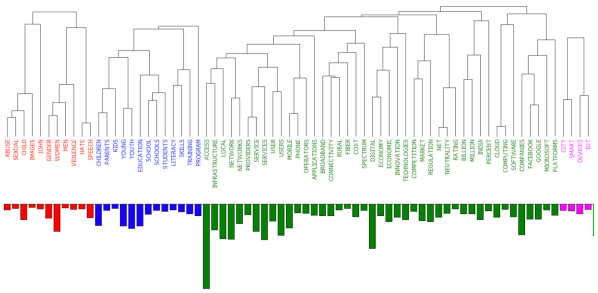
Also, Figure 10 below illustrates using the entity extraction tool, to identify the most frequently listed organizations, acronyms, countries, and people across all 12 years of IGF.

Figure 10. IGF Entity Extraction over 12 Years



Next, we conducted a hierarchical cluster analysis across all twelve years. There were initially 60 clusters identified, representing significant thematic groupings. Figure 11 below illustrates four of those clusters around: (1) child protection, (2) capacity building; (3) innovation in infrastructure (including broadband, mobile, net neutrality and cloud computing); and (4) smart cities and Internet of Things.

Figure 11. Partial Illustration of the Cluster Analysis (Highlighting Four Clusters)



When we use the inductive technique of topic modeling, looking across all twelve years of IGF, as well as looking at the middle and most recent IGFs we find the topic of Freedom of Expression and Human Rights to be the most durable and consistent topic across the IGF, with earlier topics of IDNs and mobile phones being take over in the most recent IGF with topics such as Fake News and Media Freedom and multistakeholder discussions. Table 2. Highlights this topic modeling across the IGFs.

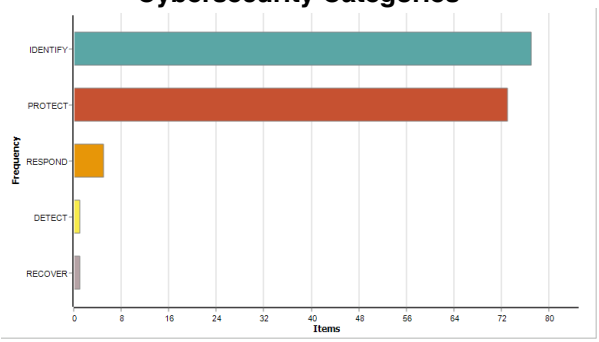
Table 2. Topic Modeling 12 Years of the IGF

2008	2009	2010	2011	2012
IPV4-6 Transition	DNS and IANA	Budapest Convention	Human Rights	Cybercrime
Child Pornography	A11y	Mobile Devices/ Wifi	IDNs	IDNs
Enhanced Cooperation		Human Rights	Mobile Phones	Disaster Risk
Freedom of Expression			A11y	
2013	**2014**	**2015**	**2016**	**2017**
Journalist/Bloggers	Human Rights	Human Rights	Human Rights	IANA Transition
Budapest Convention	IANA Transition	Net Neutrality	Wifi/Fiber	Cybersecurity
Mobile Devices	DNS	Child Abuse	Children Online	Fake News
Human Rights	CERTs/CSIRTs	IANA Transition	IXPs	Human Rights
Intellectual Property	A11y		SDGs	
IDNs			IANA Transition	

Finally, to answer our third research question, "To what extent is the NIST Cybersecurity framework represented at IGF?", we deployed a categorization model or "dictionary" that captured all the primary categories, sub-categories, and sub-sub categories of the NIST Framework. This framework is being used by numerous government agencies in the United States and many private sector organizations to structure their cybersecurity strategies. This study allowed us to see the degree to which the components of this framework were present over the 12-years of IGF debates.

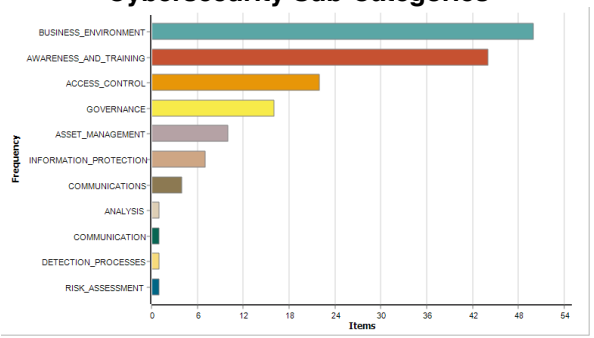
Figure 12. Below begins to illustrate that the framework is indeed present in the IGF. In particular, the first two major categories of the framework, "Identify" and "Protect." Less attention has been paid at IGF to the "Respond", "Detect", and "Recover" categories.

Figure 12. Distribution of NIST Cybersecurity Categories



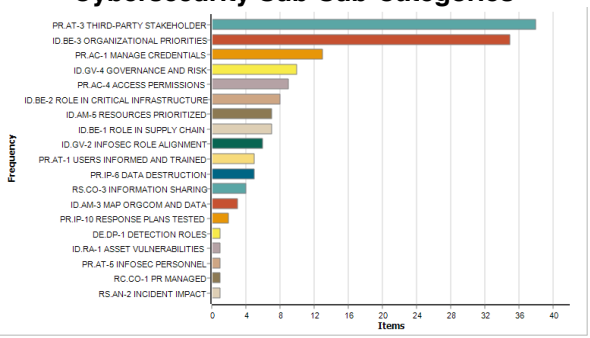
In terms of subcategories, Figure 13 illustrates how the "business environment" has received the most attention, with "awareness and training" coming a close second. Substantial attention has also been paid to "access control" issues; and "governance". Less attention has been paid to "asset management", "information protection", and "communications", with almost no discussion of "detection processes", and "risk assessment".

Figure 13. Distribution of NIST Cybersecurity Sub-Categories



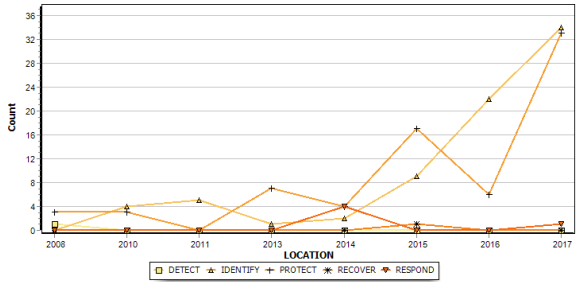
At the highest level of specificity, in the sub-sub-categories, we see in Figure 14 the primary focus on "third-party stakeholder" and "organizational priorities", with substantial attention on "managing credentials," "governance and risk and access permissions". IGF debates have paid less attention to key issues like "data destruction" and "information sharing", while not addressing other key cybersecurity issues like "detection roles", "asset vulnerabilities", or "infosec personnel".

Figure 14. Distribution of NIST Cybersecurity Sub-Sub-Categories



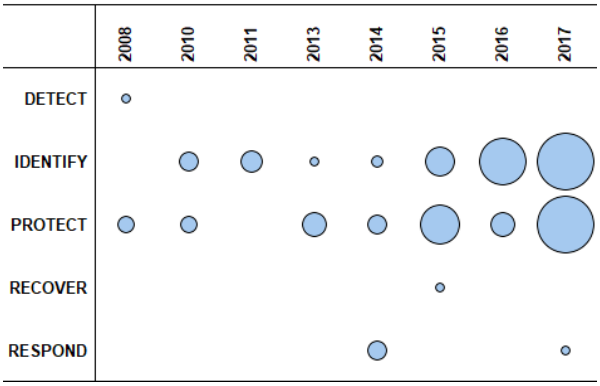
These figures show us the presence of specific elements of the NIST Cybersecurity Framework when looking at the entire twelve-year IGF dataset. However, we can see from Figure 15. below, the overall presence of the framework in IGF debates seems fairly pronounced, where in 2014, the year the framework was introduced, there was a substantial increase in two of the five components, namely "Identify" and "Recover".

Figure 15. Year-by-Year Line Chart Distribution of NIST Cybersecurity Framework



Another way of visualizing this data is via a bubble chart. Figure 16 below illustrates this distribution and correlation for these same two categories after 2014.

Figure 16. Year by Year Bubble Chart Distribution of NIST Cybersecurity Framework



9. Discussion

With this brief analysis, we have identified the key thematic focus areas of the Internet Governance Forum over its 12-year lifespan. One of the most surprising findings was that disability and accessibility issues have been included earlier and much more prominently than expected. Based on our topic modeling, it is clear this result is linked to the work of the Dynamic Coalition on Accessibility and Disability (DCAD), and its long-term coordinator, Andrea Saks.

10. Conclusions and Future Research

In this study, we believe we have accomplished both of our objectives. We have identified some interesting substantive components of the IGF, including the key thematic focus areas over its 12-year lifespan, and a year-by-year comparison. In addition, although we have only scratched the surface, we believe we have demonstrated the power of big data analytics and text mining in Internet governance and cybersecurity research. In this, and much of our other work, we have tried to highlight the importance and potential impact of these techniques as a tool for monitoring and evaluation of the SDGs and implementation of the WSIS Action Lines.

In terms of future research, we have already highlighted some of the possibilities we plan to pursue in the near term. Some of these will require adding more variables to the dataset, including type of session (e.g. main session, workshop); identifying which Dynamic Coalition organized the event; and finally, being able to identify the speaker, either by name and/or stakeholder grouping. However, before exploring some of these options, which will take some time, the nearest term studies will focus on building other categorization models, first to represent different approaches to cybersecurity in order to compare the degree to which each framework is represented in the dataset, and then to identify, represent, and compare other concepts, such as Net Neutrality and Internet Freedom. We also believe it will be very fruitful to explore the dataset to assess the degree to which the priorities of various stakeholders are represented.

Finally, once these various analyses are conducted, we plan to workshop the findings with other Internet Governance Scholars and participants. This will probably take the form of a paper submission to the GigaNet Annual Symposium, as well as perhaps a workshop proposal on big data analytics and text mining to monitor and evaluate the WSIS Action Lines and SDGs.

11. References

[1] Cogburn, D.L., Governing Global Information and Communication Policy: Emergent Regime Formation and the Impact on Africa, *Telecommunications Policy,* Vol. 27, Issue 1-2 pp 135 – 153, 2003.

[2] Milton Mueller, *Ruling the Root: Internet Governance and the Taming of Cyberspace*, MIT Press 2002.

[3] Daniel J. Paré, Internet *Governance in Transition: Who is the Master of This Domain?* Rowman & Littlefield Publishers 2003.

[4] Adam Thierer and Clyde Wayne Crews, Jr., eds., *Who Rules the Net? Internet Governance and Jurisdiction*, Cato Institute 2003.

[5] Cogburn, D.L., (2005) "Partners or Pawns? The Impact of Elite Decision-Making and Epistemic Communities in Global Information Policy on Developing Countries and Transnational Civil Society." *Knowledge, Technology & Politics.* Vol 18, # 2, Summer 2005, pp. 52-82.

[6] Cogburn, D.L.; Mueller, M.; McKnight, L.; Klein, H.; Mathiason, J. (2005). "The US Role in Global Internet Governance." *IEEE Communications Magazine*. December 2005, pp. 12-14.

[7] Jack Goldsmith and Tim Wu, *Who Controls the Internet? Illusions of a Borderless World*, Oxford University Press 2008.

[8] DeNardis, Laura, *Protocol Politics: the Globalization of Internet Governance*, MIT Press 2009.

[9] Lee A. Bygrave and Jon Bing, eds., *Internet Governance: Infrastructure and Institutions*, Oxford University Press 2009.

[10] Milton Mueller, *Networks and States: The Global Politics of Internet Governance*, MIT Press 2010.

[11] Musiani, F., Cogburn D.L., DeNardis, L., Levinson, N.S. (eds.) *The Turn to Infrastructure in Internet Governance*, Palgrave MacMillian, 2015.

[12] Cogburn, D.L., *Transnational Advocacy Networks in the Information Society: Partners or Pawns?*, Palgrave McMillian, 2017.

[13] Global Internet Governance Academic Network (GigaNet): https://www.giga-net.org.

[14] "Cybersecurity Framework Core (Excel)". NIST. This article incorporates text from this source, which is in the public domain. Found on the Internet at: https://www.nist.gov/cyberframework.

[15] Laney D. 3-d data management: controlling data volume, velocity and variety. META Group Research Note, 6 February 2001

[16] Leetaru, Kalev Mapping the global Twitter heartbeat: The geography of Twitter. First Monday, April 2013.: <http://firstmonday.org/ojs/index.php/fm/article/view/4366/3654>.

[17] S. H. Kaisler, J. A. Espinosa, F. Armour and W. H. Money, "Advanced Analytics -- Issues and Challenges in a Global Environment," *2014 47th Hawaii International Conference on System Sciences*, 2014, pp. 729-738.

[18] United Nations Internet Governance Forum: http://intgovforum.org/.

[19] Provalis Research: http://provalisresearch.com/.

[20] IEEE Intelligent Systems, 2018, vol. 33, Issue No. 02 - Mar./Apr.

[21] Schneider, Christie, "The Biggest Data Challenges You Might Not Even Know You Have." IBM AI for the Enterprise. Found on the Internet, 15 June 2018 at: https://www.ibm.com/blogs/watson/2016/05/biggest-data-challenges-might-not-even-know/.

[22] Lee, Y.-B. and S. H. Myaeng (2002). Text genre classification with genre-revealing and subject-revealing features. Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval. Tampere, Finland, ACM: 145-150.

[23] D.N. Bengston, and Z. Xu, "Changing national forest values : A content analysis", Retrieved from http://www.nrs.fs.fed.us/pubs/rp/rp_nc323.pdf. 1995.

[24] Deng, Qi; Hine, Michael; Ji, Shaobo; Sur, Sujit; Building an Environmental Sustainability Dictionary for the IT Industry. Proceedings of the 50th Hawaii International Conference on System Sciences.

[25] Liu B., Zhang L. (2012) A Survey of Opinion Mining and Sentiment Analysis. In: Aggarwal C., Zhai C. (eds) Mining Text Data. Springer, Boston, MA

[26] Rousu, J., et al. (2005). Learning hierarchical multi-category text classification models. Proceedings of the 22nd international conference on Machine learning. Bonn, Germany, ACM: 744-751.

[27] International Telecommunication Union, WSIS Action Lines, https://www.itu.int/net/wsis/stocktaking/help-action-lines.html.

[28] The CRISP-DM process model (1999), http://www.crisp-dm.org/.

[30] Site Sucker Application website: https://ricks-apps.com/osx/sitesucker/index.htm