

THE GLOBAL INITIATIVE AGAINST TRANSNATIONAL ORGANIZED CRIME

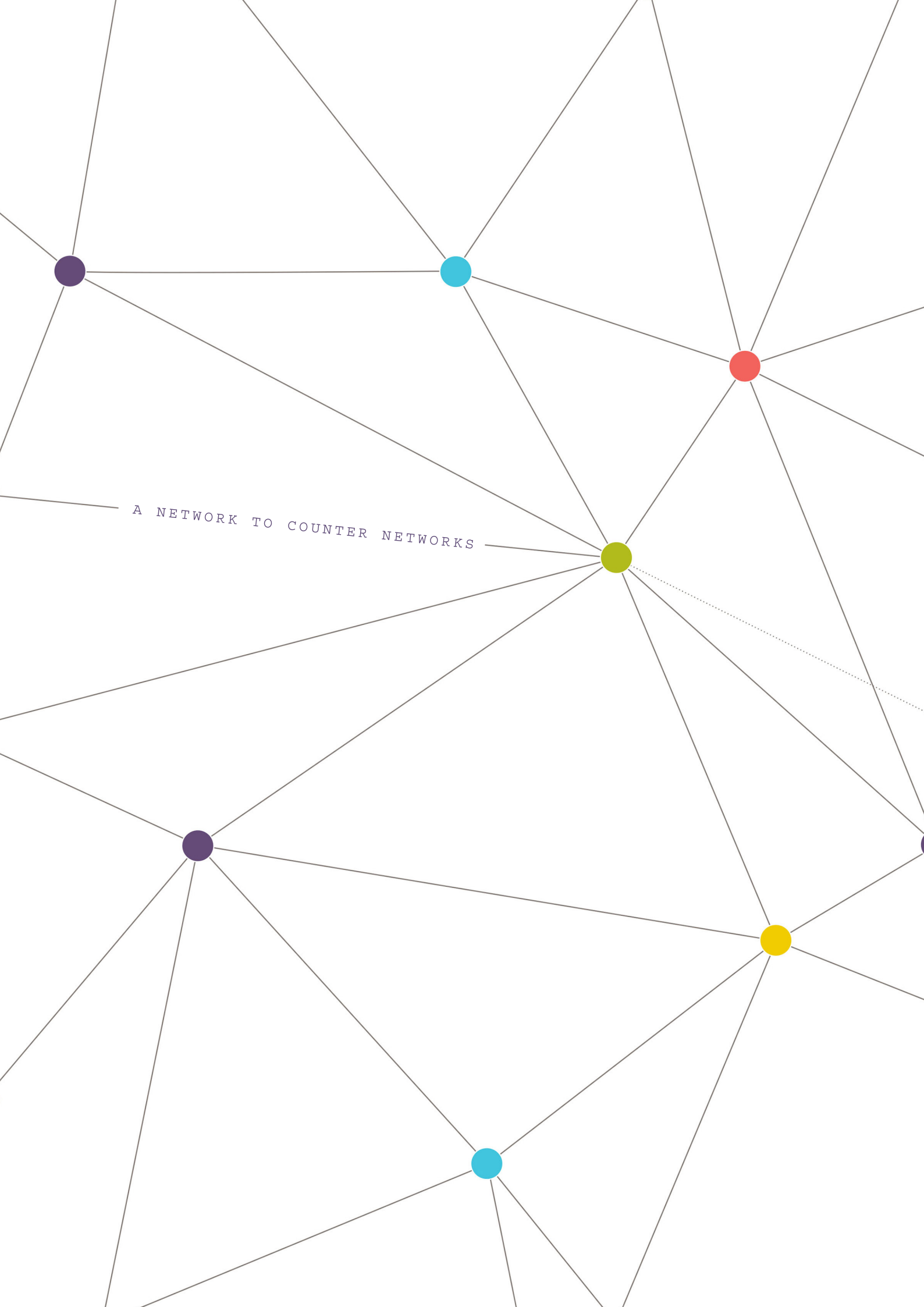


DETECTING ONLINE ENVIRONMENTAL CRIME MARKETS

Carl Miller | Jack Pay | Josh Smith



January 2019



A NETWORK TO COUNTER NETWORKS



DETECTING ONLINE ENVIRONMENTAL CRIME MARKETS

Carl Miller | Jack Pay | Josh Smith

January 2019



Cover illustration: iStock/Egor Suvorov

© 2018 Global Initiative Against Transnational Organized Crime. All rights reserved.

No part of this publication may be reproduced or transmitted in any form or by any means without permission in writing from the Global Initiative. Please direct inquiries to:

The Global Initiative Against Transnational Organized Crime
WMO Building, 2nd Floor
7bis, Avenue de la Paix
CH-1211 Geneva 1
Switzerland

www.GlobalInitiative.net



Summary

The internet is used to trade endangered species and commodities containing parts from endangered species, and more broadly hosts communities and subcultures where this trade is normalized, routine and unchallenged.

This report presents a new technical process that has been trialled to identify online marketplaces and websites involved in the trade of a selection of CITES-listed animals and plants. The technology, known as the Dynamic Data Discovery Engine (or DDDE), was developed with the aim of building upon qualitative research to produce larger, more comprehensive datasets of similar activity taking place. The report contains a description of the process and the results that it produced, its strengths and weaknesses, and some thoughts on how it might be used by others hoping to reduce the extent to which the internet can be exploited by those wishing to transact endangered animals and plants. It is hoped that the process will contribute to the creation of a more comprehensive picture of online illicit wildlife trade (IWT) activity.

Key points

- New Dynamic Data Discovery Engine technology was developed and trialled across three case studies: orchids, pangolins and ivory. Its intention in each case was to identify as many URLs as possible, and as precisely as possible, that were engaged in either the transaction of the given species or commodities containing them, or conversations about such commodities that normalized these activities.
- Although it was found the DDDE could increase the number of URLs identified to a scale impossible through manual analysis, it could not do so precisely. A significant proportion of each of the datasets produced for each of the case studies was not relevant to the purposes of this project.
- The resulting datasets are large and complex, and contain both relevant and irrelevant results. Also included in this report are a number of attempts to represent and understand the URLs that were recovered. This included qualitative analysis and narrative of smaller amounts of randomly sampled data, and data-visualization approaches to identify different 'communities' of websites within the overall results.
- It is hoped that the outputs from DDDE processing will complement and support more qualitative and immersive forms of research on online environmental crime.



Introduction and background

The aim of this project has been to compose and trial a new technical methodology, the Dynamic Data Discovery Engine (DDDE). This is a process that aims to build as comprehensive a picture as possible of how, where and when plants and animals that are listed by the Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES), or commodities containing them, are transacted over the internet. It also aims to identify the broader conversations related to these species and products.

Using the DDDE, the project aimed to quantitatively measure the nature and scale of online venues in which trade in endangered plants and animals occurs. This exercise included:

- Identifying where on the internet this kind of activity happens: what are the platforms, forums, websites, services or e-commerce portals where this activity occurs?
- Measuring, as far as possible, the total scale of the activity and how is it distributed.
- Understanding the nature of the activity: what can be learned about the nature of the commodities, conversations and networks related to this activity? This priority reflects the internet's role in allowing the formation of groups, and indeed subcultures, where the purchase and use of commodities containing endangered animals are normalized and encouraged. Therefore, the project sought not just to detect direct sales, but also wider conversations that socially and culturally underlie them.

This is a working paper that discusses a technology process. The objective of the project was to unite and use technologies developed for web scraping, machine learning and data visualization into a single, repeatable process that could find many more examples of online environmental crime than manual approaches could alone. It was hoped this kind of process could become a 'force multiplier' for the qualitative forms of research that have been carried out in this area, allowing them to inform more scaled attempts at detection.

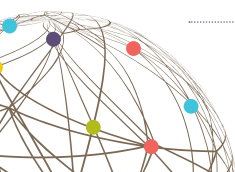
Underlying this work is the view that we do not currently have a robust, repeatable way of detecting environmental crime online, and that we need one. We do not have a comprehensive picture of the true scale of the phenomenon, where on the internet it happens, and which species it involves. Moving towards having this picture is important: for public advocacy, for engagement with the technology companies whose platforms are involved, and for governments to better understand how they need to respond.

Case studies

To develop, test and explain the DDDE, three case studies were selected, each of which emphasizes a different aspect of the overall process and project:

- A DDDE to identify the sale of CITES-listed orchids. This was conducted first to trial the initial process.
- A DDDE to identify the sale of pangolins and related products was conducted to trial the qualitative research aspects of the project.
- A DDDE to identify the sale of ivory was conducted to trial the quantitative aspects of the project and to increase the scale of data collected and analyzed.

The three case studies are discussed in more detail later in the report.



Results

The output of the DDDE in each of the case studies was a list of URLs considered to be related to the transaction of a specific animal or plant, a commodity containing that animal or plant, or a wider online discussion considered to be related to the animal, plant or commodity. In brief, the results were:

- very large in scale – numbering between 40 000 and 120 000 URLs; the datasets produced were too large to read ‘manually’ or understand in their entirety;
- a combination of both relevant and irrelevant URLs. As discussed below, the process was not able to remove all URLs that were not related to the IWT transactions under study;
- possibly relevant for various reasons: of the URLs that were relevant, some could be related to the formal sale of a commodity, others to the discussion of a commodity, while others may suggest that the transaction of a commodity was possible without explicitly saying so.

Given these complexities, the results of the project are communicated in a number of different ways:

- ‘Top-line’ results are reported in terms of the overall number of URLs collected from each case study and the number we judge to be relevant to the overall aims of the project.
- Within the write-ups of the case studies included in this report, we present more qualitative commentaries of the nature of the URLs found.
- Visualizations of the results of each case study are presented, which attempt to highlight different ‘communities’ of web pages – some more relevant than others.

Top-line results

It should be noted that the emphasis of all the case studies was to identify transaction-related activity connected to CITES-listed species, and not necessarily illegal activity. There are important legal and indeed moral nuances related to each of the cases this paper examines, and the approach was not intended, nor was it able, to tease these nuances apart. We therefore do not suggest that either the numerical results presented below, or the specific examples given, necessarily indicate illegal activity.

Online transactions of orchids

The case study collected 121 741 web pages from 3 329 sites. Of these:

- 1 068 were pages on eBay that mention a restricted species of orchid, of which we judge 58% were selling restricted species of orchids, or linked to auctions that had taken place.
- These plants were being sold by 10 vendors based in the UK, the US, Germany, China, Malaysia and Thailand.
- Most vendors were responsible for only a small number of pages, though one eBay vendor occurred 34 times in our dataset.
- An additional nine different domains selling CITES-listed orchids were found.



Online transactions of pangolins

The search found that pangolin sales tended to involve mainly drug and herbal products and brands for use in traditional Chinese medicine. The sale of these goods was widespread and often in plain sight. The results show a large number of online discussions related to their use, ownership and efficacy.

Using the DDDE, 39 823 unique URLs were gathered. Of these, a random sample of 400 web pages was drawn from the results to assess the overall distribution of sites within the dataset. (This number was calculated to be the most likely to return a representative sample of the dataset as a whole.) These pages were manually analyzed and, if extrapolated across the whole dataset, suggest that 4 878 URLs were found to be relevant to the transaction of pangolins.

Of these 4 878, we judge (again, based on manual inspections of the collected data) that the DDDE collected:

- 2 688 sites (or 55.2% of the total) that recommend or discuss the use of commodities containing pangolin ingredients in the context of traditional Chinese medicine;
- 1 891 (38.7%) sites containing articles or commentaries about products or ingredients containing pangolins; and
- 298 (6.1%) sites related to the sale of items containing pangolin ingredients.

Online transactions of ivory

A DDDE calibrated to find the online sale of ivory produced 45 319 URLs. Based, again, on a manual inspection of a smaller sample of representative data, we judge that 7 840 URLs are related to the online sale of ivory. As stated above, however, it should be noted that this did not necessarily mean such examples where illegal.

Of these:

- 3 127 sites (38.8% of the total) contain more general commentary about commodities containing ivory.
- 2 175 sites (27.7%) contain descriptions of ivory-related products.
- 2 447 sites (31.2%) are related to the sale of ivory.

The Dynamic Data Discovery Engine

Each of the case studies emphasized a different aspect of the overall method, seeking to test different research techniques to confront the multiple challenges this research entails. They were all aimed, however, at producing a unitary process that could, with some recalibration and configuration, be applied to the detection of online transaction and sales activities involving any CITES-listed plant or animal.

The aim was also to create a process that could cascade and be dynamic. A 'cascade' means that the end products of the process can be used to begin another, allowing the process to build on what has been found, and then continue to discover further things. This cascading capability allows the process to dynamically change over time. The DDDE was therefore designed, at least in theory, to be able to learn from itself and, in doing so, track important changes to the online transactions of endangered species, as this phenomenon itself continues to evolve. The overall process of the DDDE has six stages.



Stage 1: Building and cleaning an initial dataset

The first stage of the process is to build an initial dataset of activity as similar as possible to the activity the DDDE is intended to find. This can be done by:

- The collection of 'seed documents' from existing secondary-source literature and subject-matter experts.
- Using a process called cyclical extension, a small number of phrases are compiled that are considered to be relevant. Each is combined with every other, and used as queries within a small number of web searches. The URLs returned from each of these web searches are compiled and ranked on the basis of the number of times that each of these different web searches found it. This process is cyclical because it is repeated a number of times, and each time the websites are manually inspected for new relevant phrases, which are then used as queries within successive web searches.

The initial dataset produced by these steps commonly contains pages that are both very relevant and entirely irrelevant. A researcher appraises the dataset and, based on what is found, takes a number of steps to programmatically remove irrelevant data from this initial sample. Often this entails removing content in languages outside of the scope of study (in all three case studies, only English-language documents were retained), and removing reference sites, such as Wikipedia.

The outcome of stage 1 is an initial dataset of websites that contain, as far as possible, examples of transactions involving the species in question.

Stage 2: Extracting key phrases

Once the dataset has been cleaned, relevant phrases are extracted from it. To do this, the dataset of relevant activity is linguistically compared against another dataset of online web pages that discuss and describe the same species of plant or animal in question, but not in the context of a transaction or a sale. The aim of this process is to identify and extract phrases that occur within the relevant dataset, but that do not occur within the comparison dataset.

The outcome of stage 2 is therefore a sequence of phrases that occur significantly more often in activities related to the transaction of the species in question than they do in activities related just to descriptions of the species in question.

Stage 3: Web search and crawling

Each individual phrase is then combined with every other phrase to form a large number of pairs of phrases. Each of these pairs is then used as the keywords of web searches, which are programmatically conducted using the Microsoft Bing API (application programming interface). Every web search is calibrated to return a differing number of URLs, based on an appraisal of their overall relevance.

Once the URLs are returned by these searches, they are then 'crawled'. This is the retrieval of the full information contained on the page – and all other pages that are linked to it. The crawling typically occurs to what is known as 'depth two', which is where all websites linking to the websites returned by the search are also collected, and also all the websites linking to them.

The outcome of stage 3 is an extended list of possibly relevant websites (typically tens or even hundreds of thousands of URLs).



Stage 4: Finding and categorizing relevant activity

A series of analytical steps are now followed to find the relevant data within this much larger database that has been amassed. These steps include:

- **Domain mapping:** This entails separating the dataset into the different domains that make it up – for instance, social-media platforms, e-commerce sites, reference sites, such as Wikipedia, discussion forums, and so on.
- **Cross-referenced keyword analysis:** Each site is checked for the presence of keywords that it contains. These keywords are drawn from a qualitative analysis of the data, and often relate to the key themes that occur within the dataset, such as sales-related conversations, mentions of known sellers, purchasing-related discussions, and so on. The keywords are then cross-referenced: all the sites are sorted into different categories on the basis of the number and type of different themes that they contain.
- **Machine learning classification:** Machine learning classifiers are trained to categorize documents on the basis of the combinations of words and phrases that they contain. (Machine learning classifiers are algorithms that attempt to place any given item of content into categories defined by the analyst.) The classifier is trained through the provision of examples of documents that fall into each category provided by the analyst. The classifier attempts to find, through these examples, the language that most correlates with each of the categories, and abstract a series of rules that are used to classify additional documents the analyst has not seen. This is often conducted in combination with other forms of analysis, and is used to distinguish between relevant and irrelevant data.

Each of these three analytical steps aims to separate the websites that are relevant from those that are not. The outcome of this stage is consequently a list of URLs that is often a great deal shorter than that resulting from stage 3 but of greater relevance in terms of identifying sites where there is transaction of the species in question.

It is at this point that the process can be cascaded. The documents that are now judged to be relevant are subjected to the key-phrase extraction process described in stage 2. They are compared against a baseline set, and additional key phrases are extracted and inserted as search terms at the beginning of stage 3. The remainder of the process is carried out again, and can be done so a number of times.

Stage 5: Additional quantitative analysis and qualitative commentary

Next, an analyst inspects the documents collected by this process. The purpose of this is to build a contextual view of the activities that have been gathered. This stage combines a number of programmatic and qualitative approaches to clean, categorize and describe the kinds of activities that the DDDE has captured. This may include:

- Additional cleaning of data.
- Identifying additional key phrases to be cascaded as additional web searches.
- A commentary or description of the nature of the sites that have been identified.
- Describing intent and purpose of online discussions, and the products involved.
- Identifying overall patterns and trends in the data.

Stage 6: Data visualization

The resulting data from this process is then rendered into a vis-analytical dashboard to allow wider inspection and interrogation of the project's outputs. (The data visualization is discussed in more detail later in the report.)



Case studies

Orchids

This was the first case study undertaken – its primary purpose being to trial and test the initial data-collection and analysis strategies, which would then be scaled up across the two other case studies.

Researchers first gathered a list of 14 sites that are known to sell a restricted species of orchid, including pages from eBay, Instagram, Facebook and online forums. A web-scraper was configured to collect the content of these pages, where it was possible and legal to do so.

Characteristic words and phrases appearing on these 14 pages were extracted using a process called ‘surprising phrase detection’: this allows the identification of words and phrases that are present in one dataset, but not in another. This way, five lists of phrases were generated, each containing 100 terms. Two of these lists were selected as particularly relevant to the study.

Using a Microsoft search API, web searches for pages containing a combination of two of these relevant phrases were conducted. This produced a dataset of 121 741 web pages from 3 329 sites. All posts within the dataset were annotated using a pre-trained language annotation algorithm within Method52.

Table 1: Orchid case study dataset

Language	Pages	% of dataset
Total	121 741	100%
English	96 421	79.2%
Tagalog	3 279	2.7%
Russian	1 814	1.5%
Indonesian	1 506	1.2%
(Simplified) Chinese	1 290	1.1%
French	988	0.8%
Italian	972	0.8%
Catalan	971	0.8%
Romanian	913	0.8%
Thai	876	0.7%
Spanish	857	0.7%
German	842	0.7%
Farsi	697	0.6%
Hungarian	642	0.5%
Turkish	609	0.5%
Dutch	581	0.5%

Researchers put together a pipeline designed to discover mentions of species listed in the CITES appendix 1 (CITES A1), retrieved in May 2018 (from <http://checklist.cites.org>). To supplement this, a list of abbreviations for 2 800 common orchid phyla was recovered from www.orchidsplus.com/orchid-abbreviations/. All posts were then removed if they did not contain a mention of a CITES A1-listed species – either using the long-form of the species name (e.g. *Phragmipedium klotzschianum*) or the abbreviated name with or without a full stop (*Phrag. klotzschianum*).

For many of the pages in the dataset, the likelihood that they would include discussions that indicated the sale of restricted orchids could be assessed with some confidence from their domain: a site hosted on wikipedia.com, for example, is highly likely to contain reference material, as opposed to transactions. (Pages from Wikipedia alone made up 72% of the initial dataset.)

Therefore, a keyword annotator was used to sort pages into four categories based on their domain. Sites classified as 'reference' were removed; sites classified as 'social', or 'sales' were elevated to the next stage (see Table 2).

Table 2: Orchid dataset sorted into domain categories

Source type	Pages	% of dataset	Pages mentioning CITES A1 restricted species	% of source type mentioning CITES A1 restricted species
	121 741	100.0%	4 748	3.9%
Reference	87 282	71.7%	1 689	1.9%
Not annotated	18 693	15.4%	1 803	9.7%
Sales	11 610	9.5%	1 140	9.8%
Social	4 156	3.4%	116	2.8%

There were 18 693 sites that were not labelled by any of the categories above. After these sites had been filtered to be English-language only, 13 024 pages remained. The 'not annotated' category contains pages that were not from a domain known to fall under any other label, and may therefore contain previously unknown sites on which sales were transacted. To identify these, researchers trained an additional classifier to determine whether commercial activity existed within these unlabelled pages (see Table 3).

Table 3: Orchids – activities on unlabelled pages

Discussion on page	Total	CITES A1 restricted
	13 024	1 806
Retail	1 298	75
Other	11 726	1 731

This retail classifier identified a further 1 298 pages (almost 10%) within those not already caught by the 'source type' filter above that contained discussions around the sale of orchids. Among this dataset, 75 pages were found that mentioned a CITES A1 restricted species. These domains are shown in Table 4. It should be noted that a number of these sites have been misclassified: plantsoftheworldonline.org, for example, is actually a reference site run by Kew Royal Botanical Gardens. Nevertheless, this method has been effective in producing a shortlist of sites that are likely to merit further investigation, or the application of additional technological steps to more precisely separate relevant from irrelevant information.



Table 4: Pages with mentions of sales of CITES A1 restricted orchids

Page domain	Pages
plantsoftheworldonline.org	30
efloras.org	9
ecomingafoundation.wordpress.com	8
orchidgarden.co.uk	4
slipperorchids.info	4
paramountorchids.com	3
schordje.com	2
orchids.uk.com	2
paphs.net	1
ncbi.nlm.nih.gov	1
gd.eppo.int	1
aquariusorchids.com.au	1
orchidspecies.com	1
slipperorchidforum.com	1
orchids.com	1
powo.science.kew.org	1
slipperorchid.com	1
orchid-care-tips.com	1
yonggee.name	1
tropicos.org	1
cyps.us	1

Pangolins

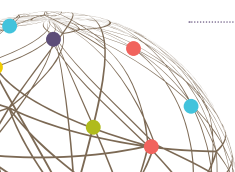
Using an initial set of search terms provided by subject-matter experts, a web search was performed to create an initial dataset. These search terms contained different names for pangolins.

The results of the search were then manually filtered, scraped for their content, and a phrase extractor was used to produce an initial list of possible key phrases. These were manually filtered and combined with the original terms to produce a full list of search terms.

A list of keywords was created that relates to five overall categories:

- Pangolin names (e.g. Latin binomials)
- Brands known to sell pangolin products
- Communication types (WhatsApp, WeChat, etc.)
- Words associated with purchasing activities ('buy', 'purchase', etc.)
- Sites known to potentially sell products containing pangolin parts

An initial attempt to create a relevancy classifier using the entire dataset failed, as it was found to be extremely inefficient at presenting relevant documents for training. Coupled with the typical size of these documents, it yielded poor results.



Instead, a classifier was trained using a dataset comprising all documents containing at least one instance of a pangolin name and at least one instance of a keyword that was related to brands, sites, purchase types or communication type. These were:

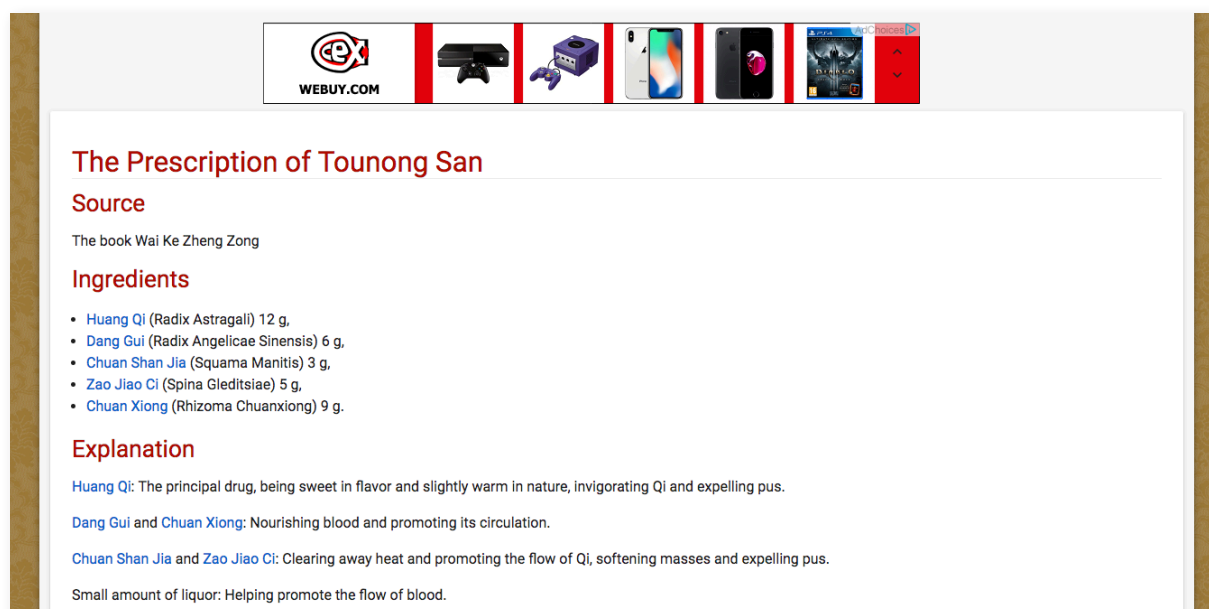
Table 5: Pages containing combinations of keywords

Keyword combination	Count	% of total
'Pangolins' and 'brands'	342	8.03%
'Pangolins', 'brands' and 'purchases'	320	7.77%
'Pangolins', 'brands' and 'sites'	3	0.07%
'Pangolins', 'brands' and 'communication types'	74	1.79%
'Pangolins' and 'communication types'	122	2.96%
'Pangolins' and 'purchasing activities'	503	12.21%
'Pangolins' and 'sites'	0.07	0.07%

Once trained, the entire dataset was then classified. Combining these forms of analysis, a final dataset of 39 823 URLs was produced. Based on a randomly selected sample of 400 of these, 18% were judged to be relevant to the transaction of pangolins. (Examples of these websites identified are shown below.)

Of the identified websites, 11.5% concerned the use of pangolin-derived products to treat ailments in traditional Chinese medicine (see Figure 1):

Figure 1: Example of traditional Chinese medicine site discussing pangolin-derived products



WEBUY.COM

The Prescription of Tounong San

Source
The book Wai Ke Zheng Zong

Ingredients

- [Huang Qi](#) (Radix Astragali) 12 g.
- [Dang Gui](#) (Radix Angelicae Sinensis) 6 g.
- [Chuan Shan Jia](#) (Squama Manitis) 3 g.
- [Zao Jiao Ci](#) (Spina Gleditsiae) 5 g.
- [Chuan Xiong](#) (Rhizoma Chuanxiong) 9 g.

Explanation

[Huang Qi](#): The principal drug, being sweet in flavor and slightly warm in nature, invigorating Qi and expelling pus.

[Dang Gui](#) and [Chuan Xiong](#): Nourishing blood and promoting its circulation.

[Chuan Shan Jia](#) and [Zao Jiao Ci](#): Clearing away heat and promoting the flow of Qi, softening masses and expelling pus.

Small amount of liquor: Helping promote the flow of blood.

Figure 2: A website listing the sale of pangolin scales

C0140	赤芍	Chi Shao	<i>Paeonia lactiflora</i> (root)	<i>Paeoniae Radix Rubra</i>	Red Peony Root	2.0/10
C0150	赤石脂	Chi Shi Zhi	<i>Halloysitum rubrum</i> (ore)	<i>Halloysitum Rubrum</i>	Red Halloysite	0.5/15
C0160	赤小豆	Chi Xiao Dou	<i>Vigna umbeuata</i> (seed)	<i>Vignae Semen</i>	Rice Bean	0.7/10
C0170	茺蔚子	Chong Wei Zi	<i>Leonurus japonicus</i> (fruit)	<i>Leonuri Fructus</i>	Motherwort Fruit	0.5/5
C0180	川貝母粉	Chuan Bei Mu Fen	<i>Fritillaria unibracteata</i> (bulb)	<i>Fritillariae Cirrhosae Bulbus</i>	Tendrilleaf Fritillary Bulb	1.5/3
C0190	川楝子	Chuan Lian Zi	<i>Melia toosendan</i> (fruit)	<i>Toosendan Fructus</i>	Szechwan chinaberry Fruit	1.0/10
C0191	炒川楝子	Chao Chuan Lian Zi	<i>Melia toosendan</i> (fruit,prepared)	<i>Toosendan Fructus Praeparatus</i>	Prepared Szechwan chinaberry Fruit	1.0/10
C0200	川牛膝	Chuan Niu Xi	<i>Cyathula officinalis</i> (root)	<i>Cyathulae Radix</i>	Medicinal Cyathula Root	2.5/10
C0211	炮山甲	Pao Shan Jia	<i>Manis pentadactyla</i> (scale, prepared)	<i>Manis Squama Praeparata</i>	Pangolin Scales (scalded)	1.4/6
C0220	穿心蓮	Chuan Xin Lian	<i>Andrographis paniculata</i> (herb)	<i>Andrographis Herba</i>	Common Andrographis Herb	0.8/10
C0230	川芎	Chuan Xiong	<i>Ligusticum chuanxiong</i> (rhizome)	<i>Chuanxiong Rhizoma</i>	Szechwan Lovage Rhizome	1.3/6
C0240	垂盆草	Chui Pen Cao	<i>Sedum sarmentosum</i> (herb with root)	<i>Sedi Herba</i>	Stringy Stonecrop Herb	1.2/15
C0250	椿皮	Chun Pi	<i>Ailanthus altissima</i> (root-bark or stem-bark)	<i>Ailanthi Cortex</i>	Tree-of-heaven Bark	0.5/6
C0260	刺五加	Ci Wu Jia	<i>Acanthopanax senticosus</i> (root & rhizome)	<i>Acanthopanax Senticosi Radix Et Rhizoma Seu Caulis</i>	Manyprickle Acanthopanax	0.8/20
C0270	磁石	Ci Shi	<i>Magnetitum</i> (ore)	<i>Magnetitum</i>	Magnetite	0.5/15

Figure 3: An online order form that lists pangolin-derived medicinal products

[www.tcm-treatment.net/images/wholesale/herb-price/herb-order-1.htm](#)

Wholesale Herb Order Form One

Your email:

 (needed*)

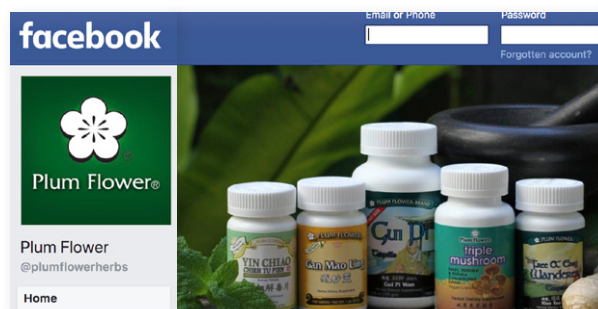
Your company (or personnel) name, address, zip code, Tel and fax so that we could export the herbs to you. Also let us know the port you prefer.

Please tick off every item you need, and fill out the weight of each item in kilogram. The price is in US\$ per kilogram.

- ☐ kg A2, a jiao zhu ass hide glue pellets Asini Corii Gelatini Pilula
- ☐ kg A3, a wei asafetida Asafoetida
- ☐ kg A4, ai di cha Japanese ardisia stem and leaf Ardisiae Japonicae Caulis et Folium
- ☐ kg A5, ai ye mugwort leaf Artemisiae Argyi Folium

6.75% of the websites discussed the use of products derived from pangolins, including Facebook (see Figure 4) and specialist information sites (Figure 5).

Figure 4: Facebook page discussing pangolin-derived products





Also in Shanghai, workers at the First People's Hospital and at the Hangkou District Hospital reported on their experience in treating 43 cases of endometriosis (7). There were three formulas: one for dysmenorrhea cases, using Neiyi Fang (endometriosis formula), comprised of tang-kuei, salvia, red peony, cyperus, calamus gum, cyathula (a substitute for achyranthes), cinnamon, pangolin scales, gleditsia, lacca, zedoaria, and sparganium; one for heavy bleeding cases, using tang-kuei, bulrush, achyranthes, salvia, peony, red peony, cyperus, ophicalcite (a mineral), rhubarb, and calamus; and a node-dispersing formula for patients with large cysts, comprised of tang-kuei, salvia, red peony, cyathula, cyperus, cinnamon twig, sargassum, anteater scales, gleditsia spine, calamus, curcuma, and lacca. The formulas were further modified according to specific symptoms that were present. The effective rate was 88%, with four of the women getting pregnant. Despite the good resolution of symptoms and the finding that cysts were reduced in size, there were no cases observed in which the cysts vanished. A three-step treatment with differing herb formulas before, during, or after menstrual bleeding, was reported in 1983 (8) by workers at the Ruijin Hospital of the Second Medical College of Shanghai. For the week following menstruation, a formula comprised of cinnamon, red peony, moutan, persica, laminaria, epimedium, cynamorium, eupolyphaga, vacarria, and the traditional prescription Xiao Yao San (Bupleurum and Tang-kuei Formula; a qi-regulating formula) was used. Then, for a week during the premenstrual phase, a formula containing bulrush, pteropus, salvia, achyranthes, frankincense, myrrh, sparganium, zedoaria, and artemisia, plus pills made from sanqi (raw tien-chi ginseng) was taken. Finally, during menstruation, bulrush, pteropus, phellodendron, limestone, carbonized cyperus, linderia, cnidium, carbonized rhubarb, astragalus,

Figure 5: Information site discussing pangolin-derived products

Chuan Shan Jia - *Manis pentadactyla*

Pin Yin
Chuan Shan Jia

Latin
Squamus Manitis

Introduction [Back to Top](#)

Squamus Manitis is the scale of *Manis pentadactyla* Linnaeus (Fam. Manidae). The pangolin is captured all the year round, killed, treated with boiling water for a moment. The scales are taken down, washed clean, and dried in the sun.

Western medical [Back to Top](#)

The drug was used to treat ischemic myopspasm, premenstrual breast distension and tenderness, and mastitis, etc..

Eastern medical [Back to Top](#)

- Pattern: Activates Blood, removes stagnant Blood, restores menstrual flow, promotes lactation, treats carbuncles and discharge pus.
- Properties: Salty, slightly cold.
- Channels entered: Liver and Stomach.

Chemical constituents [Back to Top](#)

chuan shan jia mainly contains proteins and amino acids.

Pharmacological actions [Back to Top](#)

Miscellaneous effects
Water decotion-alcohol precipitation of chuan shan jia had lower the whole blood viscosity under high shear rate, improve the erythrocytic deformability, increase erythrocytic electrophoresis ability and lower fibrinogen. 10 minutes' perfusion to isolated hearts of rats could increase the blood flow in the coronary vessel by 6.7%, and increase the contraction of cardiac muscle. chuan shan jia could elevate blood cells.

Clinical Studies [Back to Top](#)

In addition to discovering the URLs, the process also produced a body of language that was found to contain key indicators of actual instances of the sale of pangolin products. These phrases include ailments (and their remedies) for which pangolin-containing products are recommended to be used; the names of brands and companies that sell these products; and euphemistic and other indirect descriptors of pangolins. Around a third of these were suggested by subject-matter experts at the initiation of the case study, but the others were identified by the DDDE process itself. Due to the obvious sensitivity of these words, they have been withheld from the public version of this report.

Ivory

The main focus of the ivory case study was to improve the ability of the DDDE to discover larger quantities of data more automatically through a series of successive cascades.

To create an initial dataset for classification, three steps were performed:

1. Keywords were extracted from an initial collection of documents.
2. Documents that were extremely large were filtered from the process because they made it much more difficult to train algorithms to be able to sort relevant from irrelevant websites.
3. The documents were segmented by sentence so that classifiers would be able to analyze each of the sentences that each of the websites contained.

A series of web searches were made using the keywords and each of these web pages was crawled. This produced an initial dataset consisting of roughly 45 000 documents.

Next, a machine learning classifier was trained to identify documents that were considered to mention the sale, purchase or discussion of ivory products. The purpose of this classifier was to simply remove any website that had been collected but did not contain content that related to ivory or elephants.

This classifier then categorized each of the 24 261 documents that had been collected. Roughly 14 853 (61%) were categorized as related to ivory or animals ivory is sourced from. There was, however, a considerable quantity of reference material that discussed the animals in question, rather than the sale of ivory.

A second classifier was therefore built to divide the documents into four different classes:

- 'Irrelevant' – not related to ivory or animals that produce it.
- 'Product description'
- 'Discussion of ivory' (i.e. preparation and uses of ivory products)
- 'Reference material' – related to ivory and the animals that produce it, but not related to products or their transaction.

The dataset was then cascaded four times. New phrases and keywords related to ivory were extracted from the dataset and used as additional terms for web searches. These were then passed through the first classifier, and documents considered relevant were then passed through the second classifier.

Table 6 shows the new pages that were collected during each crawl, and the categories that they fell into. It can be seen that new relevant pages (related to products and discussions) were consistently identified throughout each crawl.



Table 6: Categories of web pages in ivory dataset

Crawl number	Product	Discussion	Reference	Irrelevant	Total
1	2 452	6 098	9 009	6 707	24 266
2	748	2 169	3 312	59 100	65 329
3	5 113	552	10 009	21 916	37 590
4	819	966	64	39 242	41 091

The most consistent pattern, however is the reduction in unique domains as crawls continue (see Table 7). In other words, new relevant activity was found during each crawl, but from a diminishing number of domains. This has both positive and negative implications. In one sense, the crawler is focusing on a more relevant area of the internet. For instance, the two most common URL domains are eBay and boonetrading.com – two sites that were found to be predominant sellers of ivory and bone-based products. The negative side is that there could be a reduction in diversity and a loss of information.

Table 7: Diminishing number of domains in ivory searches

Crawl number	Unique domains
1	1 460
2	549
3	314
4	90

Distinguishing between legal and illegal ivory, and between ivory and other forms of horn, bone and teeth is a significant challenge when performing the analysis. Items were found for sale that purported to be made of ivory when of course they may not have been. Websites were found whose policies prohibited the sale of ivory products but accepted product descriptions or images of ivory products. And ivory was also found alongside other forms of, at least apparently, illicit commodities.

Figure 6: Description and discussion of an elephant's tusk



Figure 7: Website listing rhinoceros horn



Consistent with the pangolin case study, the DDDE identified a body of key linguistic indicators of sales-related activity involving ivory. Again, this was a combination of euphemistic descriptors of ivory, symptoms that ivory-containing products might be used to treat, and brands that sold these commodities. They have been withheld from this report.

Visualized results

Throughout the case studies, it began to emerge that the DDDE process was identifying not just individual websites, but in fact communities of websites that shared common vernaculars and interests, and that may be explicitly linked to one another too.

After the case studies were complete, the project attempted to identify and map different communities present within the results that the DDDE produced. A community detection and visualization approach was applied to, first, cluster web pages on the basis of the kind of language that they contained; secondly, to visualize these clearly; and, thirdly, to characterize the different communities that were identified.

It was hoped that finding different communities in this way would achieve two things: first, provide a different way of differentiating between relevant and irrelevant activities and, secondly, distinguish between the different kinds of relevant activity that had been found.

The clustering of the websites was performed using an open-source program called Iramuteq. This creates a list of all words that appear on each of the websites. Each website is tested for the presence or absence of each word, and, through this process, a series of 'clusters' are generated, composed of websites that generally share the same language with one another.

To do this, each word is given a numerical value, calculated by comparing how frequently it would appear in a particular class if the words were distributed at random against how frequently it actually appears in that class. This provides a measure of which words are most strongly associated with particular classes.

The visualizations that follow show these classes as coloured word clouds, composed of the words that were most characteristic of that class – that is, those that appeared more often with other words in that class than elsewhere

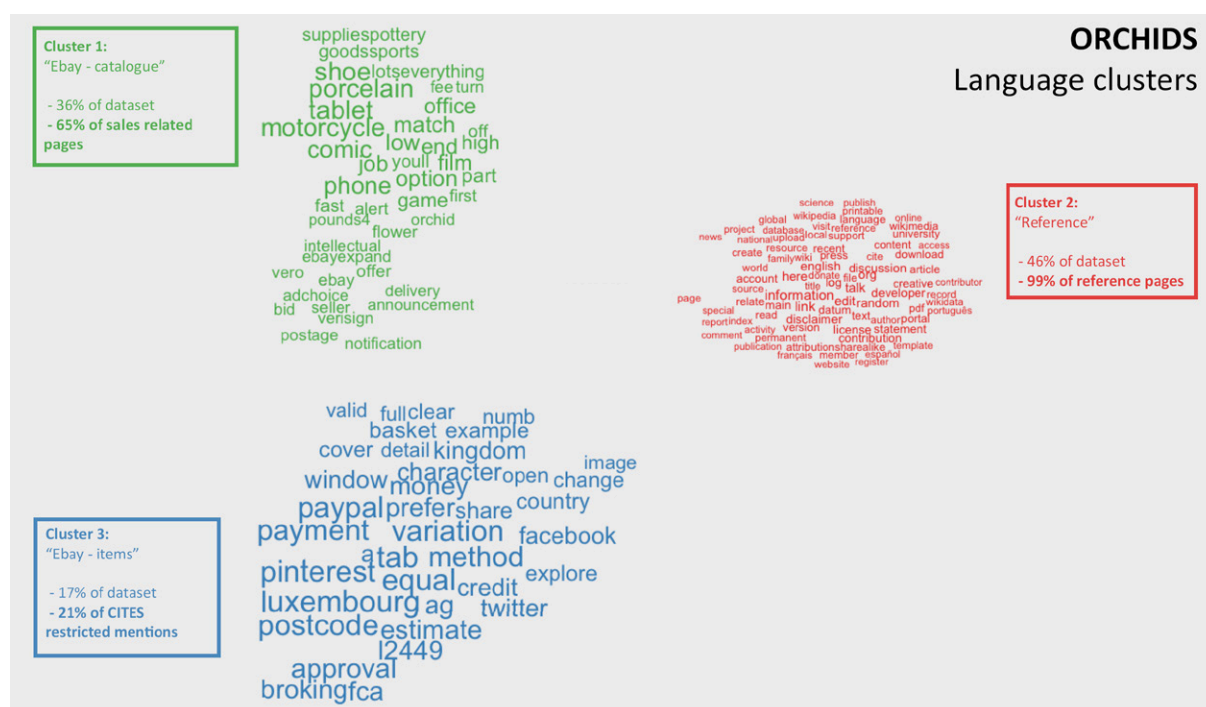
in the dataset. Classes that share terms are positioned more closely together, giving an impression of how strongly connected they are.

In the outputs below, words are coloured by class and sized by the number of times they occurred within the dataset. Clusters containing high amounts of relevant material have also been labelled according to the percentage of relevant categories, identified using Method52, that fall into them.

Case Study 1: Orchids

The orchid case study was dominated by pages from eBay and reference sites, and this made the attempt to discern clear linguistic clusters less successful. Both of these kinds of pages have high quantities of ‘boilerplate’ text that is repeated on every page, such as lists of countries and instructions like ‘add to cart’. This resulted in one cluster for reference pages, plus two separate clusters for different types of eBay pages – catalogues (i.e. lists of items) and item pages (i.e. single auctions).

Figure 8: Orchids – Clusters of language within websites produced by the DDDE



Given the dominance of eBay within this dataset, no other clear characterization of the clusters could be made. This implies that the clustering techniques developed here are more useful across datasets that have collected a range of linguistically distinct websites, such as the pangolin case study and its distinction between traditional Chinese medicine and other websites. The clustering is less useful when applied to a dataset dominated by a small number of domains with a large amount of repeated text across them.

*www.ebay.co.uk

*offer.ebay.co.uk *nbiel3.ncl.edu.tw
*zoekg.bibliotheek.be
*cgi.ebay.co.uk
*www.ebay.com.au
*www.zhiboba888.com
*orchidgarden.co.uk
*stores.ebay.co.uk
*vi.vipr.ebaydesc.com
*www.thompson-morgan.com

*www.unicat.be
*pages.ebay.com
*www.beixuezhileng.com
*www.paramountorchids.com
*vi.raptor.ebaydesc.com
*www.orchidsreprol.de
*www.slipperorchid.com
*www.schordje.com
*www.whats4eats.com
*www.propiants.com
*www.orchids.uk.com
*book.naver.com
*www.facebook.com
*www.nickyslippers.com
*www.thespruce.com
*aerialphotojab.com
*www.efforas.org
*www.gbif.org
*www.ifs.du.edu
*en.wikisource.org
*www.catalogueoflife.org
*www.wikia.com
*forgeoempres.wikia.com
*rsf.org
*legion.wikia.com
*www.ipni.org
*itunes.apple.com
*en.wiktionary.org
*www.youtube.com
*archive.is
*www.wikidata.org
*en.ce.cn
*wssp.science.kew.org
*www.orchidsmadeeasy.com
*en.wikiquote.com
*meta.wikimedia.org
*www.slipperiana.info
*www.ncbi.nlm.nih.gov
*fandom.wikia.com
*species.wikimedia.org
*www.hengduanbiotech.com
*commons.wikimedia.org
*sarawaktourism.com
*creativecommons.org
*web.archive.org
*epic.kew.org
*en.wikibooks.org
*www.mediawiki.org
*tools.wmflabs.org
*en.wikivoyage.org
*community.wikia.com
*www.plantsoftheworldonline.org
*apps.kew.org
*www.imf.org
*orchids.wikia.com
*en.wikinews.org
*wikimediafoundation.org
*www.orchid.org.uk
*ngswsweb.ars-grin.gov
*www.jstor.org
*www.citypopulation.de
*the-resident.wikia.com
*www.arkive.org
*ecomingafoundation.wordpress.com
*countrystudies.us
*www.nytimes.com
*books.google.com

*www.mofa.gov.vn
*www.unt.org
*www.cia.gov
*freedomhouse.org
*www.olympic.org
*hdr.undp.org
*www.pewforum.org
*www.state.gov
*www.ebay.fr
*www.webometrics.info
*www.abebooks.com
*www.edg.admin.ch
*www.tropicos.org
*www.prisonstudies.org

Five significant clusters were identified from the pangolin data. Cluster 4 was primarily related to academic discussions about pangolins, and cluster 5 to conservation-related discussions, and neither contained significant numbers of URLs related to sales. Clusters 1 and 3 were more ambiguous, containing some discussions of crafts and medical symptoms, respectively, both of which could be related to sales. Cluster 2 was the most significant cluster, however: more linguistically distinct from the others, it was a small cluster, representing just 12% of the URLs in the dataset, yet it contained 79% of sites identified as related to sales. This cluster was dominated by URLs related to traditional Chinese medicine.

Cluster 1:
"Crafts and sales"

- 37% of dataset

Cluster 2:
"Traditional Chinese medicine"

- 12% of dataset
- 21% of identified brands
- 79% of identified seller sites

Cluster 3:
"Medical"

- 18% of dataset

Cluster 4:
"Academic"

- 20% of dataset

Cluster 5:
"Conservation"

- 13% of dataset

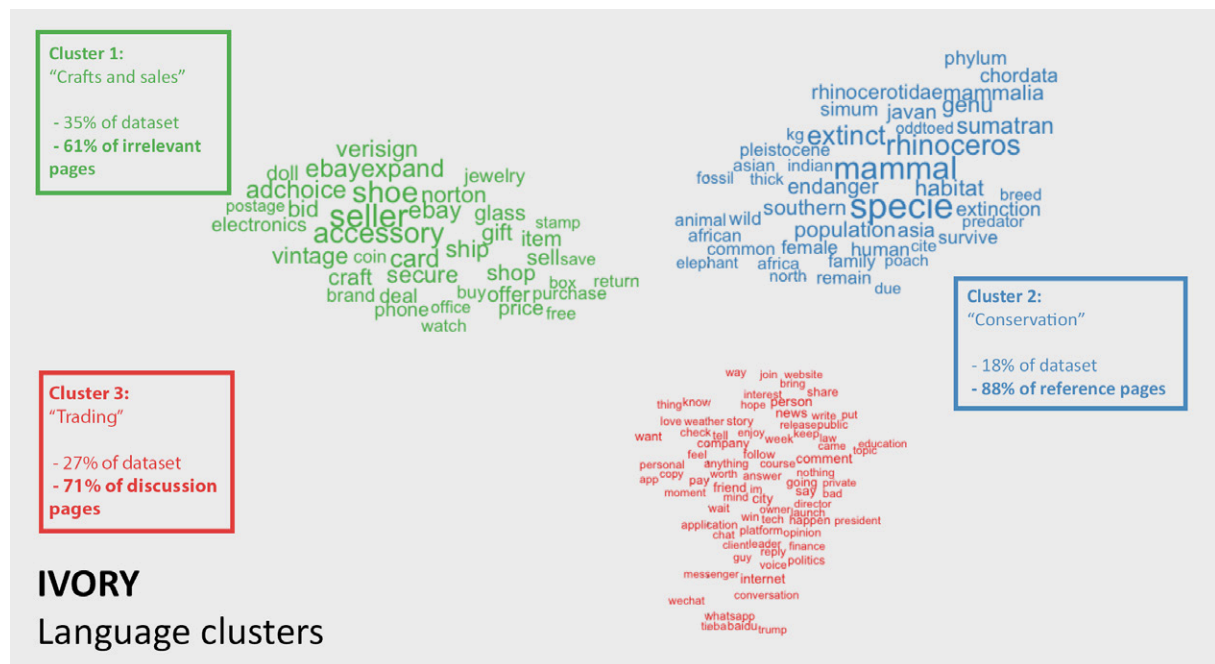
PANGOLINS
Language clusters

[illegible]

Case Study 3: Ivory

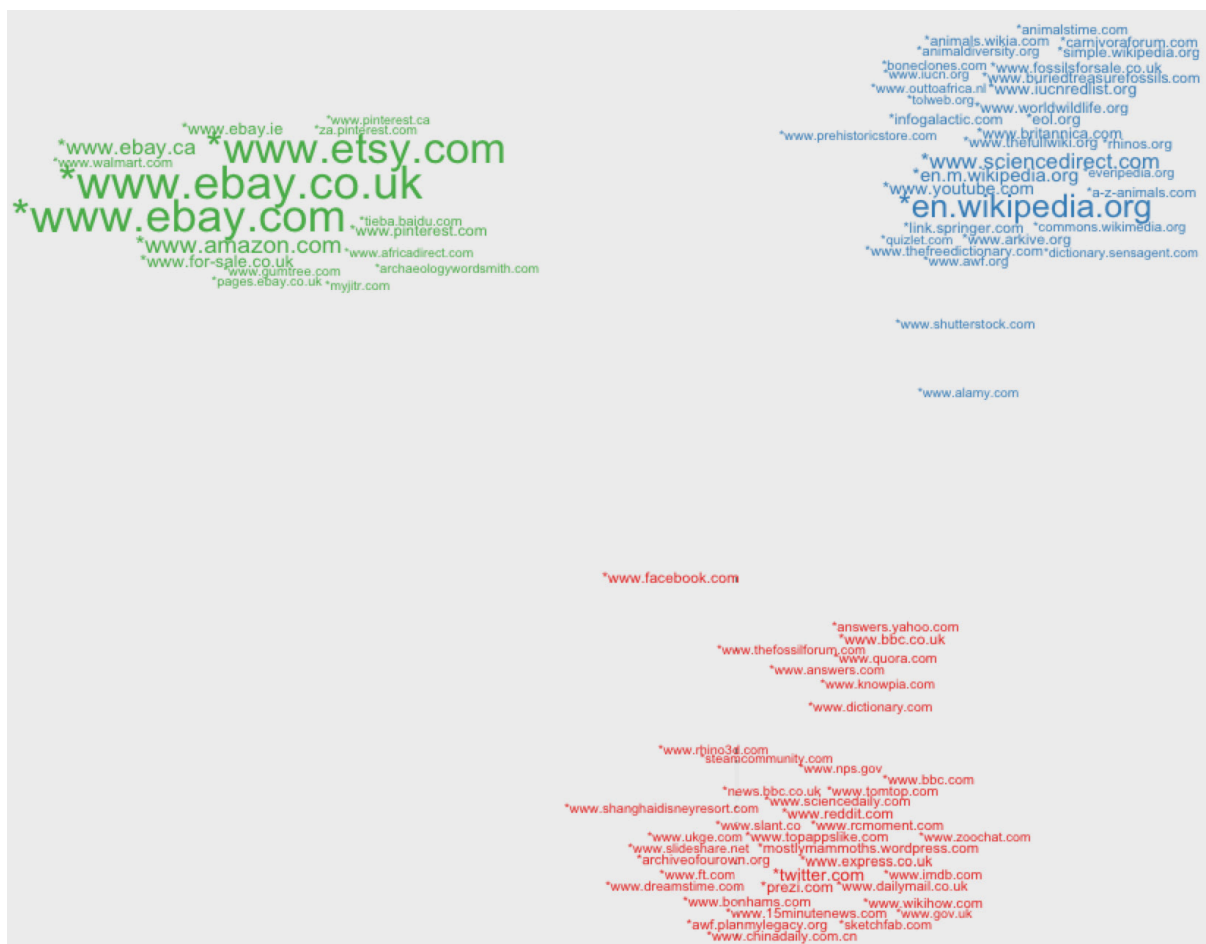
The URLs gathered during the ivory case study were separated into three different clusters. Cluster 3 was related to discussions around conservation, and largely irrelevant. Cluster 1, the largest, also contained the highest quantity of pages considered irrelevant for the purposes of this project. Cluster 2 identified the most pages related to 'discussion' – i.e. people talking about looking after, trading, finding and maintaining ivory and ivory products. This was a completely different conversation from the purely sales-related conversations that the project initially sought to find but, nonetheless, it was considered by researchers to be a significant and relevant finding.

Figure 12: Ivory – Clusters of language within websites produced by the DDDE



Once the clusters were represented by domain, it was clear that Wikipedia was the most common domain in cluster 3. A number of e-commerce platforms dominated cluster 1, while cluster 3 was more diverse, as it was not dominated by any particular domain but comprised a combination of news sites, blogs and discussion forums.

Figure 13: Ivory – Clusters of website domains produced by the DDDE



Strengths and limitations of the DDDE

Overall, the DDDE was able to identify a reasonably large number of URLs, in absolute terms, that either facilitate the sale of a CITES-listed animal or plant, or involve discussion around commodities that contain them.

It is hoped that the outputs from this process can complement and strengthen more qualitative, case-specific and immersive forms of research on online environmental crime. These forms of research are able to produce highly relevant data, which the DDDE can then use as seed documents to find many more additional examples that are similar. Likewise, the outputs of the DDDE must be interpreted and contextualized by subject matter and domain experts in order to be meaningful.

The identification of online environmental crime was found to be well suited to the language-driven way that the DDDE works. The overall process sets out to find keywords and key phrases that best identify online IWT activity, and a key subset of these are code words, or a body of language intended to mask this purpose. This was especially clear in the pangolin case study, where these kinds of code words were discovered – either for brands (i.e. Plum Flower Brands) symptoms ('blood invigoration', 'ischemic myospasm') or non-specific descriptions of the product itself (Chuan Shan Jia Anteater Scales). Although this kind of language is intended to camouflage the nature of the illicit activity, the project found that, conversely, it was highly indicative of it. Most of the code words in the case studies

were identified by the DDDE process; however, some were given to the research team by subject-matter experts as part of the initial process of gathering seed documents.

A key weakness of the DDDE process is that it cannot currently produce only URLs that are relevant. In all three case studies, a substantial amount of ‘noise’ within the dataset has remained. Therefore, despite the large absolute number of relevant URLs identified, they still constitute a relative minority within the final overall datasets.

Until techniques to remove this noise become more reliable and effective, it would be difficult for a human analyst to be able to act on the data that the DDDE produces. To become actionable or operationally useful, it is vital that the process becomes better at guiding human end-users in triaging and prioritizing the information that it produces. The visualization techniques have shown promise in being able to identify and separate relevant and irrelevant groups of URLs within the dataset, however. And, although this proved to be more useful in some datasets than others, linguistic clustering for community detection is likely to be a central technique in removing the noise that this process creates.

The potential to cascade was technically proven in principle, but it has not been shown that it can actually track a phenomenon as it changes – that would require a longer-term application and greater collaboration with qualitative work. The potential of the DDDE to work beyond the English language has also not been proven.

The project team was unable to develop a DDDE that was able to work across different domains. Although this was attempted, it was decided that developing different applications of the DDDE on a case-by-case basis was more likely to succeed, and that it may eventually be possible to connect these processes together to produce one that can cover a larger number of species.

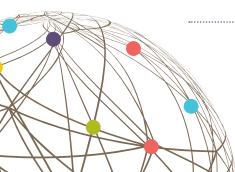
The view of the authors is that this technology shows promise, but is also a work in progress, and, to be useful, requires additional work in a number of different areas. Additional analytical work is required to increase the precision of the outputs. Work is also likely required to differentiate between the very different kinds of websites that the DDDE identifies, and to analyze them in different ways. The DDDE may also be applied more widely in a number of different ways to understand its efficacy across a range of different use cases: additional CITES-listed species, in languages other than English and over a longer period of time.

Glossary of terms

Code words – a form of keywords that the study identified. The *prima facie* intent of this language is to indirectly or euphemistically refer to illicit activity or commodities.

Cyclical extension – a technique used during the first phase of the DDDE process (*see below*) to increase the number of documents included within the initial dataset. It is a manual process whereby seed documents are read to identify keywords that are then used as the terms for a small number of web searches. The URLs returned from each of these web searches are compiled and ranked on the basis of the number of times that each of these different web searches found them. This process is cyclical because it is repeated a number of times, and each time the websites are manually inspected for new relevant phrases, which are then used as queries within successive web searches.

DDDE (Dynamic Data Discovery Engine) – the six-stage process that this paper describes. It combines a number of different technologies and techniques, described in this glossary, to attempt to allow an analyst to expand a potentially small, pre-existing series of examples of online sources related to the transaction of CITES-listed species into a larger body of examples. The first stage involves building an initial corpus. The second uses feature or phrase extraction to discover the phrases that characterize the original corpus. Thirdly, these features are combined to create search phrases; web searches are performed; and the results of these searches are collected. Next, a series of language-based



forms of analysis are used to separate the relevant from the irrelevant results in these search queries. Then, more qualitative forms of analysis are performed on the relevant outputs to produce a more textured, detailed appraisal of the data collected. In the final, sixth stage, the data is visualized. The process is recursive, where the outputs of one iteration are used as inputs for the next, allowing an analyst to return to any stage in the pipeline and update, add or filter the data further. This cyclical nature of the DDDE is intended to allow it to evolve and improve over time.

Domain mapping – a technique used by the DDDE that entails separating the dataset collected by the DDDE into the different domains that make it up – for instance, social-media platforms, e-commerce sites, reference sites, such as Wikipedia, discussion forums, and so on.

Iramuteq (interface of R for multi-dimensional text and questionnaire analysis) – an open-source software tool used to visualize the data that the DDDE produced. The software was used to compare the language contained in each website collected by the DDDE with all the other websites also collected. Websites with similar language were visually represented as closer together. This is a form of ‘community detection’, where groups of linguistically similar websites are shown as clusters on the visualized outputs.

Keywords and key phrases – words or combinations of words that, in the context of this study, occur significantly more often in activities related to the transaction of the species in question than they do in activities related to just descriptions of the species.

Keyword analysis/annotation – identifies whether a specified word or collection of words is present in any given document, and then categorizes the document on the basis of the presence or absence of this word. Within the DDDE process, this is used to identify different attributes of the documents collected, such as whether they contained mentions of known sellers, likely purchasing-related discussions and so on.

Language annotation – a technique used during the DDDE process, whereby a pre-trained algorithm analyzes the content of each of the websites collected, and categorizes them according to the primary language that they contain.

Machine learning classification – one of the core techniques used in the DDDE, and a key technological emphasis of the Method52 software (see below) that was used to conduct it. It categorizes websites using a body of technology called natural language processing. Natural language processing (NLP) is a branch of artificial intelligence that attempts to teach computers to recognize important differences in language as it naturally occurs through human discourse. Within the context of this project, Method52 was used by analysts to train a number of NLP classifiers. An analyst manually reads and places a number of randomly selected websites (normally numbering in the hundreds) into a number of categories that they defined. The classifier attempts to find, through these examples, the language that most closely correlates with each of the categories, and abstract a series of rules that are used to classify additional documents the analyst has not seen. This allows thousands (or hundreds of thousands) of documents to be placed into the categories defined by the analyst to some measurable degree of accuracy. Given that this form of analysis was one of a number that the DDDE used, it was decided to report on the precision of the process overall, rather than the accuracy of any specific stage within it.

Method52 – a software suite owned and maintained by CASM Consulting. A program and development environment, it has been created to allow researchers to conduct analysis of large, unstructured datasets, especially those drawn from digital and social-media sources, in sociologically robust ways. The software uses a graphical interface to allow researchers to build analytical ‘pipelines’, or strings of components, which perform a different analytical task or utility. The overall aim of Method52 is to equip analysts from outside of formal data-science backgrounds with the tools to collect, analyze and visualize datasets that are too large to be manually analyzed.



Seed documents – websites gathered during the first stage of the DDDE process. These are web pages that contain activity considered relevant to the project’s overall aims – in this instance, of transaction-related activity involving ivory, pangolins or CITES-listed orchids.

Web scraping and crawling – techniques to extract data from websites. Software is used to copy the data contained on a website and reproduce it in another form. Within the DDDE process, these techniques are used to reproduce websites as documents that are then analyzed. The DDDE process used a ‘crawling’ technique where websites hyperlinked to the primary sites of interest are also identified and scraped.

Acknowledgements

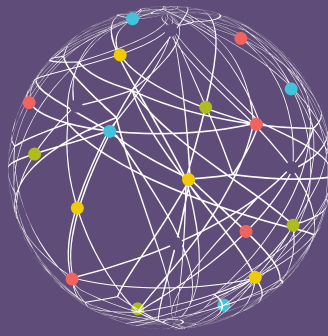
The authors would like to thank the Government of Norway for funding this report. Digital Dangers forms part of a partnership project between INTERPOL and the Global Initiative Against Transnational Organized Crime, in cooperation with the UN Office on Drugs and Crime.

Thanks to Dr Amy Hinsley and Dr Dan Challender, Oxford Martin Fellows, Oxford Martin Programme on the Illegal Wildlife Trade, for their expertise relating to the online trade in orchids and pangolins, respectively.

About CASM Consulting LLP

CASM Consulting develops commercial-grade technology in social-media analysis and natural language. It comprises 10 partners and associates, plus other consultants and partners.





THE GLOBAL INITIATIVE
AGAINST TRANSNATIONAL
ORGANIZED CRIME

www.globalinitiative.net



A NETWORK TO COUNTER NETWORKS

