06.08.2021
CNRTR

Dr Stuart E. Middleton
University of Southampton
Electronics and Computer Science

UNIVERSITY OF Southampton

Royal Botanic Gardens Kew

Border Force

OUR PROJECT

Economic and Social Research Council

- Better understanding of the Internet-facilitated online trade in endangered plants.
- Better understating of law enforcement challenges and needs.
- Develop and test a digital resource to assist law enforcement, researchers and other relevant stakeholders.

Introduction to AI

Definitions

- Artificial Intelligence (AI) >> Machines that learn or problem solve
- Natural Language Processing (NLP) >> Machines that process natural text or speech
- Machine Learning >> Machines that given a task and eval metric learn from experience
- Socio-Technical >> Machines and humans in the loop

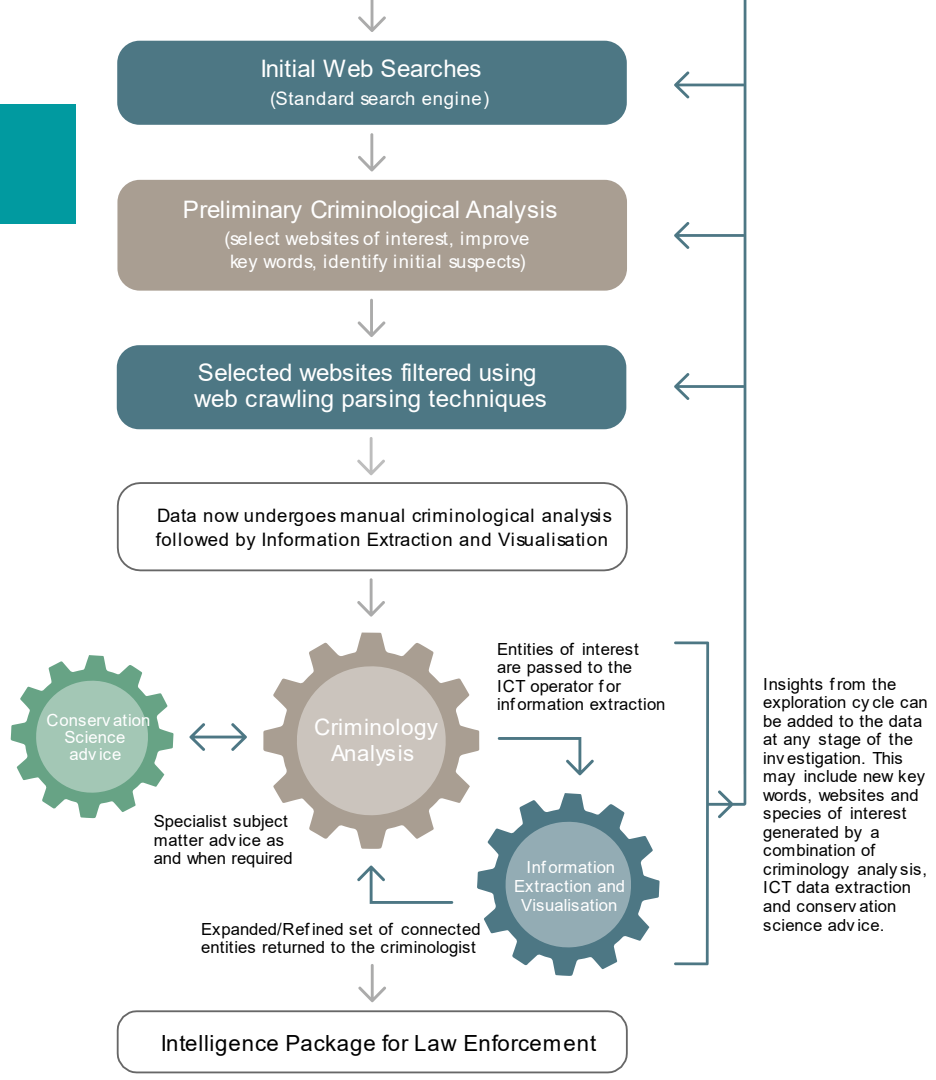
AI Landscape around cybercrime research

- Image/Video >> Facial recognition; Object and activity detection
- Audio >> Audio event detection
- Text >> Author attribution and profiling; Information extraction;
- Data >> Social network analysis; Data mining; Predicting policing
- Socio-technical >> All of the above plus Visualization; Interaction; Summarization

Step-by-Step Guide

Cyclic methodology

- Human in the loop with AI tools
- Cycles
  - Criminology analysis
  - Information extraction
  - Visualization
- Refine final intelligence package



Step-by-Step Guide

Criminology Analysis

- Subject matter expert >> Lexicon (species, trade jargon)
- Criminologist >> Manual browsing and behaviour coding
- Posts about illegal trades and CITES permits
- Subcultural examples of relevant forum user behaviour
- Coding of posts using NVivo >> Update target suspect lists
- Browsing >> Update web domain lists

Web Search

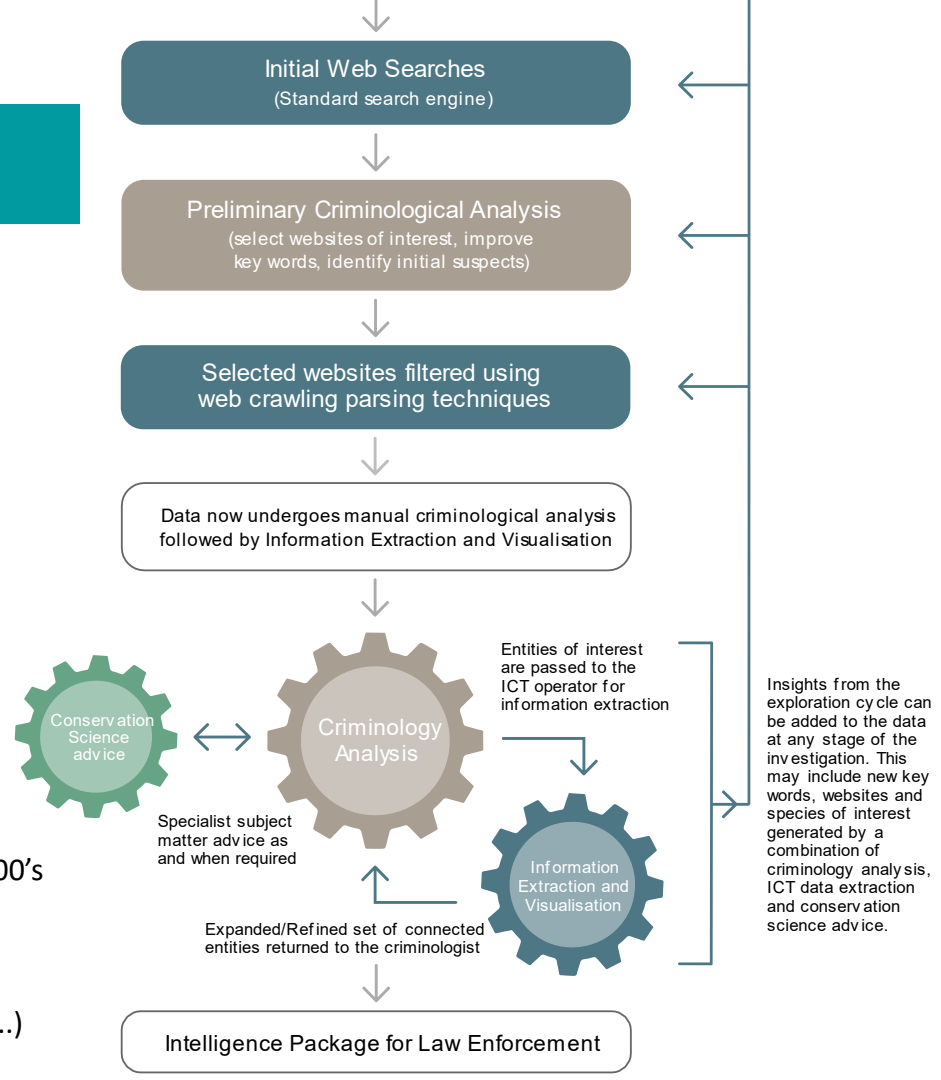
- Keywords >> Microsoft Bing >> Web pages + domains
- Find search page >> Domain + Entry URI

Crawler

- DARPA MEMEX Undercrawler
- Domain + Entry URI >> Template >> Crawl >> HTML x 10,000's

Parser

- Python3 app using html.parser and/or bs4
- HTML >> XPath pattern >> Text & Metadata (author, date ...)



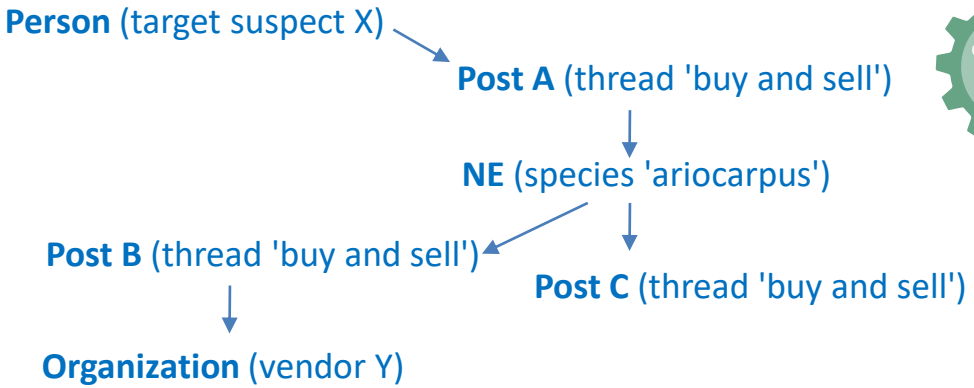
Step-by-Step Guide

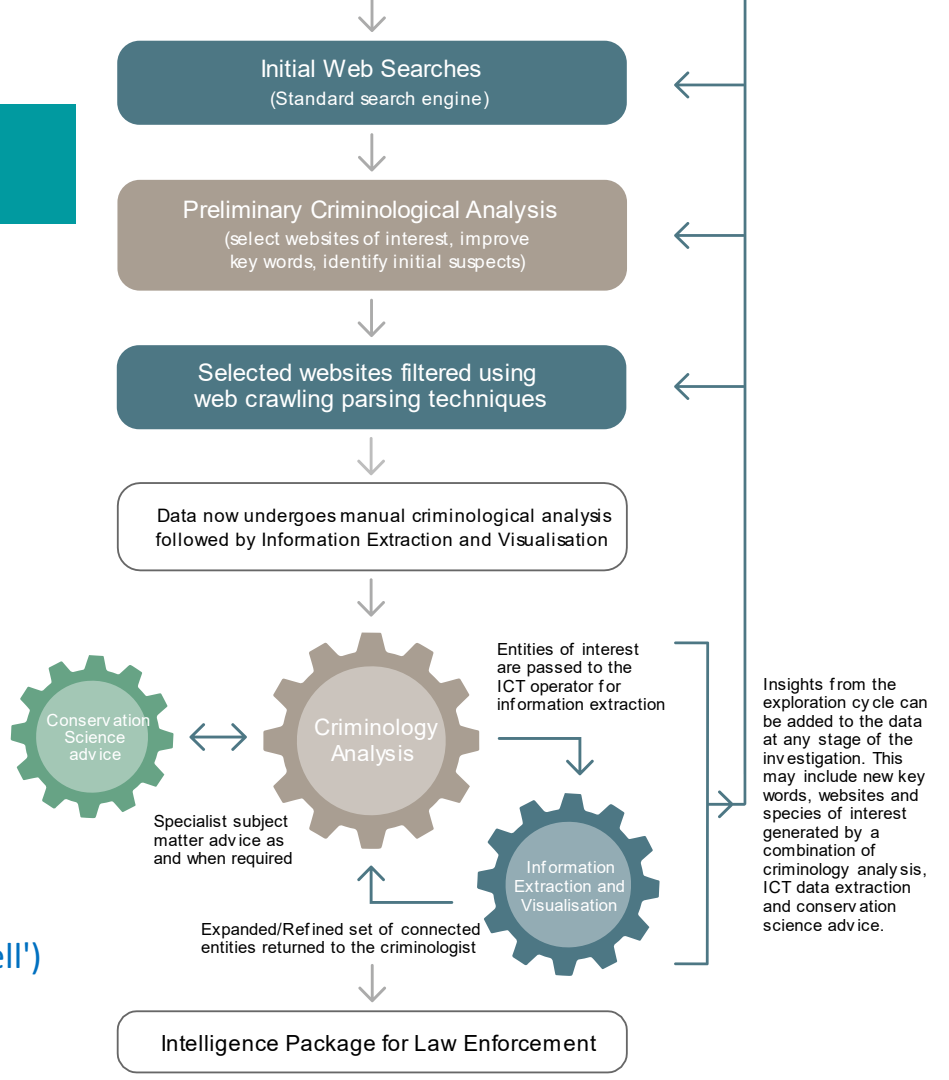
Information Extraction

- Text >> Stanford CoreNLP >> Named Entities
- Named Entities (NE)
  - Person, Location, Time, Org, Species ...
  - People, Objects, Locations, Events (POLE)

Directed Named Entity (NE) Graph

- Text, Metadata, NE >> Directed Acyclic Graph (per target)
- Discussion thread preserves conversation context





Step-by-Step Guide

Visualization

- N x target suspects >> N x Directed NE Graphs
- Python matplotlib & networkx >> Visualization
- Colour coded node types (people, places ...)
- Up to 400 entities on each graph
- Zoom, Pan, Save image to file

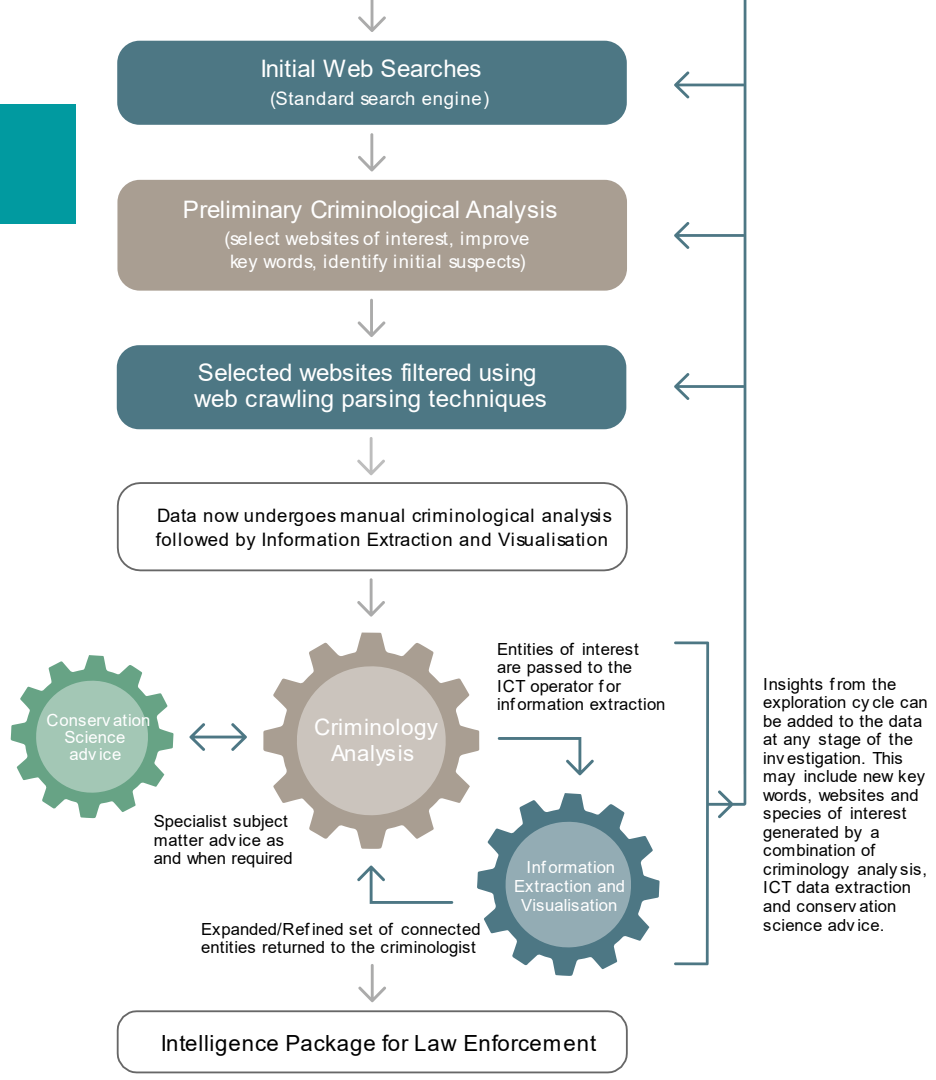
Start a new cycle ...

- Visualization >> Criminology Analysis
- Refine Targets >> Crawl / Graph / Viz

... until results are judged good enough

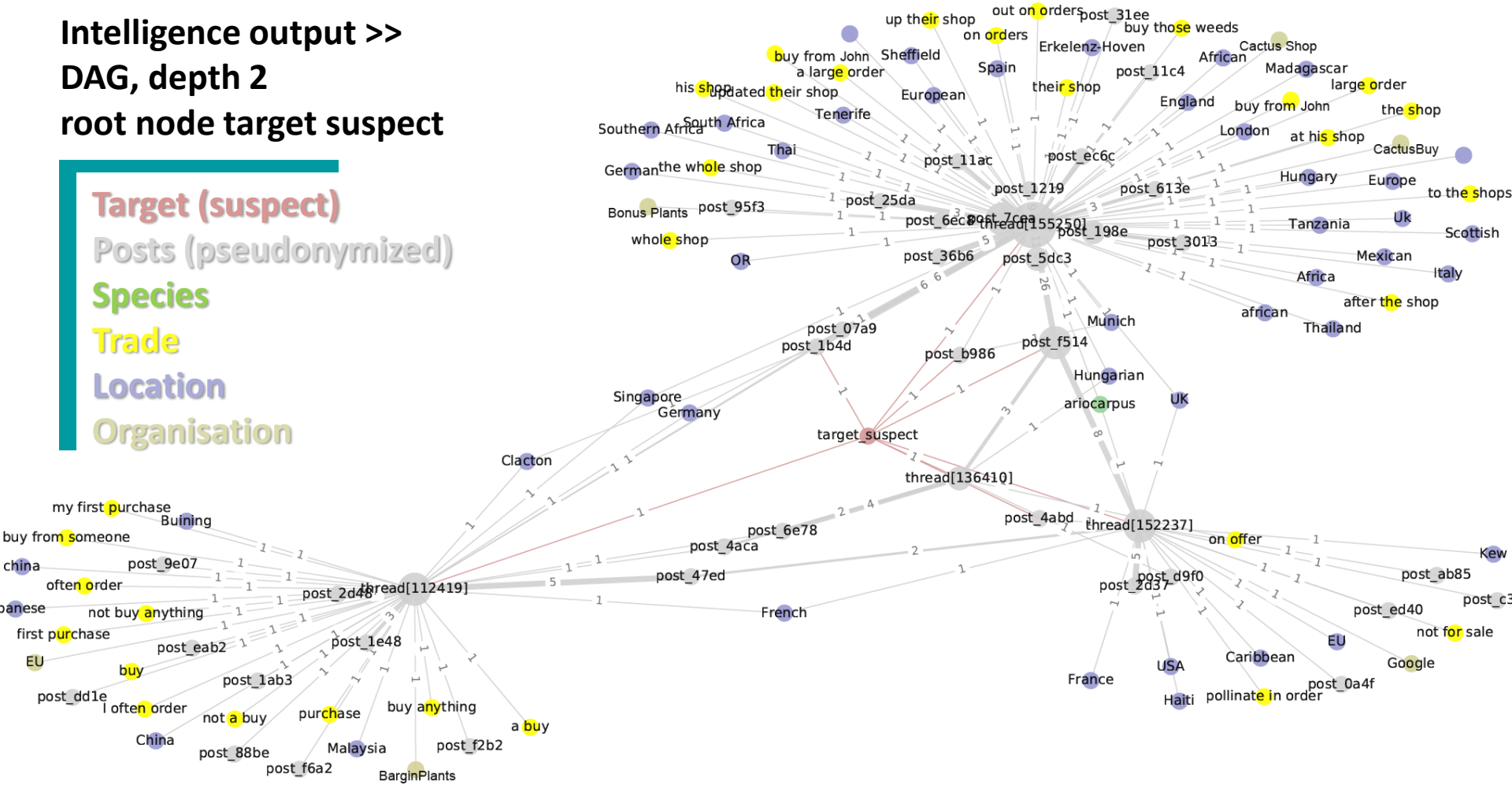
Intelligence package

- Criminologist prepares summary of evidence



Intelligence output >>
DAG, depth 2
root node target suspect

- Target (suspect)
- Posts (pseudonymized)
- Species
- Trade
- Location
- Organisation



Open-access Software Tools

- Intelligence Graph Visualization
  https://github.com/stuartemiddleton/intel_viz_entity_graph
- DARPA MEMEX Undercrawler
  https://github.com/TeamHG-Memex/undercrawler
- Stanford CoreNLP
  https://stanfordnlp.github.io/CoreNLP
- Example of a team makeup
  Criminology/Conservation expert + Computer Science intern + software tools

Key Results

Target suspects were mostly found hiding in the long tail of forums

NE graph visualization best for people and locations

Socio-technical AI methodology

- Scale up analysis with AI >> Review orders of magnitude more suspects and POLE entities
- Human in the loop >> Check automated results, look at context, make judgements

Discussion of Limitations

Data: Surface and dark web, but not deep web

- Deep web pages are computed from databases on the fly
- Hard to index by search engines, so forum / marketplace posts might be missed

Legal: NE graph evidence ready for a court of law? Not on its own

- It could be used to support target-focussed evidence packages though, with input from criminologist and maybe some relevant out of band corroborating evidence

AI: NLP can find potential suspects but human analysis needed to check suspects

- Socio-technical cycles work >> NLP + human criminological analysis
- Human in the loop AI >> Deep learning NLP where human is part of the model training method
- Wider applications >> Multi-lingual cybercrime forums (hacking, malware), online harms (cyberbullying), community moderation (trolling), disinformation (fake news bot detection)

Open Source Content

Website T&C's

- Public web search >> Find open content without login or signing website T&C's
- Websites login T&C >> Each website T&C's needs a detailed read (crawling, use, research)
- Social media API >> Facebook, Twitter, YouTube etc. >> API's come with rate limits and T&C's

Law enforcement only use open source content (unless they have a warrant)

- We follow a similar approach, avoiding private groups where a deceptive study would be needed
- This limits the coverage we can achieve for our analysis

Open Source Content

Images and videos are a minefield - avoid if you can

- We kept crawling to text and metadata only
- It is illegal in UK to store certain types of images (e.g. child abuse images) >> even accidentally
- Mental health for researchers >> images and videos can be disturbing (duty of care to staff)

Personal data captured when crawling

- Public web pages have made content manifestly public
- Personal data >> impossible to avoid crawling personal data; can pseudonymize analysis results
- Special category data >> limit accidental download by manual choice of websites to crawl
- Manual analysis provides an opportunity to flag up (and remove) special category data
- Ethical approval for study >> University ethics review panel for each study

GDPR and DPIA covered in next section (David Whitehead - Kew)

Pseudonymization

Pseudonymization and anonymization

- Browsed HTML cannot be pseudonymized (metadata too complex to remove personal data)
- Manual analysis results can be pseudonymized manually (after analysis usually)
- Crawled HTML cannot be pseudonymized (metadata too complex to remove personal data)
- JSON parsed content cannot be pseudonymized (10,000+ volume too large to manually check)
- NE graphs cannot be pseudonymized (removing personal data defeats the purpose of analysis, which is focussed on people engaging in illegal wildlife trade)
- Visualizations and infographics can be pseudonymized for publication (e.g. hash usernames)
- Statistical summaries of datasets can be anonymized

Not perfect >> Search/replace for names; Regex; NER models for PERSON type

UK Ethics panels >> 98% pseudonymized is not pseudonymized is a typical view

AI Reviews and Training

Training (law enforcement) on the future use of AI tools

- Formal AI reviews for deployments
- Help decision makers understand both the capabilities of AI and its capacity for error/bias (e.g. understanding if there is training set bias in AI tools being deployed).
- Collaborations with independent academic researchers to advise on AI reviews

Contact Info

Dr Stuart E. Middleton

University of Southampton, Electronics and Computer Science

email: sem03@soton.ac.uk

web: www.ecs.soton.ac.uk/people/sem

twitter:@stuart_e_middle

FloraGuard: http://floraguard.org/

This work was supported by the Economic and Social Research Council (ES/R003254/1)

Reflections on web crawling techniques from a conservation science perspective

David Whitehead

Royal Botanic Gardens, Kew

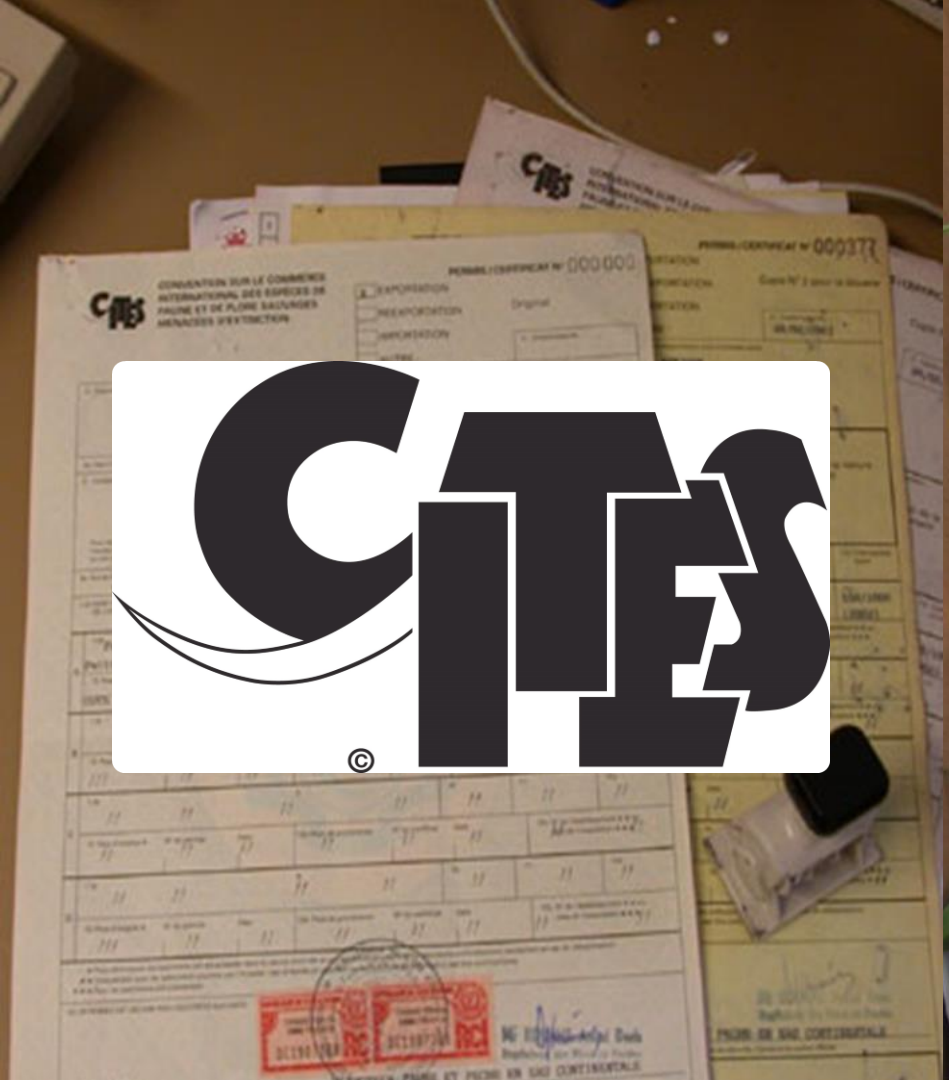
CITES@kew.org

Royal Botanic Gardens Kew

Flickr/Gary Nored

Carly Cowell

Many plants are illegally harvested and traded online. Plant poaching is leading to local or species extinctions.


CITES


© UK Border Force

Challenges include:

- Online anonymity
- Avoidance of border controls
- Spans legal jurisdictions
- Volumes of online content
- Online privacy considerations

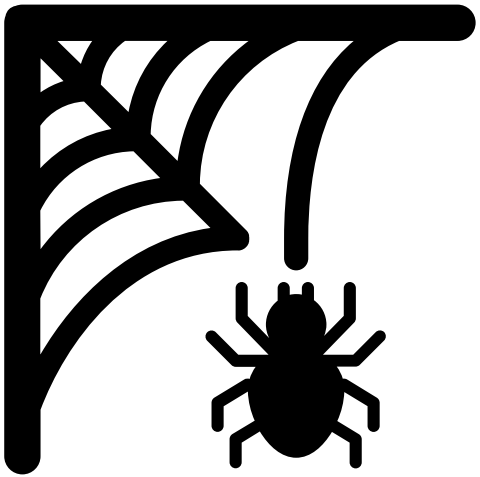
Unsplash.com

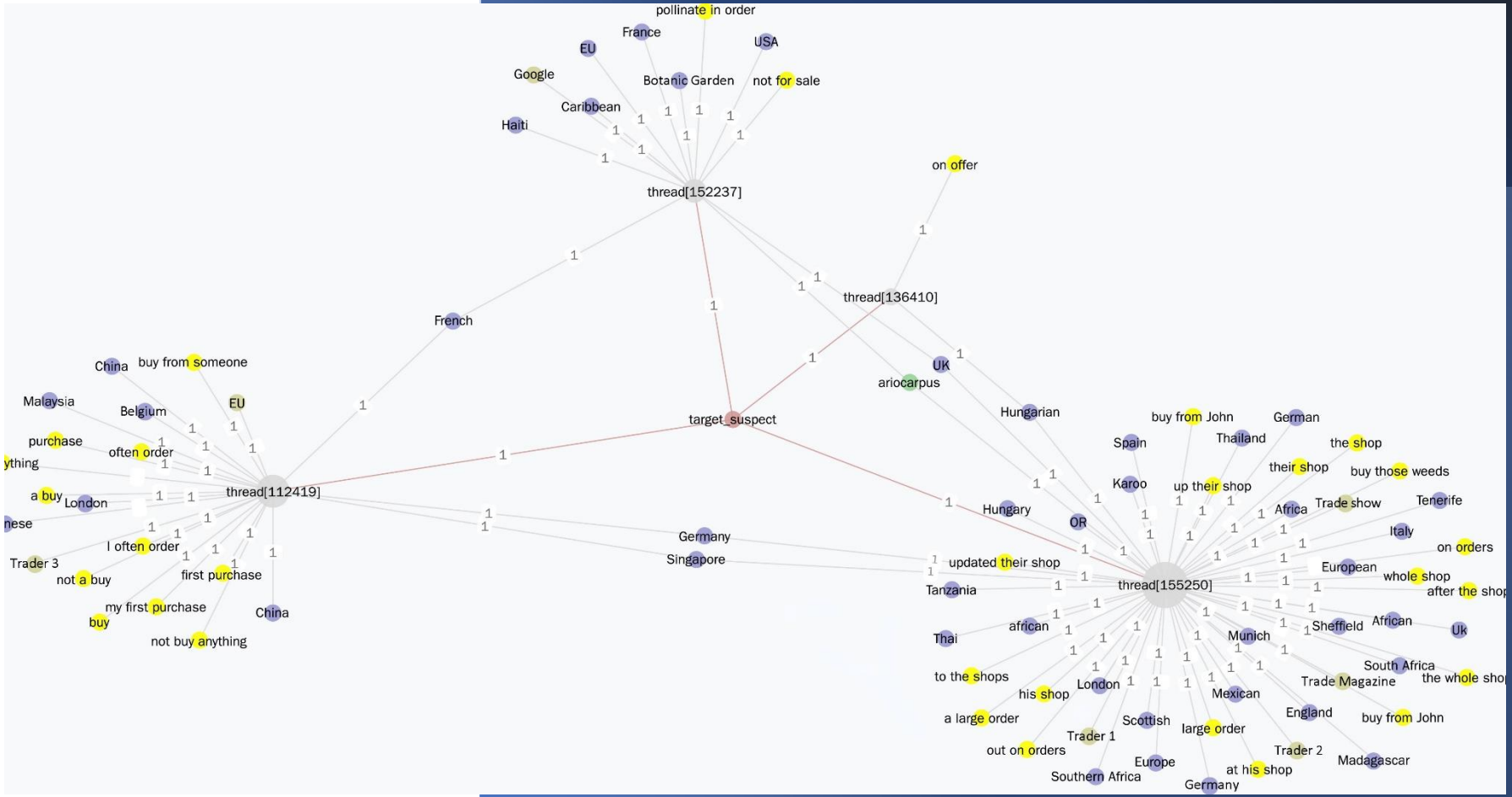
Floraguard.org

Species Names	Excluded terms	Buy/sell terms
Ariocarpus	seed	web+buy
Ariocarpus+agavoides	seeds	internet+buy
Ariocarpus+bravoanus		buy
Ariocarpus+hintonii		buy+online
Ariocarpus+kotschoubeyanus		order
Ariocarpus+kotschubeyanus		sale
Ariocarpus+kotschobeyanus		selling
Ariocarpus+kotsch		purchase
Ariocarpus+confusus		live plant
Ariocarpus+albiflorus		swap
Ariocarpus+retusus		
Ariocarpus+scaphirostris		
Ariocarpus+ scapharostroides		
Ariocarpus+scapharostrus		
Ariocarpus+trigonus		
Tamaulipas living-rock		
Tamaulipas living rock		
Nuevo leon living-rock		
Nuevo leon living rock		

Common misspellings / alternative spellings can be included.

Examples of Latin and Common names for species of interest.




pollinate in order
France
EU
USA
Google
Botanic Garden
not for sale
Caribbean
Haiti
thread[152237]
on offer
thread[136410]
French
UK
ariocarpus
target_suspect
China
buy from someone
Malaysia
Belgium
EU
purchase
often order
thread[112419]
a buy
London
I often order
Trader 3
not a buy
first purchase
my first purchase
buy
China
not buy anything
Germany
Singapore
Hungarian
buy from John
German
Spain
Thailand
the shop
their shop
buy those weeds
Karoo
up their shop
Africa
Trade show
Tenerife
Hungary
OR
Italy
on orders
updated their shop
European
whole shop
Tanzania
thread[155250]
after the shop
african
Sheffield
African
Uk
Thai
Munich
South Africa
to the shops
London
Trade Magazine
his shop
Mexican
England
a large order
Scottish
buy from John
large order
Trader 1
out on orders
Trader 2
Madagascar
Europe
at his shop
Southern Africa
Germany

- Lavorgna, A., et al. (2020). FloraGuard: tackling the online illegal trade in endangered plants through a cross-disciplinary ICT-enabled methodology. *Journal of Contemporary Criminal Justice* p. 1043986220910297.

- Lavorgna, A. and Sajeva, M. (2020). Studying illegal online trades in plants: market characteristics, organisational and behavioural aspects, and policing challenges. *European Journal of Criminal Policy and Research*. Online first.

- Lavorgna, A., Rekha, G.S. From horticulture to psychonautics: an analysis of online communities discussing and trading plants with psychotropic properties. *Trends Organ Crim* (2020).

- Middleton, S. E., et al. (2020). Information extraction from the long tail: A socio-technical AI approach for criminology investigations into the online illegal plant trade. *STAIDCC20: 1st International Workshop on Socio- Technical AI Systems for Defence, Cybercrime and Cybersecurity,* ACM WebSci 2020. Southampton.

- Whitehead et al, (2021). Countering plant crime online: Cross-disciplinary collaboration in the FloraGuard study, Forensic Science International: Animals and Env. 1

- FloraGuard Report - https://www.kew.org/science/our-science/publications-and-reports/science-reports

“All data are people, unless proven otherwise”

Zook et al, Ten simple rules for responsible big data research, PLOS Computational Biology, 2017

“Paphiopedilum canhii is now an endangered species with a precarious future. And there we have it: extinction at the speed of the Internet”

Averyanov et al, Field Survey of Pahpiopeilum canhii: From discovery to extinction, 2014

Planning an online investigation

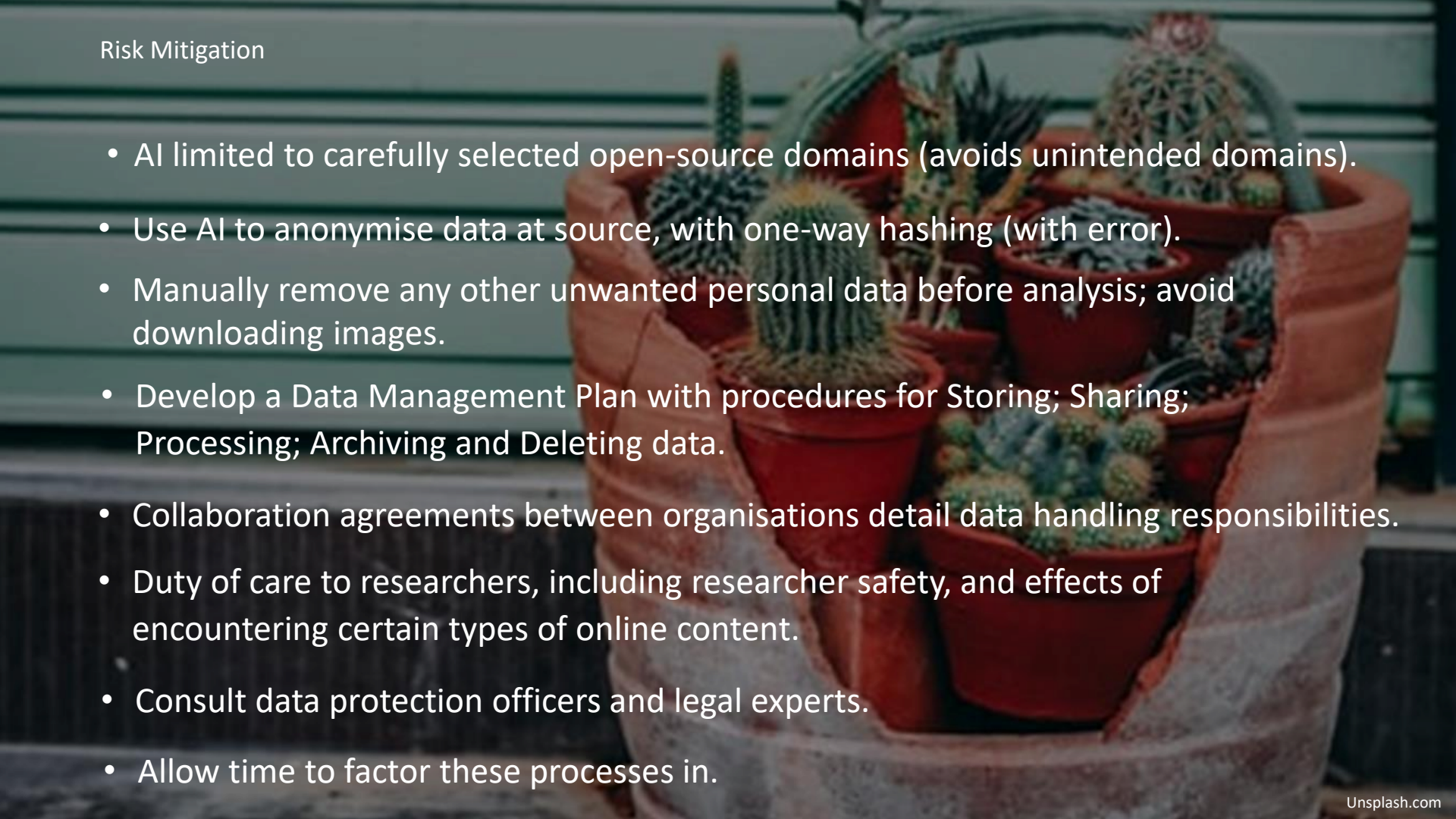
- Institutional context
- Perform a risk assessment
- Digital Protection Impact Assessment (DPIA)

Digital Protection Impact Assessment (DPIA)

- The why, what, where and how of your data collection.
- What categories of data will you be collecting, or could potentially collect?
- What are your legitimate interests for collecting the data?
- Seek specialist IT advice on hardware/software.
- Templates and advice available at ico.org.uk

Risk Mitigation

- AI limited to carefully selected open-source domains (avoids unintended domains).
- Use AI to anonymise data at source, with one-way hashing (with error).
- Manually remove any other unwanted personal data before analysis; avoid downloading images.
- Develop a Data Management Plan with procedures for Storing; Sharing; Processing; Archiving and Deleting data.
- Collaboration agreements between organisations detail data handling responsibilities.
- Duty of care to researchers, including researcher safety, and effects of encountering certain types of online content.
- Consult data protection officers and legal experts.
- Allow time to factor these processes in.

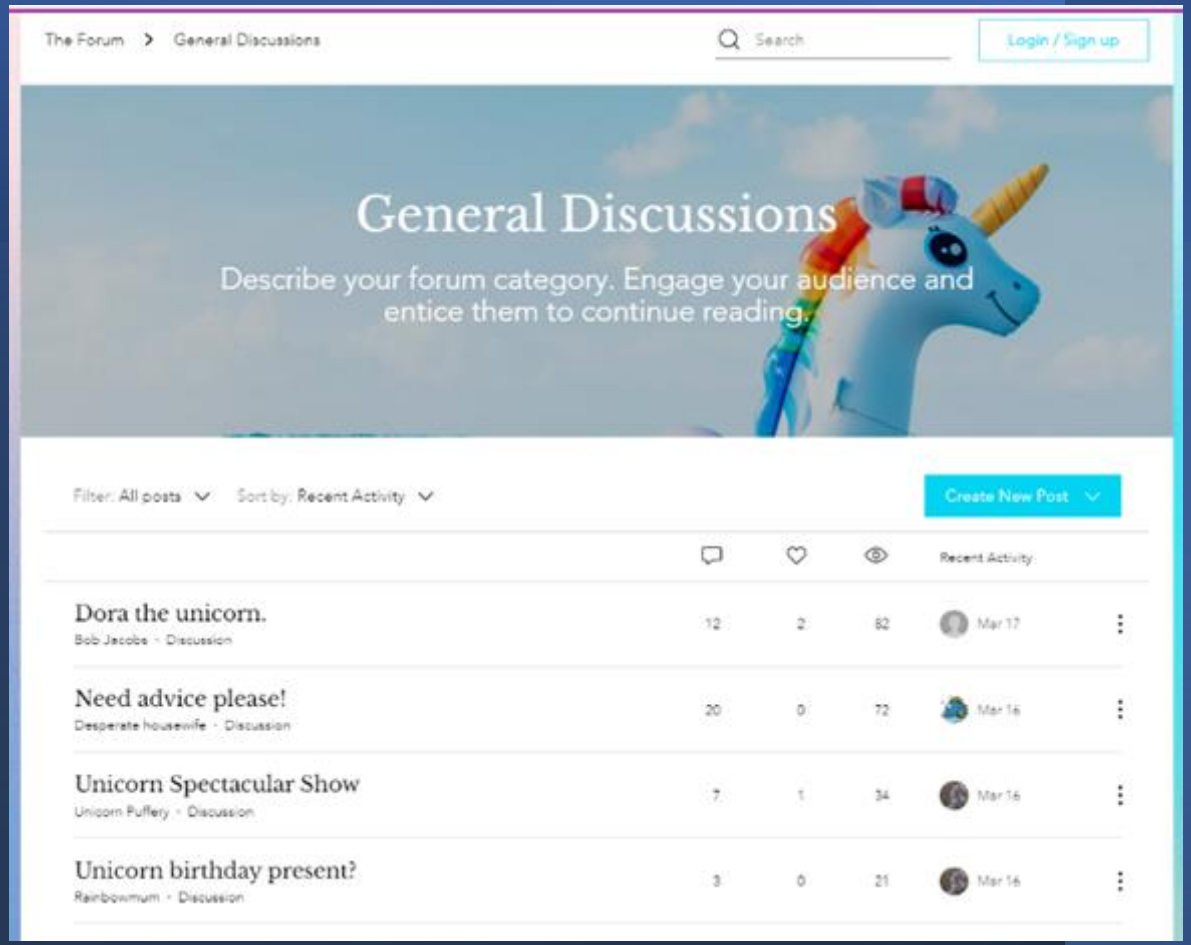


Unsplash.com

How to build capacity?

- Multi-disciplinary teams
- International partnerships must meet adequacy requirements under UK law.
- Training conducted using simulated IWT data.

Unsplash.com


The Forum > General Discussions
Search
Login / Sign up
General Discussions
Describe your forum category. Engage your audience and entice them to continue reading.
Filter: All posts
Sort by: Recent Activity
Create New Post
Recent Activity
Dora the unicorn.
Bob Jacobs · Discussion
12
2
82
Mar 17
Need advice please!
Desperate housewife · Discussion
20
0
72
Mar 16
Unicorn Spectacular Show
Unicorn Puffery · Discussion
7
1
34
Mar 16
Unicorn birthday present?
Rainbowmum · Discussion
3
0
21
Mar 16

Unsplash.com

Thank you! To get in touch, please email: CITES@kew.org

Some references we have found useful:

- Information Commissioner's Office, ico.org.uk.
- Lavorgna, A. (2020) Cybercrimes, critical issues in a global context, Red Globe Press
- Lavorgna A. (In Press) Researching Cybercrimes – Methodologies, Ethics and Critical Approaches, Palgrave Macmillan
- Markham, A. & Buchanan, E. (2012) Ethical Decision Making and Internet Research: Recommendations from AoIR Ethics Working Committee.
- Social Data Lab – socialdatalab.net
- Zook et al, (2017) Ten simple rules for responsible big data research, *PLOS Computational Biology.*
- Zimmer & Kinder-Kurlanda (ed), Internet Research Ethics for the Social Age, Peter Lang Publishing, 2017