CNS

OCCASIONAL PAPER

#27 · APRIL 2017

Searching For Illicit Dual Use Items in Online Marketplaces: A Semi-Automated Approach

Bryan Lee, Margaret Arno, and Daniel Salisbury

Middlebury Institute *of* International Studies at Monterey

James Martin Center for Nonproliferation Studies

This publication results from research prepared with the financial support of the United States Department of State, Bureau of Arms Control, Verification and Compliance (AVC), through the Key Verification Assets Fund. Additional funding for this project was provided by the Skoll Global Threats Fund.

Acknowledgements:
The authors would like to thank Stephanie Lieggi, Catherine Dill, Timothy Litts, Michael Duitsman, Yundi Luo, Bo Kim, Shea Cotton, Eduardo Fujii, and the data team at Drumwave, Inc. for their assistance in the research and preparation of this study.

James Martin Center For Nonproliferation Studies
www.nonproliferation.org
The James Martin Center for Nonproliferation Studies (CNS) strives to combat the spread of weapons of mass destruction by training the next generation of nonproliferation specialists and disseminating timely information and analysis. CNS at the Middlebury Institute of International Studies at Monterey is the largest nongovernmental organization in the United States devoted exclusively to research and training on nonproliferation issues.

Middlebury Institute of International Studies at Monterey
www.miis.edu
The Middlebury Institute of International Studies at Monterey, a graduate school of Middlebury College, provides international professional education in areas of critical importance to a rapidly changing global community, including international policy and management, translation and interpretation, language teaching, sustainable development, and nonproliferation. We prepare students from all over the world to make a meaningful impact in their chosen fields through degree programs characterized by immersive and collaborative learning and opportunities to acquire and apply practical professional skills. Our students are emerging leaders capable of bridging cultural, organizational, and language divides to produce sustainable, equitable solutions to a variety of global challenges.

James Martin Center for Nonproliferation Studies
Middlebury Institute of International Studies at Monterey
460 Pierce St., Monterey, CA 93940, U.S.A.
Tel: +1 (831) 647-4154
Fax: +1 (831) 647-3519

Table of Contents

1. Introduction

Purpose of research

Online business today takes place in a global marketplace. Drop shipments and containerized shipping have made the export and import of materials and finished goods easier than ever. In fact, the US Department of Commerce estimates that 22 percent of all wholesale trade was conducted online in 2014.[1] Anyone who has ever purchased something online is already familiar with this world from the consumer perspective. However, online marketplaces are increasingly important for business-to-business commerce as well. Suppliers of many high-tech products use online platforms for both domestic and international transfers, including products that are considered dual-use. Dual-use commodities are items which are used for civilian industry but also have potential application in the development of weapons of mass destruction. As the popularity of online marketplaces increases, so does the risk that illicit trafficking networks can misuse these platforms as a means of gaining access to critical dual-use commodities.

Well-known US-based platforms like eBay, Amazon, and Craigslist are thought to have reasonably effective procedures to screen for dual-use materials that might be controlled.[2] However, even these firms have been challenged by the upsurge in use of their platforms for trade in potentially sensitive items. For non-US based platforms, reports highlighting how some sites have been used for illicit trafficking may have brought increased attention to the issue, but it remains unclear how effective screening methodologies are.[3] Further, it is unclear if smaller sites that specialize in business-to-business transfers of industrial materials understand or implement controls on the sale of dual-use technologies.

Studies that have examined the challenge of controlling online marketplaces have been based on individual cases or small groups of items found in manual searches. The data and information provided by these cases have been useful, but they provide at most a snapshot of the problem. The lack of more comprehensive data also limits the ability of the nonproliferation community to analyze the extent of the problem and make useful recommendations to combat the spread of sensitive materials.

In order to build a more comprehensive set of product listing data, a variety of open-source software tools were used to automate search processes. Using automation, over 30,000 individual product listings were collected from several non-US based trading sites. These data were then processed and cleaned for use in exploratory data analysis and visualization.

---

[1] US Department of Commerce, Census Bureau, "E-Stats 2014: Measuring the Electronic Economy," June 7, 2016, <http://www.census.gov/content/dam/Census/library/publications/2016/econ/e14-estats.pdf>.

[2] "Amazon.com Help: Trade Control Laws," <https://www.amazon.com/gp/help/customer/display.html?ie=UTF8&nodeId=201575780>, accessed January 2, 2017.

[3] Nick Gillard and Daniel Salisbury, "The Obscure Chinese Businessman Accused of Selling Ballistic Missile Parts to Iran," Vice, July 23, 2015, <https://www.vice.com/en_uk/article/the-boring-chinese-businessman-accused-of-selling-ballistic-missile-parts-to-iran-722>.

Five project phases

This project was made up of five phases. Phase one consisted of determining the information to be collected as well as the recommended target sites. After consultation with subject matter experts in the export compliance field, five products were chosen that would be used as search terms: "autoclave"; "maraging steel"; "vacuum pump"; "carbon fiber"; and "spark gap." These items were selected because of their ubiquity and lack of arcane technical specifications.

It was clear very early on that it was best to take a naïve approach to product searching. In other words, each site would likely have different available information for each product, and a set schema should not restrict the search. For this reason, broad keyword searches were used based on common item names as opposed to detailed specifications or manufacturer names.

Phase two consisted of retrieving data, and it was in this phase that a variety of open source software tools were used. In particular, the Python programming language was heavily relied on for data collection and management.

Phase three consisted of storing and collating product listing information. This was done using the non-structured query language (SQL) database software MongoDB (MongoDB, Inc., http://www.mongodb.com). Mongo is scalable, uses a standard Java Script Object Notation (JSON) format for data storage, and readily converts data to other formats for visualization and analysis. Additionally, it has a native query language that is relatively easy to use for basic tasks like identifying duplicates or counting items. This phase also included data cleaning and a first attempt to create a standard format for analysis.

All exploratory projects should include a refinement phase—phase four. In this phase we revised search terms, identified and collected additional data such as sanctions lists, and attempted to fill holes in the data left over from the first collection effort.
In phase five, we used Drumwave, Inc's. Dynamic Data Visualization (DDV™) tool for data exploration and analysis. We used the visualization as a springboard for further analysis of the listing data and to generate our conclusions and recommendations.

The remainder of this paper is organized as follows: In section two, other published work related to this topic is discussed. Section three provides the method used to choose sites and items, as well as tools and techniques used to gather and analyze the data. Section four details findings, and section five discusses them. Finally, section six offers conclusions and recommendations for future data analysis efforts.

2. Related Work

The possiblity of dual-use items being traded online is not a recent concern. In 2008, the US Government Accountability Office (GAO) testified before Congress that undercover agents were able to purchase sensitive military equipment through popular e-commerce sites eBay and Craigslist.[4] In 2009, GAO investigators were able to purchase sensitive dual-use nuclear technology directly from manufacturers' web sites. Again, the investigation was undercover and included establishing fictitious front companies and using false identities.[5]

There have been suprisingly few attempts to look at the online trade of dual-use items more broadly, and those have largely focused on the policy implications of potential trade. A 2014 US National Academy of Sciences (NAS) report discussed the issue of the online trade of chemical warfare-related equipment, calling it a new challenge to non-proliferation activities.[6] Grant Christopher described methods to identify Business-to-Business (B2B) websites that could be relevant for nuclear safeguards, generated a list of such businesses, and reported finding many items for sale that could be of significance.[7] Christopher used the Nuclear Suppliers Group's dual-use item list as a guideline and searched manually using select keywords.

Ken Coleman and Raymond Zilinskas surveyed the online availability of botulinum toxin in cosmetic preparations,[8] and Zilinskas and Phillipe Mauger followed up with a study exploring the online trade of biotechnology focusing on Australia Group-controlled items.[9] They were able to identify a number of potentially dual-use items, but their searches were restricted to manual queries of search terms. Since there were no documented procurement cases in the literature, they describe how Iraq's former biological weapons acquisition network functioned and how the availibility of these items online would have made the network more efficient.

---

[4] US Government Accountability Office, Internet Sales: Undercover Purchases on eBay and Craigslist Reveal a Market for Sensitive and Stolen U.S. Military Items GAO-08-644T. Washington, DC, 2008 <http://www.gao.gov/assets/120/119629.pdf>.

[5] US Government Accountability Office, Military and Dual-Use Technology. Covert Testing Shows Continuing Vulnerabilities of Domestic Sales for Illegal Export, GAO-09-725T. Washington, DC, 2009 <http://www.gao.gov/assets/130/122654.pdf>

[6] Kathryn Hughes et al., The Global Movement and Tracking of Chemical Manufacturing Equipment: A Workshop Summary (Washington, DC: The National Academies Press, 2014), <https://doi.org/10.17226/18820>.

[7] Grant Christopher, "Open Source Information in Support of Safeguards," paper delivered at the Symposium on International Safeguards, Vienna, Austria, October 20-24, 2014, <https://www.iaea.org/safeguards/symposium/2014/home/eproceedings/sg2014-papers/000191.pdf>.

[8] Ken Coleman and Raymond A. Zilinskas, "Fake Botox, Real Threat," Scientific American 302, no. 6 (2010), pp. 84–89.

[9] Raymond Zilinskas and Philippe Mauger, "Biotechnology E-commerce: A Disruptive Challenge to Biological Arms Control," Occasional Paper 21 (Monterey, CA: James Martin Center for Nonproliferation Studies, 2015), <http://www.nonproliferation.org/wp-content/uploads/2015/06/biotech_ecommerce.pdf>.

The most extensive research into online nuclear-related dual-use trade has been done by Project Alpha of King's College London. Research into missile proliferation found a sanctioned missile-related entity listed as a vendor on no fewer than fifteen internet trading platforms, and five sanctioned Iranian companies were discovered advertising on one of the world's largest B2B websites, Alibaba.[10] In addition to sanctioned vendors, items advertised for sale on Alibaba included "metals suitable for centrifuge manufacturing, gauges and pumps for centrifuge cascades for uranium enrichment, metallurgical casting equipment suitable for making nuclear weapon 'pits,' and high-speed cameras suitable for use in nuclear weapon diagnostic tests."[11] Companies also had listings for uranium and gallium.[12]

Project Alpha research emphasized export and regulatory compliance, especially with respect to China, and therefore focused on Chinese company Alibaba and its subsidiary, 1688. Besides eBay, the research does not cite other specific sites of concern. With eBay, the emphasis was on China's export compliance infrastructure, and a subsequent purchase of a pressure transducer through an eBay listing was made from a Chinese company.[13]

3. Methodology

Determining Product Search Terms

The project focus was on dual-use nuclear technologies listed on the Nuclear Suppliers Group trigger list. After reviewing the list, we selected five initial product types that were thought to be broadly available online. We then used a previously-generated correlation table to link nuclear supplier commodities to standard industry commodity codes. Using the tabulated item specifications and technical descriptions (Annex 1), we conducted several manual test searches on various B2B websites. We then used the results to modify the technical language until we were satisfied that the search term was broad enough to return a reasonably large list of items. This process resulted in five keyword search terms: "autoclave"; "maraging steel"; "vacuum pump"; "carbon fiber"; and "spark gap."

After establishing the final keyword search terms, we tested them again using a variety of misspellings and alternate spellings, e.g. "vaccum pump" or "carbon fibre," but did not see significant differences in the numbers of listings retrieved. This is likely the result of the B2B sites' search algorithms

---

[10] "The Great Takedown: Alibaba.com de-Lists Factory Owned by Notorious Missile Proliferator, Karl Lee," Project Alpha, November 29, 2014, <https://projectalpha.eu/the-great-takedown-alibaba-com-de-lists-factory-owned-by-notorious-missile-proliferator-karl-lee/>.

[11] Ian Stewart, Nick Gillard, and Jonathan Brewer, "Internet-Trading Platforms: Making It Easier to Get around Sanctions?," Bulletin of the Atomic Scientists, October 31, 2014, <http://thebulletin.org/internet-trading-platforms-making-it-easier-get-around-sanctions7772>.

[12] Charles Clover, "Weapons of Mass Ecommerce: THE BIG READ: NUCLEAR PROLIFERATION," Financial Times, September 27, 2014, <https://www.ft.com/content/2a19e07c-43ef-11e4-8abd-00144feabdc0>.

[13] "Procurement of Pressure Transducer via Ebay From China," Project Alpha, September 24, 2014, <https://projectalpha.eu/procurement-of-pressure-transducer-via-ebay-from-china/>.

automatically accounting for alternate spellings. We also translated the keywords into both Korean and Russian and ran the searches on host-nation B2B sites in those languages. Surprisingly, item listings were overwhelmingly described and cataloged in English. Moreover, the retrieved results were similar to those of the English language sites. For this reason, we dropped any further effort to search using Korean or Russian language keywords.

Site selection

This research focuses on non-US based online marketplaces (Business to Business (B2B) and Business to Consumer (B2C)) that have significant amounts of high-tech commodities and industry equipment. Based on the list devised by Grant Christopher (2014) and further consultation with experts, twelve B2B and B2C sites were chosen for evaluation (Annex 2). After reviewing site terms and conditions (see *Terms of service considerations*) we selected Vorras, Like123, TradeKey, TradeIndia, ECPlaza, and EC21 for further analysis. We also gathered structured and formatted data from eBay to have a baseline set of product listings.

Terms of service considerations

We evaluated each site's terms and conditions of service to confirm automated data gathering was not prohibited. Originally, twelve sites were considered as information gathering targets. However, after reviewing all terms and conditions, only five did not prohibit automated data collection and storage. Two sites required written consent, and the remaining five had non-permissive terms of service.

The five sites with permissive terms of service were Vorras, TradeIndia, ECPlaza, EC21, and Like123. TradeKey was semi-permissive in that it prohibits data gathering for commercial purposes. eBay and Alibaba both required written consent from the company. When contacted, eBay representatives pointed us to a third-party data vendor, Terapeak. We subsequently purchased eBay listing data through this approved vendor. Alibaba could not provide an approved data vendor, and its terms of service restrict data gathering from the website. For this reason, we did not gather any data from Alibaba. Likewise, the terms of service agreements from DGMarket, MarginUp, ThomasNet and TradersCity did not have permissive terms of service and did not allow data collection or compilation into a database. Table 1 summarizes which sites did and did not allow automated data collection.

Table 1 Permissive and non-permissive terms of service

Sites with Permissive Terms of Service		Sites with Non-Permissive Terms of Service	
EC 21	ECPlaza	Alibaba	ThomasNet
Like123	TradeKey	DGMarket	TradersCity
TradeIndia	Vorras	MarginUp	
eBay (provided data through licensed vendor)			

Automated Data Collector Design and Creation

Terapeak provided eBay product listing data in a structured, comma separated value file format. Automated data collectors had to be developed individually for each of the other six sites, as each has a different website structure for product listings. We chose the Python programming language (Python Software Foundation, https://www.python.org/) to develop the collectors because it has a well-developed suite of packages designed to aid data collection from websites, and it is also compatible with several different database formats. The data collectors used the Beautiful Soup package to parse website HTML and XML, while the packages Urllib and Requests were used to work with URLs. Following the initial development, we used the Scrapy package as a framework for rapid collector development.

Information Fields

We attempted to gather as much information as possible from each listing. Specifically, we searched for company information, product description, date posted, product title, keyword, price, product origin, and URL. Company information collected included company name, location, telephone, fax, email contact, website, and contact person. The information available differed among sites. For instance, sites that organized posts by categories had no date posted. Furthermore, information varied between posts on the same site, as vendors are not required to fill out all information fields before listing an item.

Database Creation

The variation in collected information implied a need for a non-structured database. We chose to use MongoDB, a non-structured query language database, to hold all information collected. We selected MongoDB as the document database for four reasons. First, its schema-less design allows for storing different types of information from different websites in a common database. Second, MongoDB allows for data validation implementation, which permits certain elements of the data to be checked for validity even though the database itself does not have a set schema. Third, MongoDB uses Java Script Object Notation (JSON) to store documents and easily converts data to other formats for visualization and analysis. Finally, MongoDB readily interfaces with Python through the Pymongo package.

Data Workflow

Once coded, the data collectors were deployed on a cloud-based host and allowed to run until reaching the end of the item listings for each category. The raw data gathered were then manually reviewed for accuracy before being added to the final data store. Upon completion of the collection runs for all six sites, we exported the data in JSON format. Data for the seventh site, eBay, were already formatted as comma separated value files from the data vendor. We then consolidated the data from all seven sites and sent them to our data analytics partner, Drumwave, Inc., to develop a Dynamic Data Visualization™.

4. Results & Analysis

Running collection scripts for each of the targeted websites to search for product listings using the associated keywords generated site-specific data files. A descriptive summary of the raw data generated by the automated collection (i.e. not including eBay listings) is shown in Figure 1 below. The discrepancy between product listings and total listings is due to duplicates and vendors entering non-product information (e.g. company name) into the product description field. Although structured and pre-formatted, the eBay data purchased from Terapeak still contained many duplicates. After filtering, there were 43,370 autoclave listings, 187,971 listings for carbon fiber, 368 for maraging steel, 2,411 for spark gap, and 162,480 for vacuum pump. After the final collection runs, we collated and aggregated the data from all of the sites for cleaning and further analysis.

Figure 1 summary of automated collection

Total documents in database: 71,172	Total number of listings per website:
Total number of products: 28,195	EC21: 1,684
Total number of countries: 69	ECPlaza: 3,609
Total number of companies: 9,690	Like123: 7,918
Total number of product listings per keyword:	TradeKey: 5,066
Autoclave: 2,438	TradeIndia: 977
Carbon fiber: 18,717	Vorras: 9,530
Maraging steel: 6	
Spark gap: 41	
Vacuum pump: 2,113	

Data Analysis

Data Overview and Exploratory Analysis

The immediate finding after reviewing the data is that the broad search terms provided too many irrelevant findings. For example, the search term "carbon fiber" retrieved not only listings of raw carbon fiber, but also numerous consumer products made with carbon fiber, such as fishing poles and bicycle frames. This was likewise the case with maraging steel. In fact, high-quality golf clubs turned out to be a project *bête noire*, as listings frequently described "carbon fiber" shafts and "maraging steel" club heads.

The irrelevant data problem is unlikely to be solved easily as product descriptive information was often not helpful in determining if the product was or was not worth further analysis. For example, consider the description for a direct-coupled clear water centrifugal pump from Like123:

> Features: 1) Applications: a) Domestic civil agricultural and industrial clean water supply b) Water circulation systems and water pressure boosting systems

c) Pressurized water using pressure vessels (autoclaves) irrigation dairy and farm application d) To empty and fill cisterns wells e) Gardening f) Car wash g) Fountains 2) Pump body: cast iron 3) Impeller: bass 4) Motor: closed externally ventilated 5) Insulation class: B 6) Protection class: IP44 7) Duty: continuously rated

Such item descriptions led to the development of a list of more precise terms:

- For vacuum pumps: Turbomolecular pumps; Holweck pumps; molecular pump
- For maraging steel: High tensile strength (> 1.95 GPa); High performance
- For autoclave: Designed to produce materials resistant to corrosion
- For spark gap: Containing three electrodes; Gas krypton tubes; Vacuum spryton tubes
- For carbon fiber: High tensile strength; High performance

The adjusted terms were added to the keywords and the collectors were run again. The change in search terms, however, turned up only a handful of additional listings.

The second immediate conclusion we can draw is that the data are very sparse. After manually reviewing many of the entries in our database as well as on the websites, we determined that blank fields are very common in online product listings and not an artifact of the data gathering process. The visualization in Figure 2 below shows the data fields on the left and the website names across the top. Fields that had information are depicted in blue, while empty fields are shown in black.

Figure 2: drumwave visualization

Final dataset

After removing duplicates and non-product listings, the final data set contained partial product listing data on 34,369 items. Products retrieved with the "carbon fiber" keyword were the most numerous, with 12,816 listings. "Vacuum pump" was the second most numerous listing at 10,705 items. As to markets, the vendor-provided eBay data account for almost two-thirds of the total listings. This was principally due to the data being pre-cleaned and therefore most likely to survive the secondary cleaning process to remove duplicates and incomplete listings. Figure 3 below summarizes the dataset.

Figure 3: Drumwave data summary

Types of analysis

A data analytics company, Drumwave, used the raw data collected as input for its Dynamic Data Visualization (DDV™) software. After reviewing and cleaning the data, Drumwave uploaded the data to its software platform for analysis and visualization. The DDV™ software provides tools for both broad and narrow analysis. Broad analysis tools give a dimensional perspective of the problems at hand, e.g. data gaps. Four types of narrow analyses tools were given, including a multiple category comparison tool, a treemap tool to explore data hierarchies, a geographic mapping tool, and a scatterplot tool. Moreover, the software allowed for rapid generation of wordclouds and also collated and sorted product images. Additional capabilities are described in the analysis sections below.

Matrix and Treemap overview of data and data fields

As shown previously in Figure 2, the matrix tool was used to gain a better understanding of the data structure. It demonstrates both the incompleteness of many of the descriptor fields as well as the wide variety of information provided by each site. Product specifications, for example, were almost never included in the item listings, making it very difficult to determine whether a given item was dual-

use capable without a manual review. Product descriptions and product names were also frequently lacking. This was especially the case for the keywords "maraging steel" and "spark gap," implying those items would likely be missed in a routine automated scan.

Each site differed greatly in the listing information available. The pre-formatted eBay data, for example, omitted company contact information. This was likewise the case with the raw data we collected from ECPlaza. Many of the targeted online trading markets do not provide detailed company information. This means automated methods of comparing extracted company information to existing databases such as sanctions lists must be able to deal with gaps. However, some do, and the information they provide may be quite extensive. Like123, TradeIndia, TradeKey, and Vorras all provided extensive company information, often including company addresses, telephone numbers, websites, and names and email addresses of sales representatives.

Some of the websites, most notably eBay, provide product listing categories along with the item data. Not surprisingly, "No Category" is the most frequent category in the data set while "Business & Industrial" also figures prominently. The large groupings of "Sporting Goods," "Cell Phones & Accessories," and "Toys & Hobbies" reflect the astonishing variety of items made from or containing carbon fiber. "Home & Garden" and "Health & Beauty" categories are mostly due to the widespread use of vacuum pumps for items such as garden fountains or non-surgical lip augmentation. Figure 4 depicts the item categories and their rough proportions in the data set, while Figure 5 provides a detail of the "Business & Industrial" category.

Figure 4: drumwave treemap category visualization

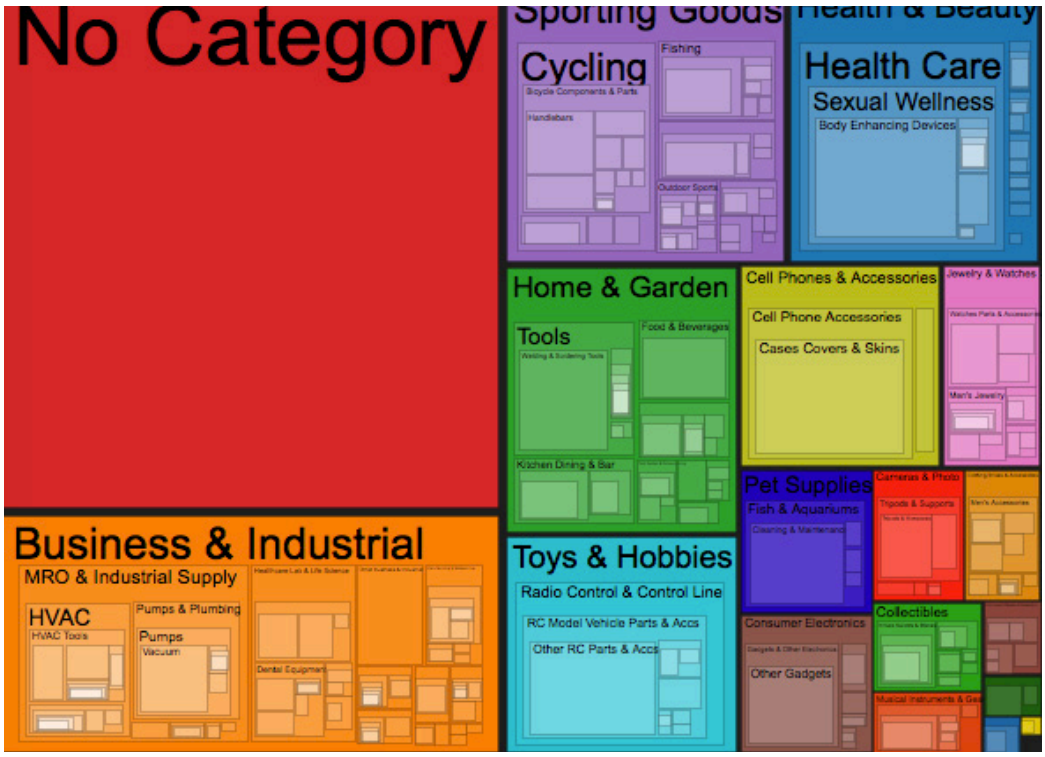
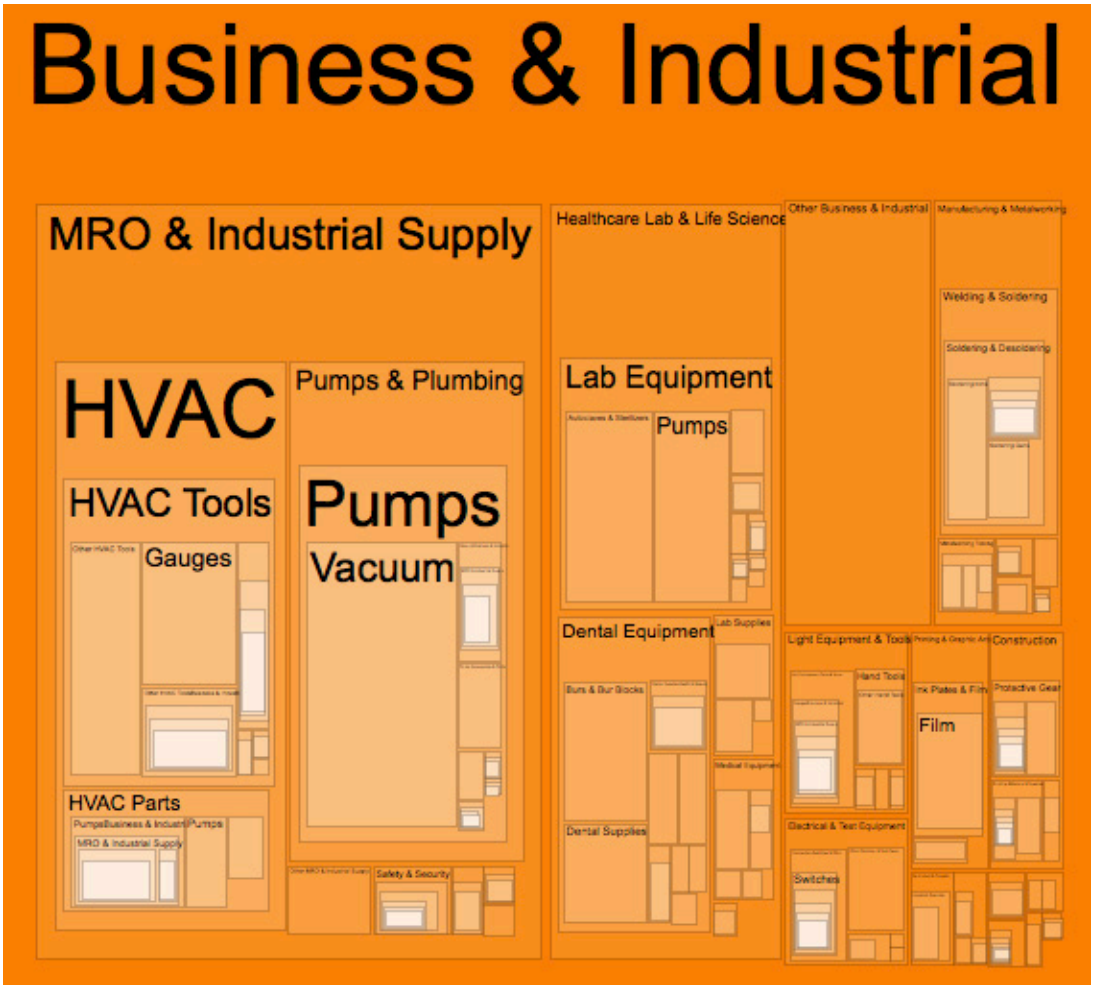


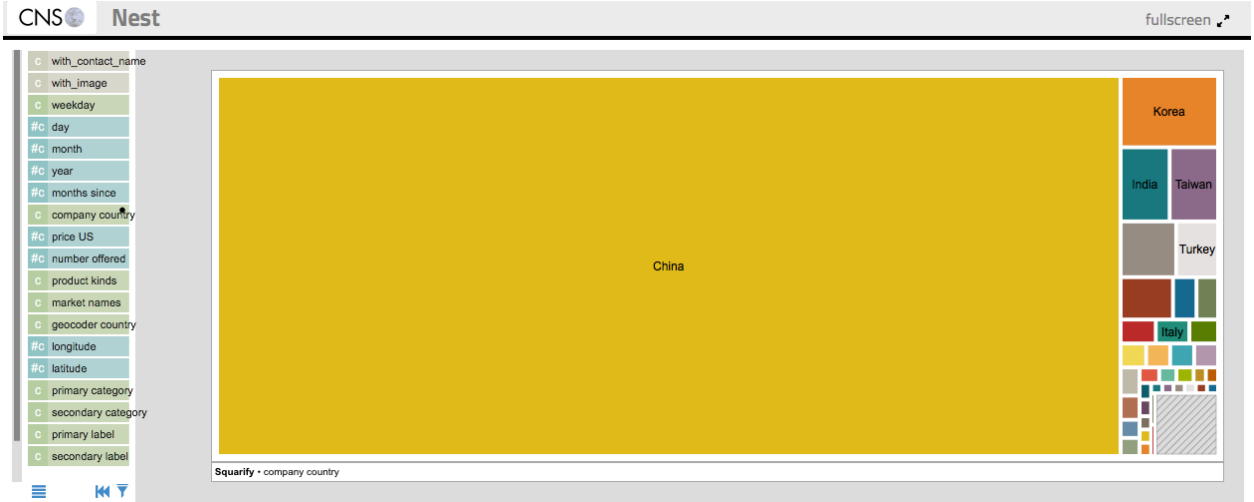
Figure 5: drumwave category detail



Geographic Data

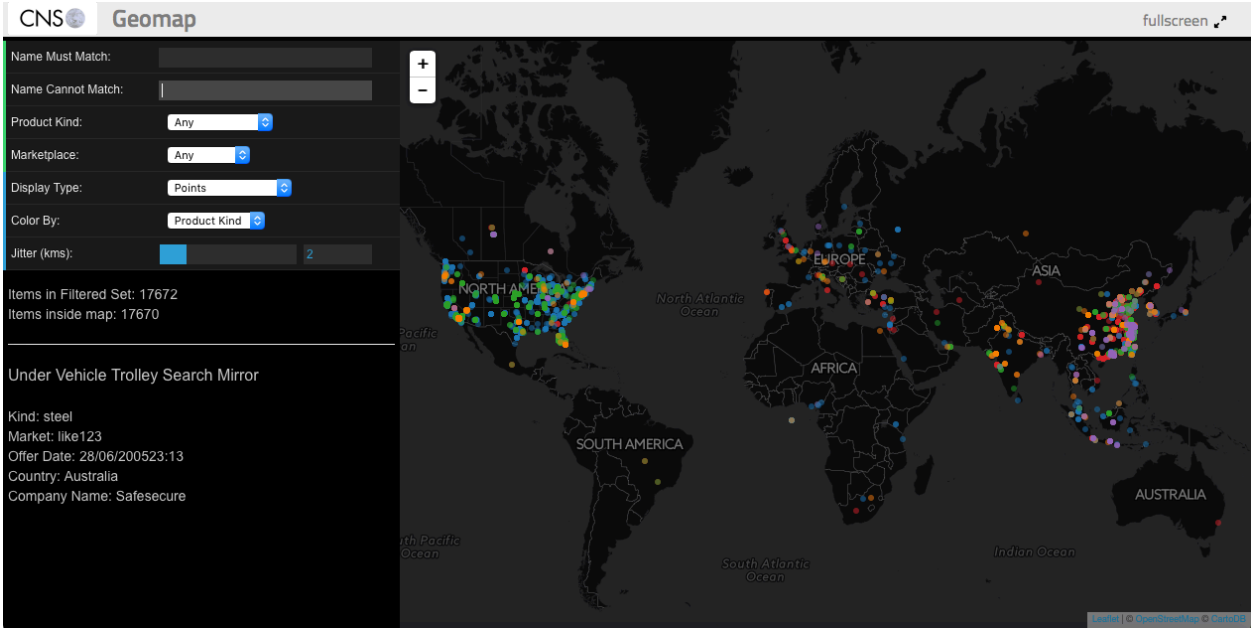
We applied two methods to determine the origin of manufactured goods. When the listing provided a company address for the manufacturer, we used that information. If a full address was not provided, we used an automated geocoder to attempt to derive a location based on the manufacturing city or country. Again, in the majority of instances, "null" was the derived location. For those listings that had adequate geographical information, our analysis confirms the intuition that China dominates the manufacturing space (Figure 6).

Figure 6: drumwave visualization of product listings by country



A more complete picture is shown when the product listings are mapped against their derived geographical coordinates. Figure 7 shows product listings by type and by country. With the exception of South Africa, the listings coded to Africa represent a failure of the geocoder-automated lookup to resolve the manufacturer location. Regulators should perhaps take comfort in the large number of products manufactured in the United States or Western Europe, but the map clearly shows that the countries of South and Southeast Asia are equally active.

Figure 7: drumwave geomap



A more refined view is provided by the category comparison tool, which can depict an "alluvial" chart (Figure 8) showing manufacturing countries and their associated online markets. This visualization shows China dominates the listings on Like123 and EC21 and that Vorras, like eBay, tends not to provide manufacturing country information. Figure 9 shows China's dominance in certain product categories, notably carbon fiber and steel, though it should be pointed out that the software does not discriminate between maraging steel and ordinary steel for this visualization. The figure also shows the overall rarity of spark gap listings in the dataset.

Figure 8: drumwave visualization of country by marketplace

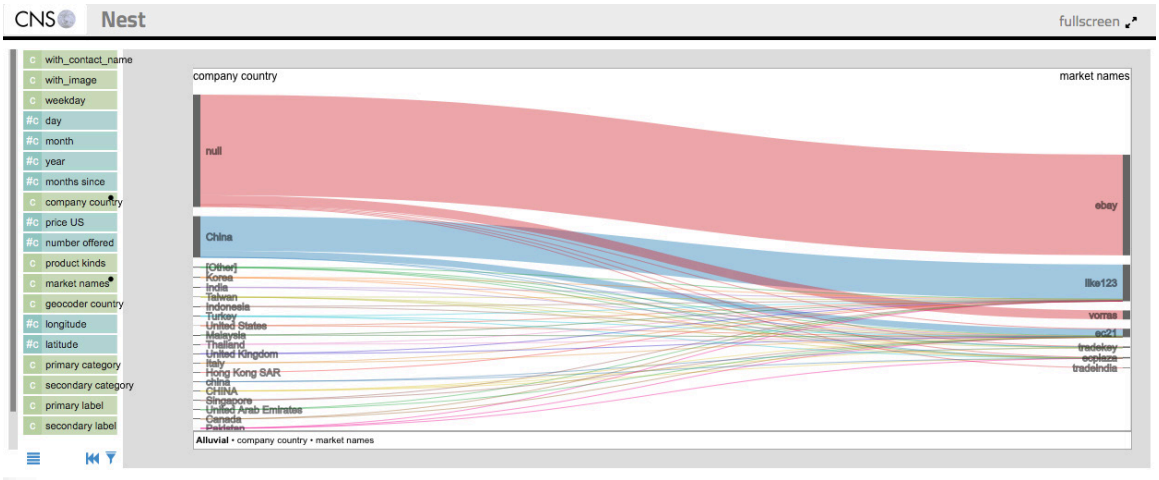
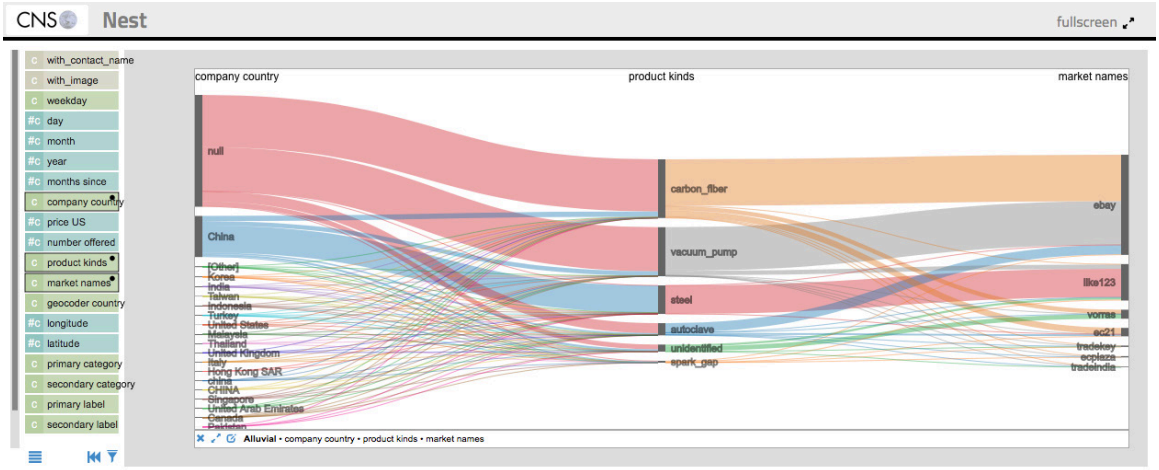


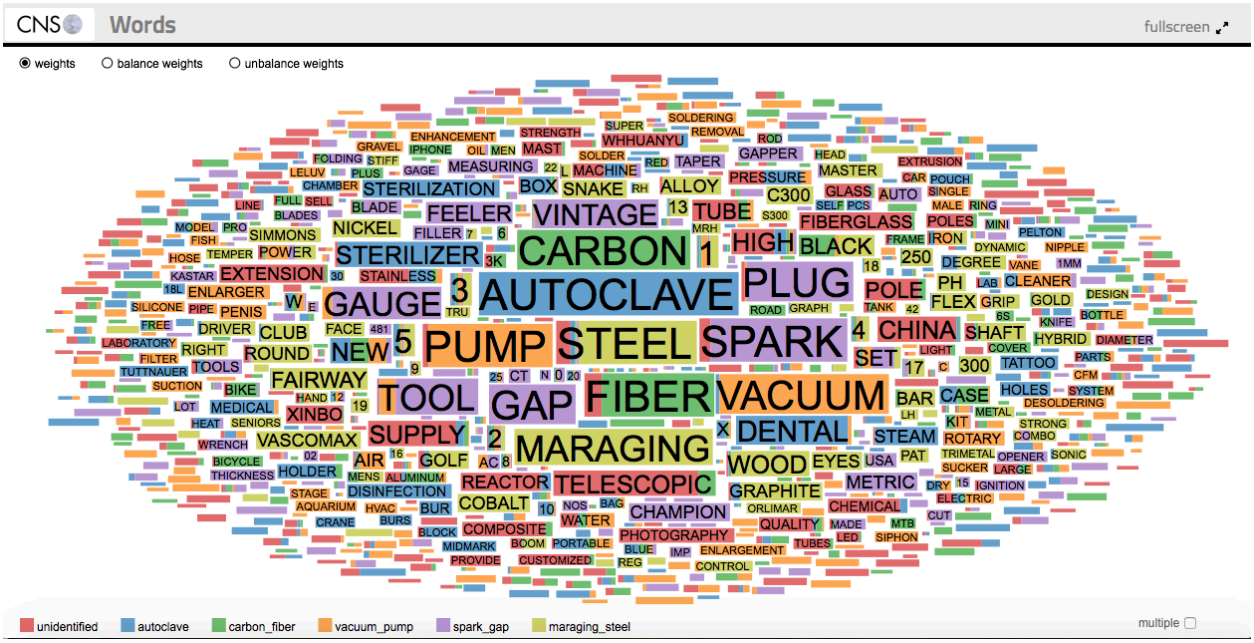
Figure 9: drumwave visualization of country by product & marketplace



Wordcloud visualizations

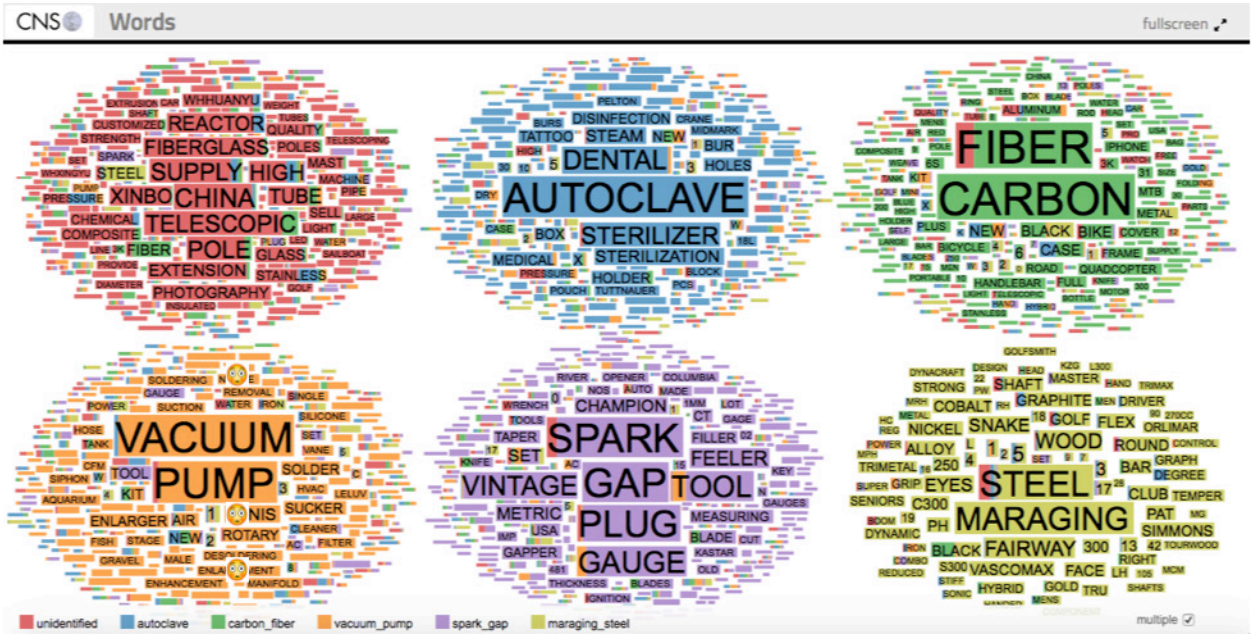
Analyzing the frequency of terms in the product descriptions can be used to create wordcloud visualizations. These terms are sized by their importance in the overall corpus of text found in the product descriptions and grouped with frequent words appearing larger and closer to the center of the diagram. Coloring the terms according to the search keyword further enhances this visualization. The result is a wordcloud (Figure 10) that depicts the most common terms used to describe the products and the proportion of times the term was used in relation to a product keyword. For example, the term "gauge" is violet and orange, indicating the word predominantly appears in descriptions associated with the keyword "spark gap," but is also sometimes associated with the keyword "vacuum pump." As expected, the major keywords are central to the diagram due to their frequency in the total description text corpus.

Figure 10: drumwave proportional wordcloud



Drilling down to the individual keyword categories (Figure 11) reveals the amount of noise in the dataset. The words associated with maraging steel, for example, refer almost exclusively to golf clubs and sporting goods. The carbon fiber listings are also heavily weighted with sporting good products, while health and cosmetic products dominate the vacuum pump keyword. It is interesting to note, however, that "China" is a word that was frequently associated with products in the "unknown" category.

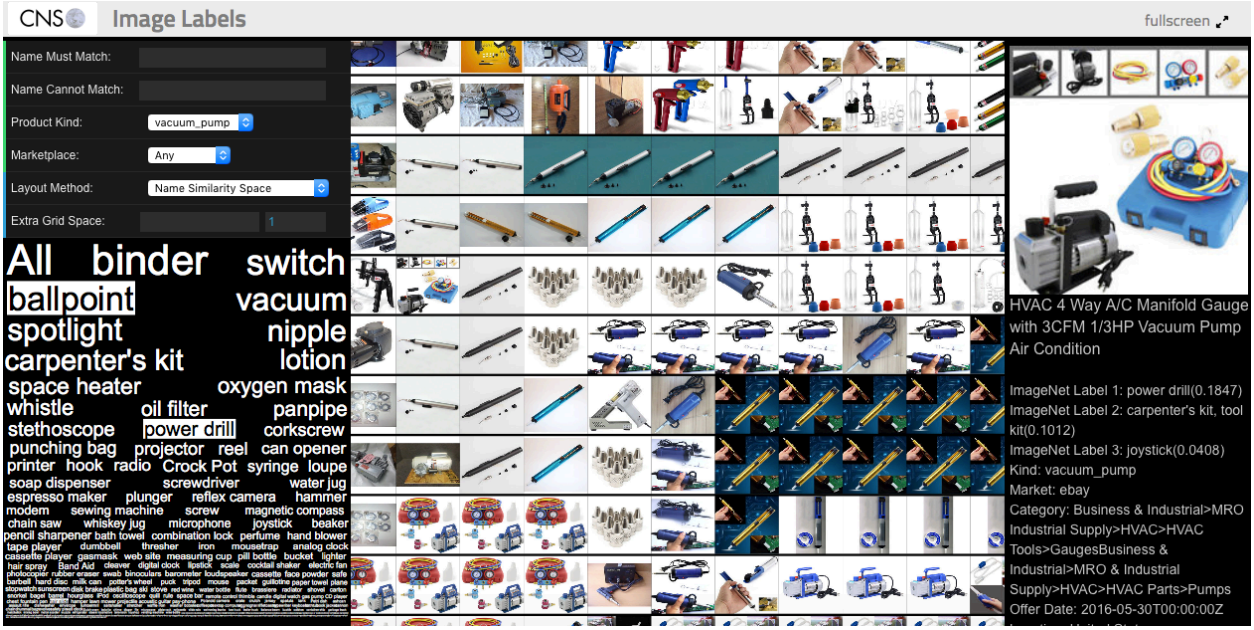
Figure 11 drumwave wordcloud keyword detail



Images

The ability of humans to rapidly recognize and decode images is well known. For this analysis, we also retrieved product images when they were contained in the product description data. Drumwave's visualization tool uses the similarity of the text in the product descriptions to cluster the images together. This allows the analyst to accomplish two things. First, it makes the detection of anomalies easier during the exploratory phase. Items that have similar text descriptions but radically different appearances may be a clue of illicit trading. Second, it provides a frame of reference to help identify whether a given item is indeed a dual-use product. Figure 12 shows an example of product clustering for various items with "vacuum pump" in the description. Clearly, the handheld items can be ignored, while the cluster of larger pumps in the lower left corner may merit further analysis.

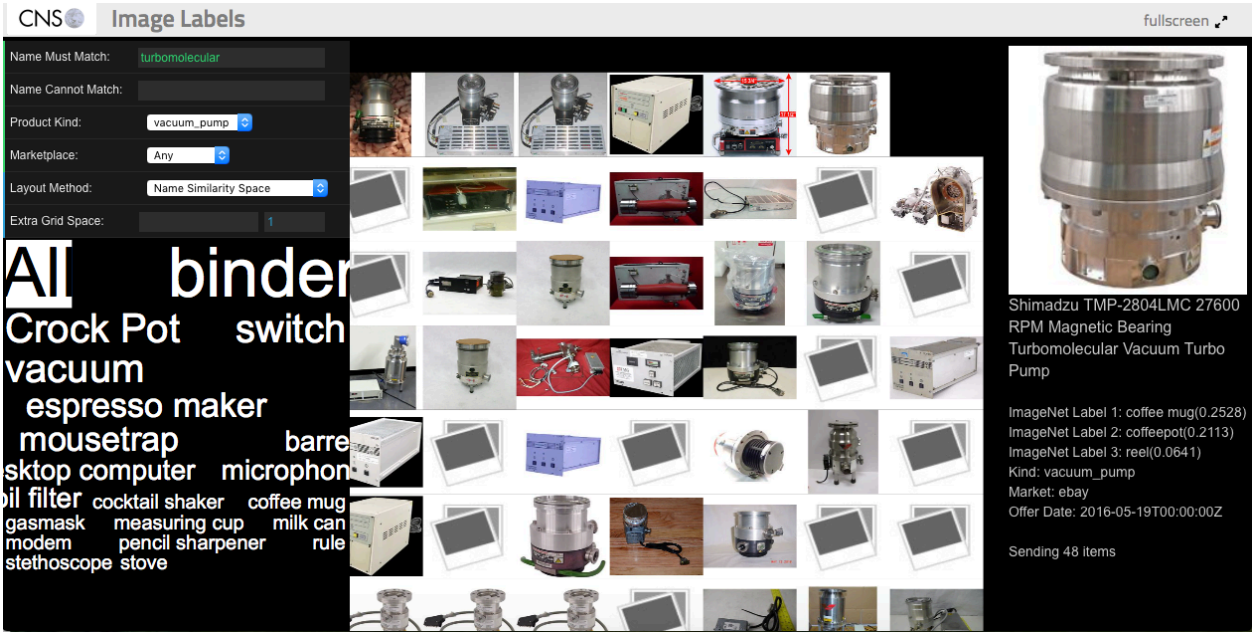
Figure 12: drumwave image clustering tool



Of even more interest is the possibility of using machine learning to augment the analysis. Figure 13 displays the data set listings using a restrictive search term of "turbomolecular." On the right under the image is the product text description from the marketplace listing. This is followed by a set of "best-guess" item descriptions based on the image analysis models of Google's TensorFlow machine-learning library.[14] The labels are often comically wrong, but they are very consistent. Items identified by the machine as a coffee mug or crockpot will all have a similar appearance, even if they are not actually coffee mugs or crockpots.

[14] Martın Abadi et al., "Tensorflow: Large-Scale Machine Learning on Heterogeneous Distributed Systems," arXiv Preprint, March 16, 2016, <arXiv:1603.04467>.

Figure 13 drumwave machine learning image labels



Using the Drumwave tool, each time the analyst selects an item as one of interest, the software updates the machine model with this preference. Over time, the software will refine the model and the labels used until it reaches a point where it is capable of scanning the data itself and retrieving the best candidates for further analysis. This technology is still in its infancy and was not a focus of this study, but it represents an intriguing use of product images. It does not take much imagination to envision an algorithm trained by an analyst on a working database being applied to assist frontline customs agents or WMD inspectors with equipment and product identification tasks.

Companies and points of contact

As mentioned previously, our collection effort attempted to gather as much company information as possible. When available, we recorded company name, address, email points of contact, and telephone numbers. We then ran several database queries, searching for companies that shared contact information. Not surprisingly, the queries turned up numerous instances of companies having multiple names but one single point of contact. For example, Shijiazhuang Dishun Metal Products Co., Ltd., Shulei International, and the Qingdao Flying Industrial Co., Ltd. all list the mysterious "Ms. Linda" as their contact person. This is not to say that the companies are operating illegally, only to point out that multiple identities appear to be as common in the online business world as they are in the online social world.

We were also able to retrieve website information for numerous listings. We used this information to identify the domain name owners using a Whois lookup to verify that the company name and owner matched. We identified one odd entry: The company Medic OSHC identifies itself as a US company in its marketplace listings, but its company website "About" page says that it is based in Hong Kong.

A check of its website domain registration, however, shows the website is hosted by the Indonesian company Ardh Global. Unfortunately for this analysis, the company also uses an Australian internet privacy service, so we have no way of determining the actual owner of the Medicoshc.com website. The multiple points of contact and other anomalies found in the company listings encouraged us to compare all company names and points of contact with the US Consolidated Screening List and the United Nations and European Union sanctions list. None of the entries in our database matched individuals or entities on any of the sanctions lists.

5. Discussion

Limitations

A number of significant limitations and caveats regarding this research must be mentioned. The first and most significant is that the search results contain product listing information only. That is, the items we were able to research represent sell-side data and do not contain any information about buyers. This precludes drawing any conclusions about the extent of illicit trade or the intent of online customers.

There are several significant limitations regarding the data itself. First, as stated above, the search was restricted to five items related to potential research or manufacturing in the nuclear field. There was no effort to search for items associated with chemical or biological weapons production.
Second, just as the number of products was restricted, this project only searched a small sample of business-to-business online trade sites and only gathered data from those sites with permissive terms of service. Since only one of the sites (eBay) made its data available in a structured format, and none of the sites offered application programming interface (API) access, there is the potential that data gathered were incomplete. Further, many vendors only offered incomplete listing information. For example, information on price and quantity offered, as well as product specifications, were frequently omitted.

Third, lack of API access together with the automated keyword-based search procedure meant a focus on breadth and not depth. For example, the products associated with keywords such as “autoclave” or “vacuum pump” are much more numerous than targeted product searches such as “Toray T700 Carbon Fiber.” Targeted searches, however, often return zero results.
Finally, the data were gathered over a limited timeframe of several weeks. This means the data are much more representative of a single search and could have easily missed product listings withdrawn before or added after the search window.

Implications in view of previous work

Previous efforts to detect potentially illicit trade of dual-use items online have relied on manual searches and reviews of product listings. In contrast, we have used keyword search terms and automation to cast a broader net across several online trading sites. The results have been mixed. Earlier and more targeted methods identified several clear-cut cases of trade in dual-use items that were outside of regulatory compliance. Our method did not identify any such cases and also did not identify any products being sold online by individuals or entities on any of the various sanctions lists.

The task of the intelligence analyst is all too frequently compared to the search for a needle in a haystack. The aggregation of data we have collected serves to more sharply define that haystack and gives us a better idea of how to approach the search in the future.

First, in terms of defining the haystack, there is currently no authoritative source for mapping product specifications, names, or common descriptions to trigger list items. This study relies on previous research conducted by the Center for Nonproliferation Studies, but as we have shown, there is little overlap between the technical language used on trigger lists and the product descriptions on online commerce sites. A similar situation exists with respect to sanctions lists. Multiple aliases and alternate spellings or transliterations make it extraordinarily difficult to run a basic comparison between entities on a database such as ours and official sanctions lists.

Next, the online marketplaces themselves could greatly simplify the process by offering access to their data in a structured format using an application program interface (API). This would allow vendor names or points of contact to be extracted easily and compared to the relevant sanctions lists. Mandating that vendors provide complete product specifications for their listings would also make it much easier to identify potential dual-use items.

The study further shows the futility of the purely manual search effort. While it may be possible to identify compliance issues with previously suspect vendors, the sheer volume of online trade dooms a naïve (i.e. routine) search effort to failure. Even with this study's timeframe and marketplace restrictions, the database contains product listings from 9,690 companies based in sixty-nine countries. A naïve manual search through so much data is very likely to miss something. This is especially likely to be the case given the fact that so much of the data online are incorrect, ill-formatted, or simply missing.

It seems clear, then, that some combination of manual and automated searching would be the ideal. There is a large gap between the raw data collected and the final analytical conclusion, and the best path forward is not always intuitive or obvious. Powerful and flexible visualization tools, like those provided by Drumwave, simplify the analyst's task with a streamlined overview of the data and a series of structured viewpoints that can be freely combined to generate insight.

Beyond visualization, however, the incorporation of machine learning tools seems to be the way of the future. Clustering algorithms, such as those used to group product images together in the Drumwave software, could be easily adapted to data elements such as product listing category, countries, price, or product size dimensions. This could effectively shrink the universe of product and company listings to a manageable handful for human review. Moreover, the inclusion of a self-teaching element such as that shown with TensorFlow-based image recognition could result in a benign circle where human review and feedback improves and perfects the automated selection mechanism.

Suggestions for further research

The information provided here has shown a simple method to gather data from online marketplaces and conduct basic analysis. We have identified several additions to our method that may benefit future research projects.

Search refinements

The current search effort was conducted a limited number of times over the course of several weeks. A scheduled and more systematic search effort could generate additional insights. Additionally, incorporating other relevant data such as known equipment manufacturer or vendor names or addresses could help to analyze patterns and trends in the online trading.

Additional data

Combining this dataset with others such as sanctions lists, court cases, trade flows, or buy-sell transaction data should also be investigated. Such a layered technique could lead to new conclusions or generate new analytical methods and approaches to the problem.

Development of an analyst workflow

Data analysis is not always intuitive and can require long periods of study before an analyst is comfortable with the techniques and methods. A review of the data with the goal of developing a basic analytical workflow would benefit both analysts and researchers.

Machine augmented analysis

Further exploration of the potential of machine learning to enhance human analytical efforts would be very beneficial. Computer science research into the field of artificial intelligence is booming,[15] but there has been little effort to explore how progress on topics such as automated image captioning, image recognition, predictive analysis, or deep learning can directly aid nonproliferation analysts.

[15] Sumit Das et al., "Applications of Artificial Intelligence in Machine Learning: Review and Prospect," International Journal of Computer Applications 115, no. 9 (2015).

6. Conclusion

This project has demonstrated that large-scale harvesting of online marketplace information is possible using off-the-shelf open source technologies and basic programming skills. Freely available script-style programming languages such as Python and non-structured query language databases such as MongoDB can be deployed at very low cost and can quickly generate large amounts of online trade data for analysis. Data gathered through such harvesting methods, however, are of generally poor quality. Information in data fields is likely to be missing or incomplete, and a minor change to a website's underlying HTML code can easily frustrate the collection effort.

Despite the limitations of this study, we can draw some tentative conclusions from the data gathered. First, illicit dual-use trade in online marketplaces does not appear to be a large-scale problem in our data, even on less well-known trading sites. There were simply very few items that could be clearly identified as dual-use. This leads to our second conclusion, namely, that most data gathered through automated searching is likely to be irrelevant. Companies do list industrial equipment on online marketplaces, but wholesale products ultimately destined for consumer markets vastly outnumber these listings. Finally, the combination of irrelevant data coupled with the strong suspicion that illicit trade exists argues strongly for the continued development of analytical tools. Machine learning techniques such as data clustering and classification or automated image identification paired with a human analyst would seem to offer the best prospect of consistently finding the ever-elusive needle.

Finally, and perhaps most importantly, closer cooperation between online marketplaces and compliance officials is likely to improve data analysis results profoundly. Something as simple as marketplaces mandating that vendors must provide complete product specification data would allow internal and external compliance officers to easily flag trigger list items. Providing a structured interface to the data through an API would likewise facilitate analysis and make it easier to scan for illicit trading. The computer and the Internet have created the online marketplace revolution. A few cooperative data sharing measures between industry and export control regulators would be an effective and low-cost way to keep the marketplaces safe and open for business.

Annex 1 Product search terms

Item	*Dual-Use Qualifier*	*Korean*	*Russian*
Vacuum pumps	Suction capacity of 5 m3 per minute	真空泵	Вакуумные насосы (plural); Вакуумный насос (singular)
	Resistant to corrosion by UF6: rotary or positive, may have displacement, fluorocarbon seals, working fluids		
	Designed/prepared vacuum manifolds, headers and pumps for service in UF6-bearing atmospheres		
	Made of/protected by UF6-corrosion resistant materials		
	Separator module housings with vacuum pump connections		
	Input throat size equal to or greater than 380 mm		
	Pumping speed equal to or greater than 15 m3 /s		
	Capable of producing an ultimate vacuum better than 13.3 mPa		
	The pumping speed is determined at the measurement point with nitrogen gas or air		
	The ultimate vacuum is determined at the input of the pump with the input of the pump blocked off		
	Bellows-sealed scroll-type vacuum pumps		
	An inlet volume flow rate of 50 m3 /h or greater		
	Pressure ratio of 2:1 or greater		
	Having all surfaces that come in contact with the process gas made from any of the following materials: Aluminium or aluminium alloy; Aluminium oxide; Stainless steel; Nickel or nickel alloy; Phosphor bronze; or Fluoropolymers		
	In a bellows-sealed scroll compressor or vacuum pump, the process gas is totally isolated from the lubricated parts of the pump and from the external atmosphere by a metal bellows. One end of the bellows is attached to the moving scroll and the other end is attached to the stationary housing of the pump		

Item	*Dual-Use Qualifier*	*Korean*	*Russian*
Spark gap plugs	Containing three or more electrodes	火花间隙插头; “Spark gap” = 火花间隙; “plug” =插头	Искровой промежуток пробки; “spark gap” = Искровой промежуток; “plug” = Пробки
	Anode peak voltage rating of 2 kV or more		
	Anode peak current rating of 100 A or more		
	Anode delay time of 10 μs or less		
	Includes gas krytron tubes and vacuum sprytron tubes		
	Anode delay time of 15 μs or less		
	Rated for a peak current of 500 A or more		
	Turn-on time of 1 μs or less		
Maraging Steel	Capable of an ultimate tensile strength of 1.95 GPa or more	马氏体时效钢	
	Ultimate tensile strength of 1950 MPa or more at 293 K (20 °C)		
	Maraging steel before or after heat treatment		
Carbon fibre	--	碳纤维	Углеродное волокно (single strand); углепластики or карбон (composite)
Autoclaves	Especially designed or prepared process systems or equipment for enrichment plants made of or protected by materials resistant to corrosion by UF6	高压灭菌器	Автоклавы
	Autoclaves used in heating UF6 to evaporate solids		

Annex 2 Initial Online Marketplace Listing

Name	*Country*	*Category*	*Description*
Vorras	United States	B2B	Provides a forum for business owners to post offers and requests on topics such as trade, investments, and services.
TradeIndia	India	B2B, B2C	An e-marketplace that specializes in data acquisition and online promotion and provides exporters, manufacturers, importers, and buyers with an online catalog and trade leads.
ECPlaza	ROK	B2B	Provides an e-marketplace, overseas market research, trade infrastructure establishment, on and off-line overseas marketing, and trade consulting services.
EC21	ROK	B2B	Provides a marketplace which specializes in global buyer sourcing, social media marketing, international market research, search engine marketing, and search engine optimization.
TradeKey	Saudi Arabia/ Pakistan	B2B	An e-marketplace that connects traders with wholesalers, buyers, importers, exporters, manufacturers, and distributors in over 240 countries.
Like123	United Kingdom	B2B	An e-marketplace that offers a directory of suppliers, exporters, importers, manufacturers.
E-bay	United States	B2B, B2C	An e-marketplace that focuses on connecting buyers and suppliers as well as on connected commerce.
Alibaba	China	B2B, B2C	Provides an e-marketplace that provides one-stop sourcing for buyers and suppliers.
DGMarket	United States	B2B	Provides an e-marketplace for buyers, suppliers, and partners.
MarginUp	United States	B2B	Provides a real-time trade environment to trade companies, traders, trade brokers or agents, freight carriers, and storage companies worldwide.
ThomasNet	United States	B2B	A supplier discovery resource which allows customers to find specific products and suppliers to advertise their goods.
TradersCity	United States	B2B	A trade import and export trade leads and services marketplace which offers importers exporters, and international traders unlimited trade lead submissions, competitive analysis reports, product photos, and leads' external search keywords and statistics.

Appendix: Summary and Recommendations

Summary of Research and Conclusions

This report explored the potential value of semi-automated approaches to gather and analyze data from online marketplaces to further understanding of the prevalence of dual-use technologies and the potential that these platforms could be utilized by proliferators to illicitly procure them.

- The data analyzed as part of this project suggest that illicit dual-use trade in online marketplaces is not a large-scale problem, even on some of the lesser well-known trading sites.

- The research—although admittedly focused in nature—found very few items that could be clearly identified as dual-use and no examples of designated or sanctioned entities trading on these platforms.
- The combination of irrelevant data coupled with the strong suspicion that illicit trade exists argues strongly for the continued development of analytical tools.
- The research provides insights for the further development and utilization of semi-autonomous tools and methods:
    - Machine learning techniques, such as data clustering and classification or automated image identification paired with a human analyst, would seem to offer the best prospect of identifying proliferation risks.
    - Closer cooperation between online marketplaces, compliance officials, and governments is likely to improve data analysis results profoundly.

More specific recommendations are provided for governments and policymakers, online marketplaces, and civil society below.

Recommendations for Governments and Policymakers

- Governments should further consider how online marketplaces—alongside exporters—could form part of the "first line of defense" against WMD proliferation. Effectively engaging these platforms could have a multiplier effect for industry engagement efforts. Approaches for policymakers in this respect could include:
    - Ensuring that online marketplaces and trading websites headquartered within their jurisdiction are including in national industry outreach strategies;
    - Ensuring that online marketplaces are included in efforts "to work with and inform industry and the public regarding their obligations" under UN Security Council Resolution 1540;
    - Further exploring how the semi-automated approaches described in this report could best be utilized to understand potential proliferation risks in online marketplaces;

- Utilize national technical expertise and understanding of export control lists to facilitate a mapping activity to match possible product descriptions with names and common descriptions of trigger list items.

Recommendations for Online Marketplaces

- Consider ways to make data more user-friendly to researchers and national authorities (the dual-benefit should be emphasized—that such measures will also likely make the website more user-friendly for potential customers):
    - Consider making certain fields mandatory in company and product listings—this could help to ensure that a fuller dataset is available for nonproliferation purposes.
    - Consider making categorization of product listings clearer and mandatory—for example, ensuring that companies listing golf clubs tick a "leisure or sporting goods" box or similar.
    - Offer access to data in a structured format using an application program interface (API).
- Alternatively, consider ways to encourage entities using the platform to enter data in more fields, making the data more user-friendly (again emphasizing the potential dual-benefits for business—e.g. more complete company profiles are more likely to receive business enquiries).
    - This could include implementing a system of "red flags"—e.g. "the authenticity of this profile cannot be verified because the owner did not list an address or phone number."
- Encourage online marketplaces to include generic references to export controls in their terms of use statements and when potential customers submit business enquiries (e.g. "please be aware that this product may be subject to export controls").

Recommendations for Civil Society

- Further develop research in this area to improve understanding of the potential use of online marketplaces for illicit purposes. Several areas for future research have been identified in the report:
    - A scheduled and more systematic search effort along the lines of that undertaken as part of this project could generate additional insights. Additionally, incorporating other relevant data such as known equipment manufacturers or vendor names or addresses could help to analyze patterns and trends in the online trading.
    - Combining this dataset—or those generated using similar means—with others such as sanctions lists, court cases, trade flows, or buy-sell transaction data should also be investigated. Such a layered technique could lead to new conclusions or generate new analytical methods and approaches to the problem.

- Data analysis is not always intuitive and can require long periods of study before an analyst is comfortable with the techniques and methods. A review of the data with the goal of developing a basic analytical workflow would benefit both analysts and researchers.
    - Further exploration of the potential of machine learning to enhance human analytical efforts would be very beneficial.

- Further areas of work for civil society could also help to encourage online marketplaces and governments to take a more proactive role:
    - Civil society could undertake a benchmarking activity—considering which online marketplaces have the best measures in place to prevent illicit WMD-related trade. This benchmarking activity could include metrics relating to the sites' efforts to prevent illicit trade, as well as its willingness to provide structured data for use by researchers.
    - Civil society could work to act as a neutral third party to facilitate discussion between the platforms and governments to further understand illicit trade risks and how to mitigate them.

OCCASIONAL PAPERS AVAILABLE FROM CNS

online at http://nonproliferation.org/category/topics/cns_papers

#33 **Evaluating WMD Proliferation Risks at the Nexus of 3D Printing and Do-It-Yourself (DIY) Communities**
Robert Shaw, Ferenc Dalnoki-Veress, Shea Cotton, Joshua Pollack, Masako Toki, Ruby Russell, Olivia Vassalotti, Syed Gohar Altaf • 2017

#32 **Taiwan's Export Control System: Overview and Recommendations**
Melissa Hanham, Catherine Dill, Daniel Salisbury, P. Alex Kynerd, Raymond Wang • 2017

#31 **Revisiting Compliance in the Biological Weapons Convention**
James Revill • 2017

#30 **Crowdsourcing Systems and Potential Applications in Nonproliferation**
Bryan Lee • 2017

#29 **The Verification Clearinghouse: Debunking Websites and the Potential for Public Nonproliferation Monitoring**
Bryan Lee, Kyle Pilutti • 2017

#28 **Geo4nonpro.org: A Geospatial Crowd-Sourcing Platform for WMD Verification**
Melissa Hanham, Catherine Dill, Jeffrey Lewis, Bo Kim, Dave Schmerler, Joseph Rodgers • 2017

#27 **Searching for Illicit Dual Use Items in Online Marketplaces: A Semi-Automated Approach**
Bryan Lee, Margaret Arno, Daniel Salisbury • 2017

#26 **2016 Symposium Findings on Export Control of Emerging Biotechnologies**
Steven Fairchild, Caroline R. M. Kennedy, Philippe Mauger, Todd J. Savage, Raymond A. Zilinskas • 2017

#25 **Outlawing State-Sponsored Nuclear Procurement Programs & Recovery of Misappropriated Nuclear Goods**
Leonard S. Spector • 2016

#24 **Strengthening the ROK-US Nuclear Partnership**
Miles A. Pomper, Toby Dalton, Scott Snyder, Ferenc Dalnoki-Veress • 2016

#23 **Replacing High-Risk Radiological Materials**
George M. Moore, Miles A. Pomper • 2015

#22 **A Blueprint to a Middle East WMD Free Zone**
Chen Kane, PhD • 2015

#21 **Biotechnology E-commerce: A Disruptive Challenge to Biological Arms Control**
Raymond A. Zilinskas, Philippe Mauger • 2015

#20 **Countering Nuclear Commodity Smuggling: A System of Systems**
Leonard S. Spector, Egle Murauskaite • 2014

#19 **Alternatives to High-Risk Radiological Sources**
Miles Pomper, Egle Murauskaite, Tom Coppen • 2014

#18 **Stories of the Soviet Anti-Plague System**
Casey W. Mahoney, James W. Toppin, Raymond A. Zilinskas, eds. • 2013

#17 **Ugly Truths: Saddam Hussein and Other Insiders on Iraq's Covert Bioweapons**
Amy E. Smithson, PhD • 2013

#16 **Rethinking Spent Fuel Management in South Korea**
Ferenc Dalnoki-Veress, Miles Pomper, Stephanie Lieggi, Charles McCombie, Neil Chapman • 2013

Older Papers

#15 Engaging China and Russia on Nuclear Disarmament • *2009*

#14 Nuclear Challenges and Policy Options for the Next US Administration • *2009*

#13 Trafficking Networks for Chemical Weapons Precursors: Lessons from the 1980s Iran-Iraq War • *2008*

#12 New Challenges in Missile Proliferation, Missile Defense, and Space Security • *2003*

#11 Commercial Radioactive Sources: Surveying the Security Risks • *2003*

#10 Future Security in Space: Commercial, Military, and Arms Control Trade-Offs • *2002*

#9 The 1971 Smallpox Epidemic in Aralsk, Kazakhstan, and the Soviet Biological Warfare Program • *2002*

#8 After 9/11: Preventing Mass-Destruction Terrorism and Weapons Proliferation • *2002*

#7 Missile Proliferation and Defences: Problems and Prospects • *2001*

#6 WMD Threats 2001: Critical Choices for the Bush Administration • *2001*

#5 International Perspectives on Ballistic Missile Proliferation & Defenses • *2001*

#4 Proliferation Challenges and Nonproliferation Opportunities for New Administrations • *2000*

#3 Nonproliferation Regimes at Risk

#2 A History of Ballistic Missile Development in the DPRK • *1999*

#1 Former Soviet Biological Weapons Facilities in Kazakhstan: Past, Present, and Future • *1999*

nonproliferation.org

Middlebury Institute *of*
International Studies at Monterey
James Martin Center for Nonproliferation Studies