D2.3 Evaluation of iEcology workflow output

2026-03-31

Author(s): Reynaert Simon, Alam Azharul, Hulme Philip, Pipek Pavel, Novoa Ana, Billiet Niels, Meeus Sofie, & Groom Quentin

Funded by
the European Union

Prepared under contract from the European Commission
Grant agreement No. 101180559
EU Horizon Europe Innovation Action
European Research Executive Agency

Project acronym:	**OneSTOP**
Project full title:	**OneBiosecurity Systems and Technology for People, Places and Pathways**
Project duration:	01.01.2025 – 30.06.2028 (42 months)
Project coordinator:	Dr. Quentin Groom, Agentschap Plantentuin Meise (MBG)
Call:	HORIZON-CL6-2024-BIODIV-01-1
Deliverable title:	Evaluation of iEcology workflow output
Deliverable №:	D2.3
WP responsible:	WP2
Nature of the deliverable:	Report
Dissemination level:	Public
Licence of use:	CC BY
Lead partner:	Meise Botanic Garden (MBG)
Recommended citation:	Reynaert, S., Alam, A., Hulme, P., Pipek, P., Novoa, A., Billiet, N., Meeus, S., & Groom, Q. (2026). ***Evaluation of iEcology workflow output.*** OneSTOP project D2.3.
Due date of deliverable:	Month № 15
Actual submission date:	Month № 15

Deliverable status:

Version	Status	Date	Author(s)
1.0	Draft	27 February 2026	Reynaert S.; MBG Alam A.; LU Hulme P.; LU Pipek P.; IBOT-CAS Novoa A.; CSIC Billiet N.; MBG Meeus S.; MBG Groom Q.; MBG
1.0	Review	4 March 2026	Golivets M.; UFZ Kumschick S.; SU

2.0	Final	31 March 2026

Table of contents

Key takeaway messages

- 14 internet platforms were screened for their utility to provide invasive alien species (IAS) occurrence data and early warning signals in an automated, low-cost fashion.
- The primary barrier was the inability to access geolocated data, as major internet platforms (X, Google, TikTok) rejected researcher API access, limiting automation.
- For five platforms (Google Trends, Wikipedia, Flickr, YouTube and Facebook), geolocated IAS-related internet activity data was evaluated using repeatable workflows bundled into a software package.
- The collected iEcology data was biased and of variable quality, with overall poor predictive accuracy of real-world IAS occurrence patterns, despite the potential of platforms that allow automated verification of geolocated observations (Flickr, YouTube) as complementary data sources.
- Privacy legislation complicates the publication of coordinate-level iEcology observations, limiting their utility for fine-grained IAS modelling (e.g., occupancy modeling) or publishing onto publicly available databases (e.g., GBIF).

Executive summary

This report presents the outcomes of our investigation into the utility of iEcology data—the practice of using publicly available internet information to answer ecological questions—for the early detection and monitoring of Invasive Alien Species (IAS) in Europe. The core goals were to develop automated, open-source data mining workflows and to assess the quality and utility of data from various online platforms.

We evaluated the suitability of 14 internet platforms based on criteria such as data type, cost of API access, geolocation detail, and researcher authorization. A critical finding is that, despite the potential, the technological readiness of iEcology is severely hampered by significant barriers to low-cost researcher API access. Applications for non-profit access to X (formerly Twitter), Google Ads, Google Search API, and TikTok were all rejected. Automated data mining for Google Trends was found to be unreliable.

A subset of platforms that granted API access—including Meta (Facebook), YouTube, Flickr, and Wikipedia pageviews—was incorporated into repeatable, open-source workflows. Data collected from these platforms, however, exhibited a number of challenges:

- Bias: Data was skewed towards charismatic species in densely populated areas, with platform-specific differences in how well it represents real-world IAS occurrence patterns.
- Geolocation: Detail was highly variable, ranging from country-level (Facebook, Wikipedia) to precise coordinates (YouTube, Flickr).
- Quality issues: It was impossible to verify activity or individual observations for many platforms in an automated fashion (Wikipedia, Google Trends, Facebook); some data had to be normalized due to small sample sizes (Facebook), and other data sources

contained a substantial amount of AI-generated content, further inflating false positive rates (YouTube).

The quality and utility of the collected data were tested in three case studies: Case Study I evaluated the utility of manually downloaded Google Trends data to track invasive alien plant occurrences in the USA, proving limited accuracy for using this data source to map real-world IAS occurrence patterns. Case Study II constructed an automated workflow to repeatedly extract, analyze, and visualize geolocated social media data on IAS of Union concern, evaluating it against trusted observations from GBIF and EASIN. Even in the best scenarios, changes in IAS-related internet activity only picked up real-world invasions in ~ 50 % of cases. Case Study III developed a proof of concept for an automated pipeline to extract and identify plant species from geolocated Flickr images using the Pl@ntNet API, illustrating the potential of the platform as a complementary data source for low-cost mining of IAS occurrences.

Based on our investigation, IAS-related internet activity data trends poorly reflect real-world occurrence patterns. Nonetheless, iEcology holds potential to complement existing monitoring efforts within early warning systems, raising its technological readiness. Facilitating its widespread operational use requires overcoming substantial hurdles, including the removal of barriers to researcher access to platform data and APIs, the need for improved automated data validation capacities to address human bias and low-quality content and clarification regarding the use of iEcology data within a GDPR-safe context.

Non-technical summary

Invasive Alien Species (IAS) — plants and animals introduced to new areas that cause harm to people and the environment — are a major problem, and finding them quickly is crucial. Traditional monitoring methods, such as field surveys by experts, are slow and limited. iEcology offers a modern alternative: using public information from the internet, like social media posts, search trends, and photos, to rapidly detect new species outbreaks.

This report explored whether this internet-based tracking is reliable and practical for use across Europe. Our work focused on two questions: Can we easily build automated tools to collect this data? And how trustworthy is the data once we get it?

We looked at 14 popular platforms, including X (formerly Twitter), Facebook, Google, Wikipedia and TikTok. The biggest and most consistent hurdle we faced was access. Despite seeking free access for non-profit research, major platforms like X, Google Search, Google Ads, and TikTok rejected our requests. Automated data gathering from Google Trends also proved unstable. This means that a crucial step for widespread, low-cost monitoring is currently blocked by the companies that own the data.

We successfully built working, automated data collection systems using low-cost platforms that granted us API access, namely:

- Facebook: Used to track post activity about IAS.
- YouTube: Used to find geolocated videos.
- Flickr: Used to find geolocated photos.
- Wikipedia: Used to track how often species pages are viewed.

The data gathered is not perfect. We found that internet information is naturally biased: people post more often about charismatic or easily recognized species, and most of the data comes from densely populated areas. We also ran into technical quality problems, such as low sample sizes from Facebook and a surprising number of fake (AI-generated) videos on YouTube that could easily be mistaken for real sightings.

iEcology has limited potential as a stand-alone, automated early warning system. Its use seems to be primarily complementary within other early warning systems and should revolve around harvesting individual occurrences rather than search or pageview internet activity trends. For iEcology to reach its full potential, three major things must happen:

1. Platform owners must allow researchers low-cost access to their (geolocated) data.
2. Better automated tools are needed to filter out bias and low-quality information, such as bot posts and AI-generated content.
3. Clarification regarding the use of iEcology data within a GDPR-safe context is required, as current legislation complicates the publication of coordinate-level observations.

List of abbreviations

EU	European Union
MBG	Meise Botanic Garden (BE)
LU	Lincoln University (NZ)
IBOT-CAS	Botanický ústav AV ČR, v.v.i. (CZ)
CSIC	Agencia Estatal Consejo Superior De Investigaciones Científicas (ES)
API	Application Programming Interface
GBIF	Global Biodiversity Information Facility
UL	List of Invasive Alien Species of Union Concern
IAS	Invasive alien species
FAIR	Findable, Accessible, Interoperable & Reusable (data)
DSA	Digital Service Act
GDPR	General Data Protection Regulation
EASIN	European Alien Species Information Network

1. Introduction

iEcology, i.e., the practice where publicly available information is obtained from various internet sources (e.g., Reddit, Google Trends) and used to answer ecological questions, has recently been put forward as a promising avenue for accelerating detection of invasive alien species (IAS) outbreaks and better understanding their spread and population dynamics (Jarić et al., 2021; Novoa et al., 2025; Tomojiri & Takaya, 2025). For instance, taxonomists previously identified the presence of invasive fish and flies from images on angler websites and enthusiast Facebook groups across different EU countries (Kalous et al., 2018; Schifani & Paolinelli, 2018).

Despite the abundance of potential information available in the online public domain, iEcology data are fragmented, of highly variable quality and usually require thorough manual polishing before they are suitable for ecological analyses (see e.g., issues with bots mentioned in Tomojiri & Takaya, 2023). Moreover, while many different public platforms have proven their applicability for specific cases in the past, such as whale monitoring in Spain (Morais et al. 2021) or plant monitoring in Israel (Vardi et al., 2021), no iEcology studies thus far have thoroughly evaluated to what extent iEcology is applicable at a continental scale, and whether it could directly complement evidence-based informing of policy makers and other stakeholders regarding IAS decisions.

Naturally, data generated from social media mentions will be inherently biased (Olteanu et al., 2019). For iEcology analyses, data are skewed towards larger-bodied, charismatic and easily recognizable taxa from densely populated areas with sufficient internet access (Jarić et al., 2021; Novoa et al., 2025). Thus, it is virtually impossible to use iEcology data alone for answering typical ecological questions related to IAS dynamics, such as reliably estimating population size. Despite the increasing number of citizen science projects with relatively accurate classifying mechanisms; e.g. iNaturalist (López-Guillén et al., 2024), monitoring efforts in many parts of the world are limited to a relatively small number of trained taxonomists and delays in data publishing may hamper rapid IAS detection (Kourantidou et al., 2022). Hence, monitoring and detecting changes in online activity regarding IAS shows potential to become a useful extra component in early warning systems tracking IAS spread (Cardoso et al., 2024).

This report presents the outcomes of our investigation on the utility of iEcology for IAS occurrence detection within the context of the OneSTOP project. Our goals were to:

1. develop automated, repeatable and open-source workflows for mining IAS data from positively evaluated online platforms to raise the technological readiness of iEcology from 4 (validated in lab) to 6 (demonstrated in relevant environment).

2. assess the quality of data obtained from different platforms by comparing them with known reliable data sources on IAS occurrences.

3. evaluate which (if any) online platforms can aid monitoring of IAS expansions and occurrences.

2. Suitability of online platforms for automated geolocated iEcology data mining

Based on previous studies and the current digital landscape (Novoa et al., 2025), we evaluated the suitability of 14 different internet platforms for efficient and low-cost collection of data on IAS in Europe between 01/01/2025 and 31/12/2025. From all platforms, different aspects were summarized and evaluated on their applicability (Table 1). These included i) the type of accessible data (txt, jpg or vid), ii) the cost of accessing the data, iii) the existence of geolocation information associated with the data, iv) the query limit for the respective application programming interfaces (APIs), v) the temporal resolution of the data, and vi) whether low-cost data mining for research purposes was authorized by the platform.

In principle all platforms explored either incorporate free or researcher API application procedures or did so in the recent past. Moreover, communication through API's is in most cases already supported by platform-specific Python and R libraries. However, as discussed in more detail below, four platforms rejected our researcher (non-profit) applications (X, Google Ads, Google Search and TikTok) and data from one platform (Google Trends) was only incorporated using manual data downloads since the automated mining libraries in R and Python turned out to be highly unreliable and likely against the platform's terms of service.

Table 1: Qualitative characteristics utilized to evaluate the suitability of different internet platforms for mining iEcology data through APIs. Text or numerical values, images and videos are abbreviated by txt, img and vid, respectively.

Platform	Data types	API Cost	Geolocation detail	Query limit	Temporal resolution	Mining authorized?
X (Twitter)	txt/img/vid	free*	user dependent	NA**	any	No
Meta	txt/img/vid	free*	country	60/min	any	Yes
Google Ads	txt	free*	country/region/city	NA**	monthly	No
Google Trends	txt	free*	country/region/city	NA**	hourly	No
Google Search API	txt/img	free*	country	NA**	any	No
YouTube	txt/img/vid	free*	full coordinates	10k/day	any	Yes
TikTok	txt/img/vid	free*	user dependent	20k/day	any	No
Reddit	txt/img/vid	free	language based	100/min	any	Yes
Wikipedia	txt	free	country	unlimited	daily	Yes
Flickr	txt/img	free	user dependent	60/min	any	Yes
Bluesky	txt/img/vid	free	language based	600/min	any	Yes
Mastodon	txt/img/vid	free	language based	60/min	any	Yes
iNaturalist	txt/img	free	full coordinates	60/min	any	Yes
eBay	txt/img	free	country/region/city	5k/day	live listings	Yes

*Platform-specific limitations affecting utility apply, see detailed descriptions below.

**NA values indicate lack of relevant information about query limits due to poor documentation or rejected (researcher) API applications.

After assessing the accessibility and overall suitability of different platforms, we downloaded and evaluated the IAS internet activity data quality using three case studies focused on the most promising platforms (Fig. 1). In the first case study, repeatedly downloaded data from Google Trends on 100 invasive plants was compared with a US dataset and activity signals evaluated for how well they represented true occurrence patterns (Alam & Hulme, 2026). In the second case study, we constructed an automated workflow incorporating a subset of internet platforms to repeatedly extract, analyze and visualize geolocated social media data on IAS of Union concern (Commission Implementing Regulation (EU) 2022/1203 of 12 July 2022 Amending Implementing Regulation (EU) 2016/1141 to Update the List of Invasive Alien Species of Union Concern, 2022). For this workflow, an updated version of the Union list of concern (UL) was programmatically extracted from Wikidata and further used as input data for API queries on geolocated internet activity across a variety of platforms where we obtained API access (Facebook, Flickr, Wikipedia, YouTube). After initial cleaning and processing, these datasets were evaluated against observations from GBIF and EASIN and checked for anomalous activity increases surrounding the moment of species expansions into new EU countries. Finally, for the third case study, we developed a proof of concept to automatically download geolocated Flickr images and identify plant species present using the Pl@ntNet API, to evaluate the feasibility of automatically extracting high-quality species observations from this platform. All workflows were set up to be repeatable, and those of case study II and III were incorporated into a software package (Fig. S1; https://doi.org/10.5281/zenodo.19235707). In the following sections, we further explain platform-specific details as well as their adherence to the evaluation criteria.

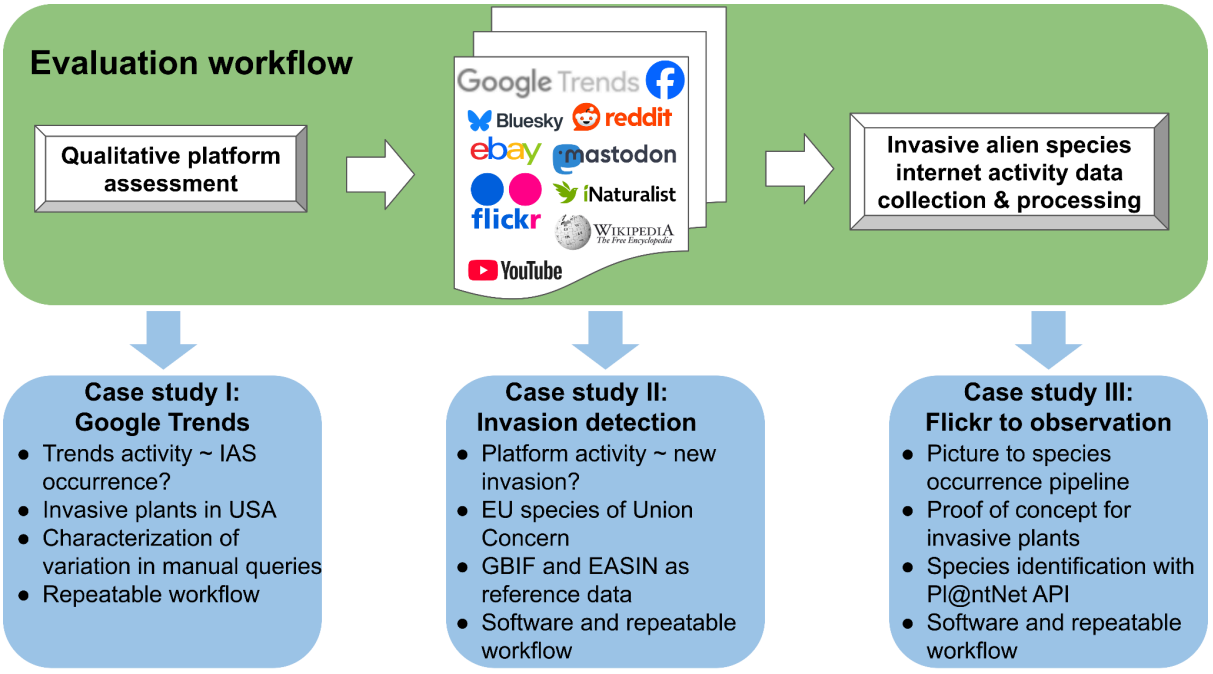

Figure 1: Overview of the workflow followed to evaluate the use of iEcology data for invasive alien species monitoring.

2.1. X (Twitter)

Due to the high number of short communication messages (text, images and/or videos) being “tweeted” daily, X (formerly known as Twitter) has proven itself as a valuable data source for iEcology research (Bhatt & Pickering, 2021; Tomojiri & Takaya, 2025).

Since Twitter was bought in 2022 by Elon Musk (and renamed to X in 2023), the social media platform, its moderation practices, business philosophy and the societal representation of active users have changed considerably (Bisbee & Munger, 2025; Novoa et al., 2022). This is also noticeable in the way API access for developers (and especially researchers) has changed. As exemplified by previous studies (e.g., Tomojiri & Takaya, 2025; Novoa et al., 2022), obtaining free API access to perform data mining of sufficient tweet volumes within the context of academic research used to be quite straightforward. Yet, this is no longer the case. Despite the implementation of EU legislation which should allow researchers to access the twitter API for free under certain conditions (EU, 2022), the Meise Botanic Garden (MBG) application for researcher API access (X, 2025b) within the context of the OneSTOP project was rejected on 20/3/2025, with the reason given that “it does not appear that your proposed use of X data is solely for performing research that contributes to the detection, identification and understanding of systemic risks in the EU as described by Art. 34 of the Digital Services Act”.

Hence, we looked for alternative ways to obtain X API access. While X currently offers a free API tier to all developers, this free tier only allows automation of tweet posting (1500 / month) and no reading, though the latter is required for IAS data collection (X, 2025a). To actually retrieve post reading info, one has to go for the paid API option with a ‘pay as you go’ schedule, which can rapidly become very costly when querying systematic keywords list for thorough analysis. Given the considerable pricing of this tier for conducting research, especially over longer periods of time given its intended use for active monitoring of species names and associated terms, we deemed X not suited for IAS monitoring and detection within the context of the OneSTOP project.

2.2. Meta (Facebook)

Meta, formerly known as ‘Facebook’, has expanded in recent years to incorporate an array of platforms including Facebook, Instagram and Threads where users can post online content including text, images and videos. Previous studies show the value of these platforms for early detection of IAS (Schifani & Paolinelli, 2018). Meta has a meta-hosted API service for which we did not receive access, but researchers can also apply for access to the Meta Content Library which aggregates data from Meta’s platforms through the Social Media Archive (SOMAR) of the University of Michigan in a secure research environment (SRE) (Sweet, 2025). However, this API introduced a new pricing scheme (a $1,000 registration fee + a $400/month usage fee) starting in March 2026.

After obtaining researcher access to the SRE, we were able to fetch Meta online activity data by filtering out the surface_countries for Facebook posts containing exact matches with scientific names of species across all EU countries. However, due to small sample sizes and increased risk of exposing poster identity related to these (internal communication), data had to be summarized to normalized (monthly post count / maximum monthly post count across the sampled period) post counts before export out of the SRE. These normalized temporal data (similar to Google Trends) were used in case study II (see section 3.2).

2.3. Google

As one of the bigger players online globally, Alphabet (mother company of Google) collects and hosts a wide array of data across its many platforms. Within the context of OneSTOP, we explored collecting data on IAS from four of these platforms, including Google Trends, Google Ads, Google Search and YouTube.

2.3.1. Google Trends

Google Trends tracks the popularity of individual keywords or topic searches performed through its search engine for a specific country, region or city over time and has shown its value for iEcology research (Schuetz & Johnston, 2019). Although Google Trends data are normalized and relative (expressed as percentage of the maximum interest a specific term/topic across the time of interest per query), the data are quite well moderated and of relatively high quality because it e.g., removes duplicates, bot searches and even accounts for typos made by the searcher (Google, 2025b). On the other hand, the data received are based only on a subsample of total search volumes.

Although Google Trends has been used as a tool for the infodemiology and infosurveillance of human diseases for more than a decade), it has only recently caught the attention of invasion scientists (Alam & Hulme 2026). Nevertheless, despite this short history, Google Trends has proven to be useful in investigating public attention and awareness towards various alien species. In general, Google search volume appears to be correlated with the spatial distribution of alien plants, insects, and fish within a region (Fazzari et al., 2024; Fukano and Soga, 2019; Maciejewski et al., 2025; Proulx et al., 2014; Woodworth et al., 2023), suggesting that data from Google Trends could be a suitable tool for surveillance and early warning (Jarić et al., 2020; Novoa et al., 2025). However, despite the rising popularity of using Google Trends to examine the spatial patterns of biological invasions, much of the research to date has paid scant attention to three major pitfalls of using this tool.

While manual querying and downloading .csv files from the Google Trends website is free, relatively fast and within the terms of use, we were unable to construct a workflow that was reliable enough to perform automated data collection on a regular basis. In principle, Google Trends has a (hidden) API that can be accessed utilizing packages in both R (*gtrendsR*; Correia 2024) and Python (*pytrends*; Pytrends 2025), but we found both of these unreliable when making larger data queries, as this would result in blocked requests for some queries and not others, often without a clear reason. These problems were somewhat in line with our expectations, given that the Google Trends endpoint these packages are accessing is not mentioned in the official Google API documentation, and hence likely no longer actively maintained (even if it was so in the past), or work via scraping the Google Trends site directly, which would be against Google's terms of service. During the 2nd half of 2025 Google announced the alpha testing phase of a new Google Trends API, available through their cloud platform. However, the MBG application to become an Alpha tester never received a response. Given these issues, we decided not to incorporate Google Trends data into our automated workflow. Instead, we manually downloaded data containing activity patterns across space and time in .csv format for further analysis in case study I (see section 3.1).

2.3.2. Google Ads

Google provides a variety of other data services through their cloud platform. One of them is Google Ads, which developers use to track marketing trends and to manage and tailor their

advertisements to the intended audience over time. Within Google Ads, there is a feature called Search Term Report that allows developers to collect monthly search volumes for specific keywords as well as their estimated market values across different geographic scales (country, city, region) (Google, 2025a). Given the previously discussed issues with Google Trends data, we pursued obtaining API access to mine data on monthly IAS search volumes since we presumed these would give similar info as Google Trends, but our MBG application within the context of the OneSTOP project was rejected on 14/3/2025, since "tools that offer only keyword research are not allowed by the Google Ads API Policy". As such, we also deemed Google Ads unsuitable for integration into our automated IAS monitoring and detection workflow.

2.3.3. Google Search

For researchers specifically, Google developed the Search Researcher Result API, returning the rankings and associated metadata of top Google search hits, which could aid in investigating and quantifying geographic differences in IAS search activity and (search results algorithm dependent) popularity of individual data sources. However, the MBG application to receive API access was rejected due to the "inability of the institute to prove that sufficient data protection measures were in place".

2.3.4. YouTube

The fourth and final Google service with proven iEcology relevance is YouTube (Dylewski et al., 2017). The Google Cloud Platform provides access to the YouTube database where YouTube videos, thumbnails and comments as well as their associated metadata for specific keywords can be retrieved via YouTube Data API (v3)(YouTube, 2025). After setting up a trial account in Google Cloud (3 months only) and having created a test project, authentication credentials and a free API key can be obtained. Specifically interesting for iEcology applications, is that the YouTube API allows searches in a radius (max. 1,000 km) around a specific location, which can be used to geolocate videos, albeit dependent on user privacy settings (i.e., explicit location info needs to be filled in in order to be able to retrieve it) (Wright, 2023; YouTube, 2025). An important limitation is that a maximum of 10,000 YouTube quota are available in the free trial everyday (100 quota per unique search query), which is approximately equal to maximally extracting 5,000 videos on a daily basis (100 API calls with maximally 50 videos per page). However, after contacting YouTube, our quota got upped to 200k, and we were able to collect geolocated videos in circles approximately covering all EU countries and summarized them for further evaluation in case study II (section 3.2). Within this context, it is important to note that while our queries returned many videos referring to geolocated content about the real species of interest, a substantial amount were AI generated videos that inflated false positive rates, and are likely difficult to distinguish from 'real' observations via automated processes (e.g., using machine vision models; see section 3.3). Moreover, exact query matches on YouTube are inconsistent, necessitating post-hoc filtering by exact species names to avoid hyperinflation of false positives.

2.4. TikTok

As one of the most popular social media platforms for younger generations, TikTok is a valuable secondary data source for obtaining information about species distributions (Balestrieri et al., 2023; Nascimento et al., 2024). The primary format of TikTok content is characterized by short videos, accompanied by descriptions and comments.

MBG applied for access to the TikTok API on 18/02/2025 and received a response in June 2025, stating that MBG is not considered an academic institution and can therefore not be granted access to their database. Additionally, the local Flemish government prohibits its employees from accessing the app on their personal devices due to concerns over its safety (El Bakkali, 2023), which made it practically impossible to evaluate this platform. Because of these reasons, we did not further explore incorporating TikTok into our final workflow.

2.5. Reddit

Reddit is an online platform where people discuss topics in a forum-style format, through “subreddits”, i.e., subsites specializing in specific subjects. Posts may contain text, images and/or videos. Although geolocation info is limited and can only be derived indirectly from user or post content (Reddit, 2025), the platform has proven its usefulness for iEcology research in the past (Fox et al., 2021). In particular, the ability to search less broadly for more tailored info within ‘expert’ subreddits seems to have much potential.

We obtained standard API access for Reddit using the free developer app platform through the actively maintained Reddit package *praw* in Python (Bryce, 2025) and set up a workflow to mine all relevant IAS data from publicly accessible subreddits by specifying exact matches with queries in the cloud backend of the API. While the resulting dataset contained interesting information regarding IAS sentiments across EU languages (i.e., useful for culturomics approaches), accurate geolocation was virtually impossible to automate without manual verification, resulting in limited use of these data for iEcology and mining of IAS occurrences. Therefore, we did not further explore incorporating Reddit into our final workflow.

2.6. Wikipedia pageviews

The number of searches redirected to individual pages on Wikipedia, i.e., pageviews, has proven its usefulness in previous iEcology-oriented research, for example, to investigate temporal changes in pageviews in relation to plant phenology or invasive species outbreaks (Cerri et al., 2022; Henke et al., 2024; Mittermeier et al., 2019; Vardi et al., 2021). As a widely used platform that has existed for over two decades, we thus also pursued obtaining (geolocated) pageview data through the Wikimedia Analytics API. In the end, we set up two different workflows to evaluate Wikipedia page view info, both with their respective benefits and shortcomings.

The first method fetches individual pageviews on a daily basis over time from all linked pages in all available languages to the Wikidata species Q identification numbers of all unique species on the UL. While this yields a large dataset including many different EU languages, the accuracy of this method for reliably obtaining info on a country level is low, as exemplified by Zachte (2018). According to their study, pageviews on e.g. the Finnish page of *Alopochen aegyptiaca* may indeed predominantly originate from Finland, whilst those from its corresponding English page may originate from anywhere across the world (Zachte, 2018).

Given the limitations of this approach for research, the Wikimedia Foundation started collecting geolocated daily pageview data from 2017 onwards for each unique Q identifier in its database. However, to protect the personal information of page visitors, only pageview counts above 90 are recorded on a daily basis for specific Q identifiers, hampering time-series analysis of pages/identifiers that get relatively low traffic volumes. Nonetheless, we set up a workflow to extract the geolocated pageview data of species with sufficiently high search volume from the Wikimedia endpoint that performs daily data dumps in the form

of .tsv files. Finally, although beyond the scope of this report, internal communication with the Wikimedia team indicates that there will soon (within the coming months; communication from March 2025) also be a publicly accessible API that can be directly called and integrated into future software products to fetch geolocated pageviews more efficiently.

We compared the two currently available Wikipedia pageview data sources (language based vs geolocated filtering) between 2017 and 2025 and found relatively good correlations and trends for most countries with sufficient data points (Fig. 2; Reynaert et al., 2026). As expected, accuracy declined for English (Ireland) and other languages (e.g., Portuguese, Spanish) that are spoken worldwide (Zachte, 2018). Another reason for differences, especially regarding peaks at low language-based pageviews, could be that the geolocated searches aggregate searches per country for the same species across different language subsites and/or pages. For instance, certain species may only have an English page, and thus all searches from any country will be immediately redirected to that one. At the same time, we were unable to test correlations for many of the less visited pages/languages (e.g., Latvian) since they do not show up in the geolocated pageview statistics, limiting their use. Nonetheless, it is likely that even less visited pages would pass the threshold of 90 pageviews during times when a species is gaining popularity locally due to redirected views, even if no page exists in the language of the Google search, further highlighting the importance of the geolocated pageviews database within the context of tracking IAS internet activity. Due to the overall relatively good correlations between Wikipedia language-based and geolocated pageviews, these data sources were both further evaluated in case study II (section 3.2).

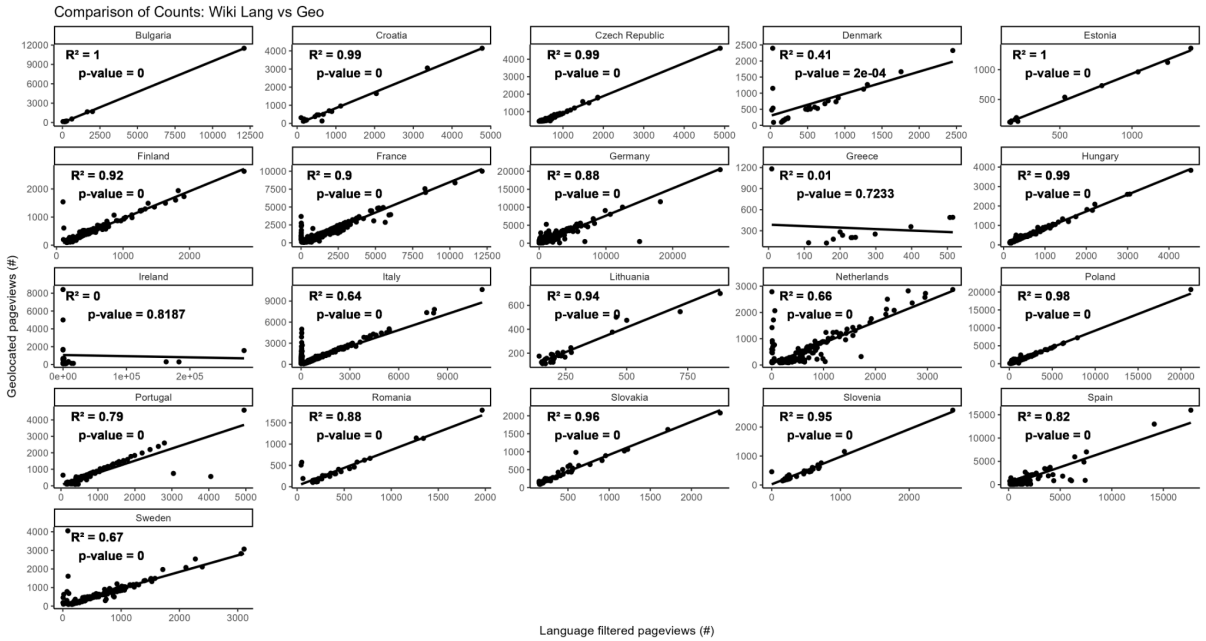

Figure 2: Correlations between Wikipedia pageviews per country based on language filtering vs direct geolocation (> 90 pageviews per day) for all species on the UL. This figure was adopted from Reynaert et al. (2026).

2.7. Flickr

As an online platform where people share and store photos, often with geolocation tags, previous studies indicate that Flickr can be used for gathering information on IAS

occurrences in European countries (Allain, 2019; Cardoso et al., 2024; Edwards et al., 2021). After obtaining access to the Flickr API free tier, we specified an approximate EU bounding box (25° W to 40° E longitude and 34° N to 72° N latitude) and collected all geolocated pictures (either including user-entered coordinate info or image EXIF derived metadata) through text matches with our IAS queries. Important to note is that while searching utilizing machine tags is in principle more accurate, they seemed to be very much underassigned for the species of interest, resulting in us opting for “free text” search (incl. “free text” tags). Whilst the latter option increases false positives, we believe this is outweighed by the net gain in informative data points. Moreover, to minimize false positives, we post-hoc performed exact query matches to only retain those pictures where the associated text included the full species names. Note that the free Flickr API tier only gives access to low-resolution images, thus limiting its reach within the Flickr database. Nonetheless, given the accurate geolocation of the resulting dataset, this data source was further evaluated in case study II.

2.8. Bluesky

Although not part of the Fediverse (Theophilos, 2024), Bluesky adheres to many of the principles behind the philosophy of decentralized social media, such as being maximally transparent and open (incl. open source code) and storing a minimal amount of sensitive user data (Failla & Rossetti, 2024). Moreover, previous studies highlight its twitteresque potential for sentiment analysis on a wide variety of topics (Failla & Rossetti, 2024), which has applications in iEcology and culturomics. We performed Bluesky post text mining that queries for exact matches with our invasive species dictionary using the *atproto* Bluesky Python library. After fetching results, we filtered out those posts in languages spoken across the EU regions. In line with data from Reddit, these posts contained interesting, albeit more niche information regarding IAS sentiments across EU languages (i.e., useful for culturomics approaches). However, accurate geolocation was virtually impossible to automate without manual verification, resulting in limited use of this data for more iEcology-focused analysis and automated IAS occurrence data mining pipelines. Therefore, we did not further investigate incorporating Bluesky into our final workflow.

2.9. Mastodon

Similar in rationale to Bluesky, but an actual part of the Fediverse, Mastodon is one of the more popular decentralized social media platforms (Theophilos, 2024). In contrast to other platforms, different Mastodon subsites and communities are fully disconnected and located on different servers or ‘instances’. Nonetheless, previous research indicates its potential for enriching social media-focused datasets (Persico, 2024). Since the more segregated nature of Mastodon made collecting all available posts (‘toots’) on IAS more challenging, we settled for collecting public posts from the *mastodon.social* instance, which is most similar to X or Bluesky since they may contain text, images and/or short videos.

In line with Bluesky, geolocation info is not directly retrievable through the API due to high privacy standards, so any geolocation must be conducted post hoc via text analysis. As such, the resulting dataset not only had a very limited reach (small userbase) but also primarily returned information regarding IAS sentiments without providing sufficient information for automated IAS occurrence mining pipelines.

2.10. iNaturalist

Biodiversity focused citizen science internet platforms are a known and relatively reliable source of ecological information for species around the globe (Callaghan et al., 2021). One

of the most commonly used citizen science platforms is iNaturalist, where users upload geolocated pictures of species with a taxonomic guess, after which other users can verify the classification and validity of this observation (Callaghan et al., 2021). The usefulness of this platform is also recognized by GBIF, which directly pulls data of 'research grade', i.e., verified by multiple users, observations from iNaturalist on a regular basis (GBIF, 2025).

To verify if non-research grade iNaturalist observations contain useful information regarding IAS occurrences, we constructed a workflow that extracts non-research-grade iNaturalist observations. The observation categories extracted included "casual", related to cultivated/captive/poorly documented species and "needs ID", of which the classification has not been confirmed by others. Given the unknown quality yet accurate geolocation of the resulting dataset, this data source was further evaluated in case study II.

2.11. eBay

Despite generally in breach of their policies, alien and invasive species are commonly traded on internet trade platforms, including eBay (Heinrich et al., 2019; Humair et al., 2015). As a platform providing public information about the location where listings are posted and sold, eBay may thus give insight into where IAS are potentially being held and bred. For exploratory purposes, we used the eBay API (5,000 calls / day using a free account) to collect data about active listings mentioning IAS from the UL.

Although exact text matching works well and the platform enables geolocation of listings, an important shortcoming of eBay is that the API only returns data regarding public and live listings (preventing timeseries analysis), and that their terms of use vastly limit which types of data can be retained and analyzed externally due to conflicts with their terms of service regarding eBay market analysis and manipulation (eBay, 2026). Moreover, whilst plant listings were often interesting, referring to e.g., sellers of IAS seeds across EU countries, listings from other species groups mostly referred to objects other than actual species (e.g., books, posters, etc.), resulting in high false positive rates. For these reasons, we decided to not further analyse these data.

3. Evaluation of collected iEcology data quality to monitor invasive alien species

Due to lack of authorization or technical issues related to obtaining reliable access to their database through maintained APIs (Google Search API, Google Ads, X, TikTok), lack of a temporal dimension in the data (eBay) and missing country-level geolocation information (Reddit, Bluesky, Mastodon), only 6 out of the initial 14 explored platforms were retained and further evaluated. These included Google Trends, YouTube, Wikipedia, Flickr, Facebook and iNaturalist (non-research-grade observations). We further evaluated the data quality of the collected data and compared datasets regarding scope, false positives and utility for monitoring IAS in the EU.

As reference data for evaluation we used the USDA PLANTS database (case study I; Alam & Hulme (2026)), a dataset containing all recorded invasions of IAS from the UL retrieved from EASIN (case study II; Reynaert et al. (2026)) and all available European per-country observations of species from the UL from GBIF between 01/01/2022 and 15/07/2025 collected utilizing the *pygbif* library (GBIF, 2026). Out of all GBIF observations <1 % (n = 2,630,741) was discarded since they represented a date range > 24 h, making it impossible to know when exactly the species was observed.

3.1. Case study I: Utility of Google Trends to track invasive alien plant occurrences across the USA

Google Trends has long been used to track the epidemiology of human diseases; however, its application to address biological invasions has been quite limited to date. We developed a workflow for best practice in the use of Google Trends to study biological invasions that accounts for the underlying pitfalls inherent in data from Google searches. We illustrated this workflow by examining the extent Google searches adequately depict the state-wide occurrence of 100 alien plant species in the United States (Fig. 3; Alam & Hulme, 2026).

While manually downloading data from Google Trends is straightforward (see section 2.3.1), the data itself proved to inconsistently reflect real-world trends of invasive plant species. Because each Google Trends search returns results based on only a sample of all Google searches, searches on different days or IP addresses present different results. For this reason a single Google Trends search can be misleading. It is recommended that the same Google Trends search is repeated at least 10 times to better capture search interest. This is rarely done, suggesting many previous studies only described a partial story. A further complication is the search term used. Google Trends can enable searches by scientific name, common name or provide more semantic searches using a Topic. These different search terms produce different results that vary considerably in their accuracy in describing species distributions. A final complication is that the search algorithm for Google trends has been updated on several occasions since its inception, meaning that temporal analyses are challenging due to the lack of comparability in different time periods. Accounting for these constraints sets up a complex workflow which would be difficult to automate. Hence, we undertook extensive manual searches that amounted to almost 4000 separate Google Trends searches. This is unlikely to be a feasible option for regular updating of distribution records using Google Trends (Alam & Hulme, 2026).

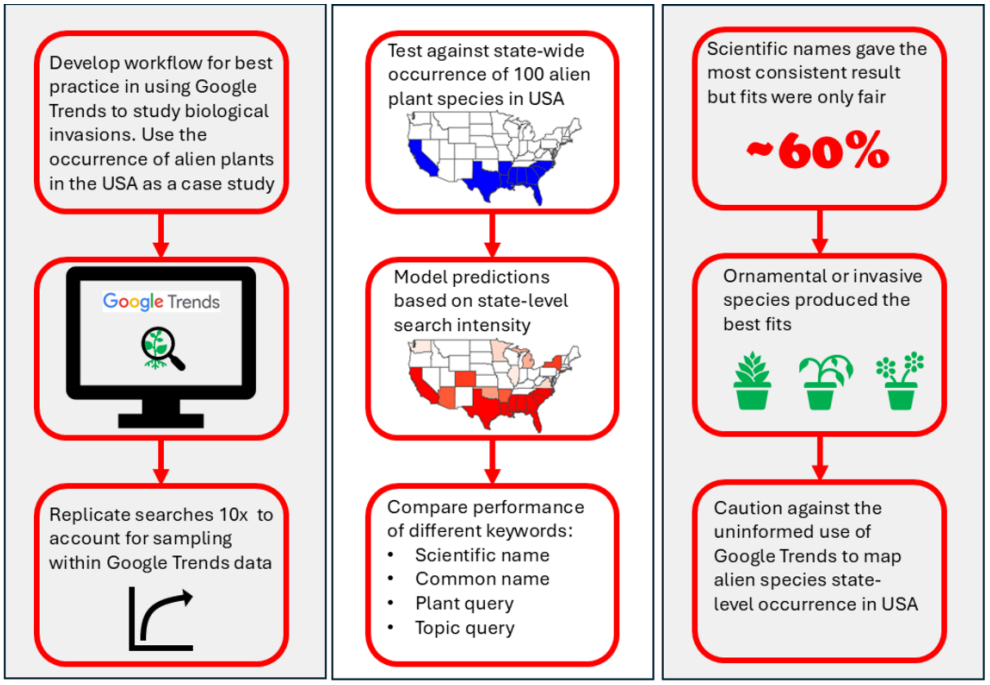

Figure 3: Graphical description of the workflow and primary outcomes from the Google Trends case study. This figure was adapted from Alam & Hulme (2026).

In general, results from Google Trends provided only a moderate goodness of fit to the known state-wide occurrence of alien plant species, and then only when the scientific name was used as a keyword. Other keywords, such as the common name or a Topic query, performed poorly. The goodness of fit between the observed species occurrence and that predicted using Google searches was higher for ornamental species or those officially classed by the USDA as noxious or invasive in at least one state. Those states with a greater alien plant richness and higher average education level of their citizens performed best. Given this data heterogeneity, we highlight an objective workflow (Fig. 4) to assess the value of Google Trends in order to ensure that greater scrutiny is applied when using this tool in invasion science (Alam & Hulme, 2026).

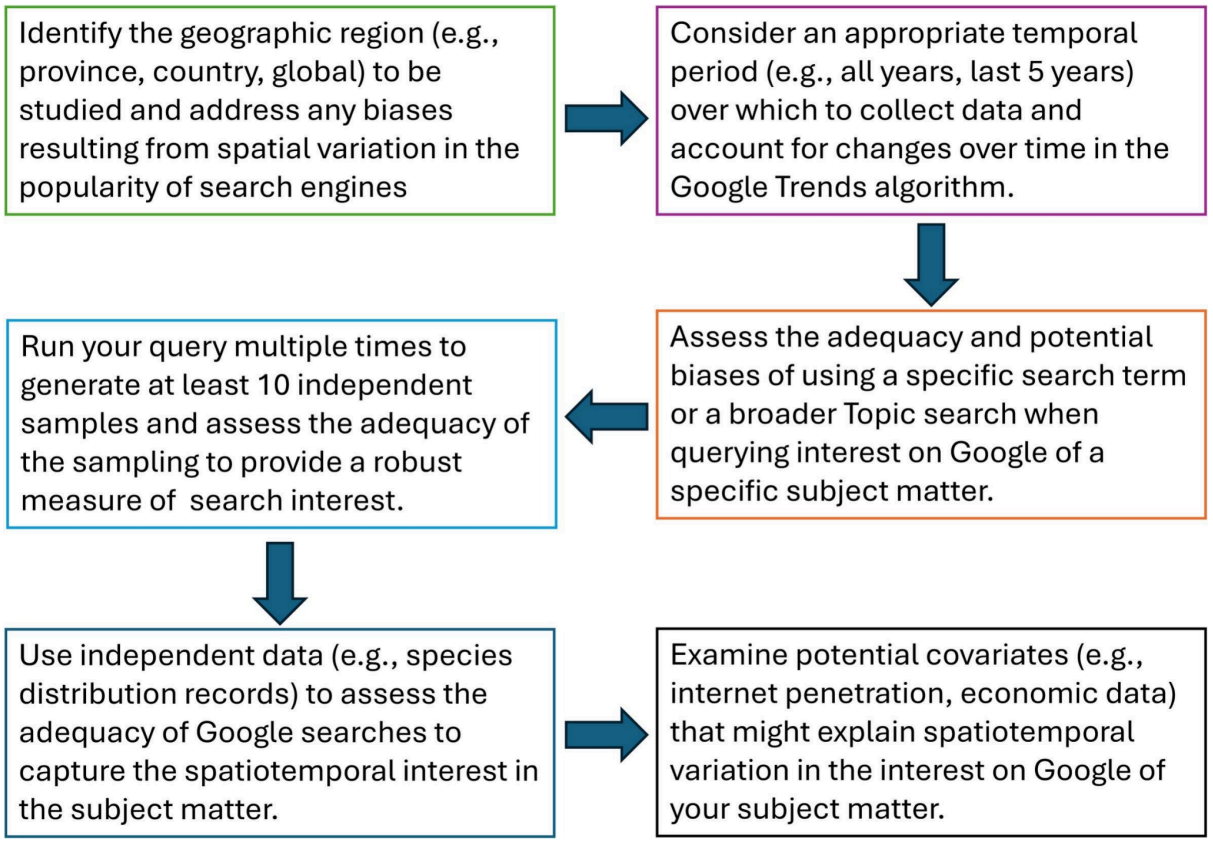

Figure 4: Schematic workflow of the issues that need to be considered when using Google Trends to investigate interest in a particular subject matter. The workflow has been designed to be generic for any subject matter though it has been developed from an examination of the ability of Google searches to capture the state-wide distribution of alien plants in the United States. This figure was adapted from Alam & Hulme, 2026.

3.2. Case study II: Capacity of geolocated internet platforms to detect IAS expansions within the EU

Other platforms were explored in a more automated fashion. After collecting relevant mentions, pictures, videos and activity on IAS of Union Concern from Facebook, Flickr, YouTube, Wikipedia and iNaturalist (non-research grade observations) (see section 2) through their respective API's using our software package, we combined all geolocated (per country) data by counting the number of mentions/activity per month between 2016 and 2025 (Fig. S1). These activities were then compared to the species observation data mined

from GBIF to evaluate how well they correlated with real-world occurrences over time (Fig. S1). Despite correlating poorly in general, our results indicated that Wikipedia, followed by iNaturalist (non-research grade observations) and Facebook, best represented real-world country-level IAS occurrence patterns across the EU (Fig. 5; Reynaert et al., 2026). Moreover, on average, insect, flatworm, plant and reptile observations correlated better with real-world occurrences compared to other groups (Fig. 5; Reynaert et al., 2026).

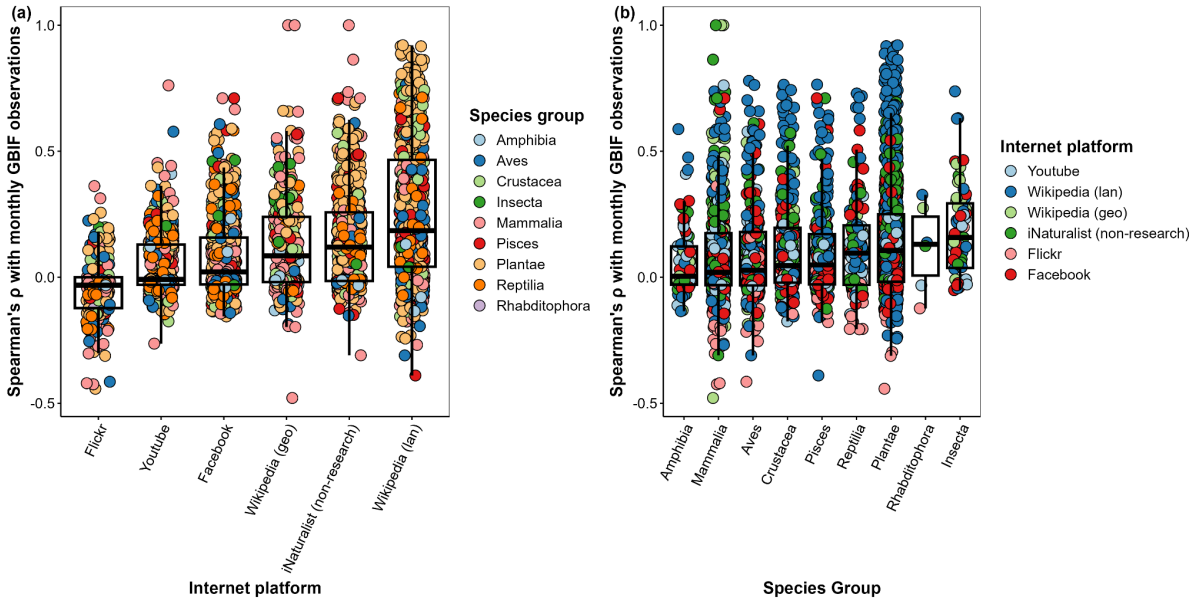

Figure 5: Correlations between internet activity per month per country and occurrences mined from GBIF between 2016 and 2025 for all species on the UL. Wikipedia (geo) and Wikipedia (lan) refer to geolocated and language-based pageviews respectively. This figure was adopted from Reynaert et al. (2026).

To further explore the ability of platform activity to detect IAS expansions into uninvaded EU countries, we used two different methods to detect anomalous increases in internet platform activity in the two-year window surrounding the EASIN report year of IAS of UL invasion for all EU member states (112 unique invasion scenarios across all species between 2016 and 2025). Analysis on individual platforms showed that even for the best performing platforms (Wikipedia and Facebook), only approximately 50 % of invasion scenarios with sufficient data showed anomalous activity increases during the two year invasion window. To improve detection capacity and reduce false positive rates, we then normalized IAS internet activity data (activity divided by maximum activity over the entire period) and pooled data from all platforms together to see if this increased detection performance. However, even using this method (identifying the first period of anomalous activity and variance increase across the timeseries; Reynaert et al., 2026), detection capacity was still only modest to poor across all species groups, with only slightly more true positives than false positives across all explored scenarios (Fig. 6).

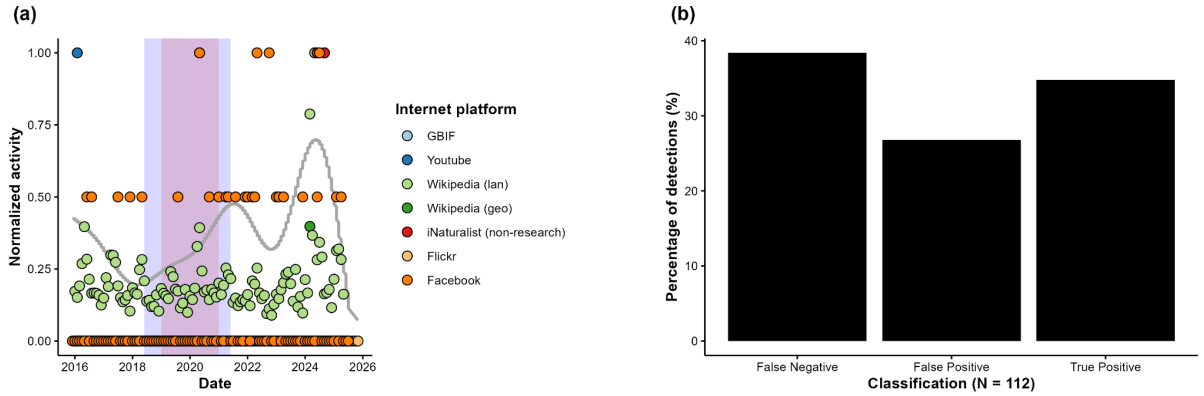

Figure 6: Example of a) the mean fitted GAM predictions (grey line) on the sum of normalized monthly activities for the combined internet platform activity data (excl. GBIF) of *Pontederia crassipes* in Poland. The red box outlines the 2-year invasion window (reported EASIN invasion year in 2020) and the purple box indicates the first region of strong activity increase in the summed normalized data timeseries based on our pre-defined threshold. GBIF data was not used in the models but is shown for illustrative purposes. Wikipedia (geo) and Wikipedia (lan) refer to geolocated and language-based pageviews respectively. In b) the sensitivity of the detection algorithm is summarized across the total of 112 unique EASIN reported invasion cases of species of the UL between 2016 and 2025. This figure was adopted from Reynaert et al. (2026).

3.3. Case study III: Flickr image to pl@ntnet identification to high fidelity occurrence pipeline

Although Flickr proved a poor predictor of recent IAS expansions into new EU countries (see 3.2), it previously proved useful for harvesting new IAS occurrences when combined with machine vision tools (Cardoso et al., 2024). Thus, we created an automated workflow as proof-of-concept that mines geolocated images from Flickr using an invasive plants keyword search, and validates them as 'real' occurrences by sending them to the Pl@ntNet API for species identification. Unique observations were first filtered by retaining pictures taken at different times and or places. Next, we identified high-quality observations where the species name queried through the API matched with a species in the top five responses of the Pl@ntNet machine vision model (this ruleset can be modified). By doing so, this workflow enables harvesting of 'high-quality' occurrences across the EU for a wide variety of plants (see e.g. example for *Reynoutria japonica* below; Fig. 7), with the primary bottlenecks being i) species data availability and image quality on Flickr and ii) the identification accuracy of the Pl@ntNet model. Given the modular approach of this software package, it can easily be expanded to include other machine vision models to cover more taxa. Notably, because the Pl@ntNet model is also still trained on pictures labelled with previously accepted species names and synonyms (e.g., *Fallopia japonica*), these should all be included in the validation dictionary to not reject correct identifications due to taxonomic inconsistencies. Despite these shortcomings, this type of workflow could serve as a low-cost extra data source for IAS occurrences when the original data are processed in a secure and GDPR-compliant fashion (see section 4.3).

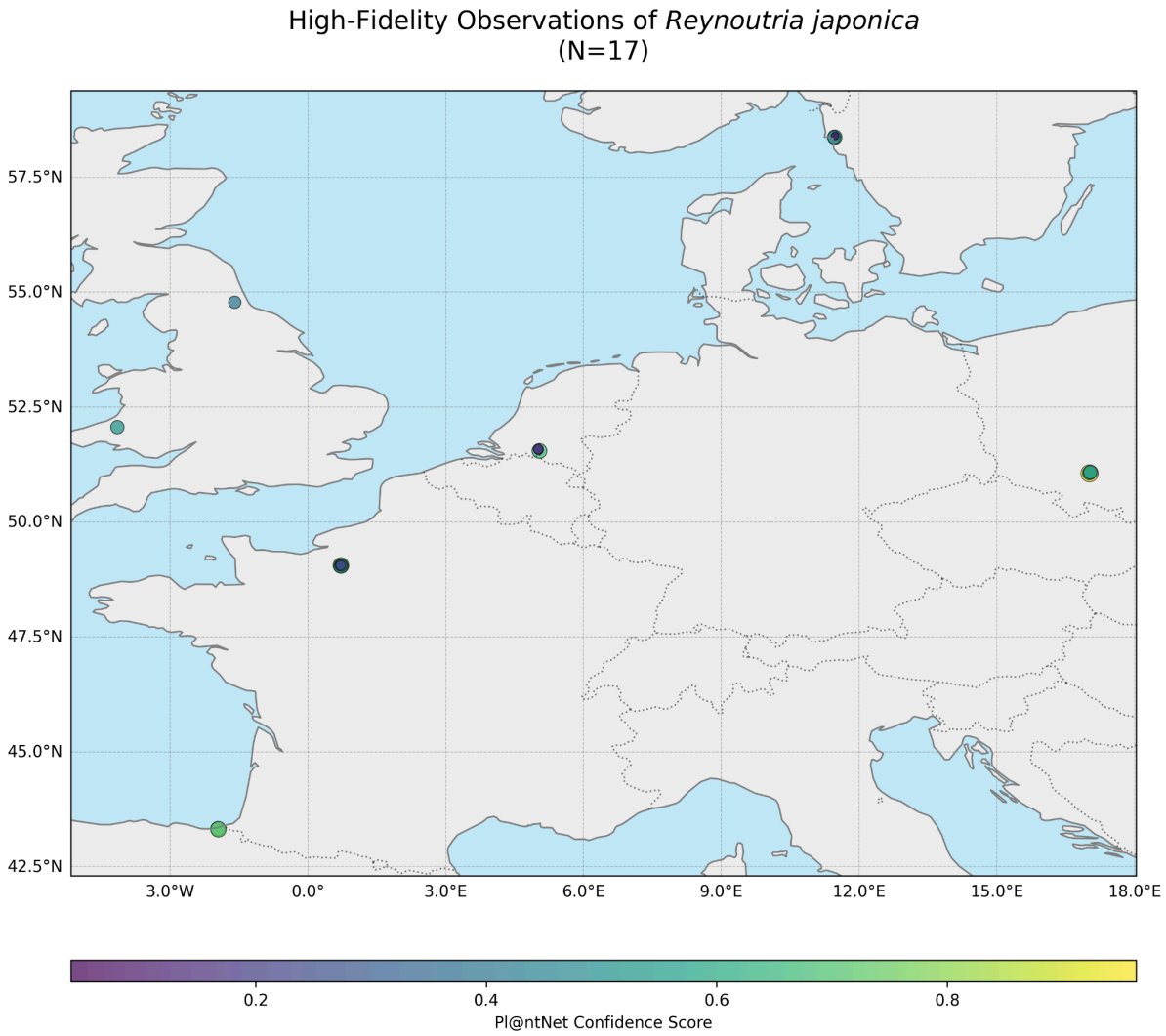

Figure 7: Pl@ntNet verified Flickr images of *Reynoutria japonica* across the EU with the confidence score given by the machine vision model. Note that some observations overlap, resulting in fewer than N unique points on the final figure.

4. Recommendations for future use of iEcology to monitor invasive alien species

4.1. Remove barriers to low-cost researcher internet platform data and API access

After exploring available data sources for the automated mining of iEcology data, we found that only six platforms (eBay, Flickr, Facebook, Google Trends, YouTube and Wikipedia) out of the 14 explored were able to provide geolocated iEcology data at no cost, despite that other platforms (e.g., X) provided such services to researchers in the past (see e.g., Tomojiri & Takaya,2025). This was despite the mention within the API applications that our research focused on 'investigations of IAS-related risks (e.g., public health, illegal pet trade) derived from and related to internet platform posts, mentions and activity data', as outlined within the EU legislation of Article 40(12) of the EU Digital Services Act (DSA). Many platforms still rejected our researcher applications because of i) disagreement that IAS-related risks could

directly emerge from the platform's data (X), ii) lack of proven 'safe data storage and processing measures in place' (Google Search Result API), (iii) unintended use case for the platform (Google Ads) and iv) lack of affiliation with a university (TikTok). Moreover, while access to Facebook data through the SOMAR portal was still free for academic researchers in 2025, new fees coming into effect in March 2026 indicate a one-time 1000 dollar 'registration' fee combined with a monthly API access subscription fee of ~ $400 dollar/month. This trend, limiting free (researcher) API access possibilities, is spreading across many platforms and e.g., also reflected in the May 2025 decision to make high-resolution Flickr images no longer available through the free tier of their API. While most of these platforms still provide high-quality paid API access to researchers, putting a price on such data does not only prevent researchers from studying platform associated risks due to lack of sufficient budget but also undermines the open and repeatable nature of iEcology workflows and research in general.

To resolve these issues, we propose a few measures. First, given the lack of explicit mention regarding low cost researcher access to internet platform data within the DSA, leading to internet platform misconduct (see e.g., the recent case against X also related to unreasonable pricing for researcher API access; European Commission, 2025), there is an urgent need to legally clarify this aspect in the DSA guidelines. Second, since current 'systemic risks from digital data' needed for researcher API access are only explicitly defined in relation to public health, mention of environmental risks should also be incorporated within the DSA text. In other words, the DSA legislation should be clarified to actually enable low-cost investigation of systemic EU risks using internet platform data beyond what can directly be caused by the algorithms/data itself. Third, it is important that platforms provide sufficient data for performing meaningful analysis of risks, as long as sufficient commitments to protection of sensitive data in line with GDPR are in place. While fine-grained IP-derived geolocation is immediately available to those advertising on platforms such as Facebook, researchers are only allowed to extract normalized counts of Facebook posts mentioning the IAS of interest with country-level geolocation from the SOMAR SRE (Reynaert et al., 2026). This strongly limits the use of this data for e.g., national-level risk management and creates a two-tiered system, where the parties least likely to exploit sensitive data are blocked from accessing it. Finally, the DSA should better define which parties are allowed to gain low-cost access to such data so that systemic risks from internet platform data can not only be studied by academics from universities, but also by e.g., government-associated researchers working for biosecurity agencies.

4.2. Improve automated data validation capacities

Given the lack of reliable georeferenced data streaming through API for many platforms, the utility of these data sources for tracking IAS dynamics without extensive data processing and external validation with other sources (e.g., GBIF, EASIN) remains limited. Especially for smaller platforms (e.g., Mastodon), the relatively small number of active users likely increases platform-specific biases regarding the context in which IAS are mentioned, which could both positively or negatively affect data quality. Previous studies have circumvented these issues by manually verifying observations and selecting 'high fidelity' geotagged observations based on expert opinion (Allain, 2019; Canavan et al., 2025; Chowdhury et al., 2024; Edwards et al., 2022), but this process is too time-consuming and inefficient for processing large volumes of data as required within the context of early warning systems and actively mining IAS occurrences across platforms, countries and taxa. Nonetheless, one way of limiting manual labor and reducing the number of false positives could be to not only query for species names in API queries, but also simultaneously include IAS-related

keywords (e.g., 'invasive') or location names. While searching for matches using extended dictionaries may increase data volumes collected (by e.g., collecting data on species synonyms), our investigation showed that it also resulted in proportionally more irrelevant data points. This is because many vernacular species names have multiple meanings and many APIs (e.g., YouTube) perform fuzzy matching with queries in the back end. One way to somewhat circumvent this issue is by collecting more metadata tags and performing platform specific filtering afterwards. In this way, a larger proportion of irrelevant mentions may be circumvented. However, another way of improving the utility of unstructured text platform data lacking geolocation tags could be by training LLM's or other machine learning tools to more accurately extract IAS and location information from free text posts and mentions (Castro et al., 2024). By doing so, both the data processing step, geolocation steps and final identification step (using machine vision models) could be fully automated.

Even for those platforms that provided geolocation information, our case studies showed that internet platforms providing general IAS-related internet activity data (i.e., Google Trends and Wikipedia) are relatively poor predictors of country/state level IAS expansions and occurrence patterns due to technical drawbacks in platform-specific internet activity calculations and various biases within the data itself (Alam & Hulme, 2026; Reynaert et al., 2026). In addition, tracking of individual post/video/mention volumes over time (i.e., Flickr, Facebook and YouTube) also performed poorly as a predictor of country-level occurrence patterns and IAS expansions (Reynaert et al., 2026). Nonetheless, platforms providing verifiable IAS-related records (Flickr, Facebook, Youtube and eBay) show potential to be a useful extra data source complementing more traditionally used data sources (e.g., EASIN, GBIF) when combined with automated IAS occurrence validation methods (see section 3.3; Cardoso et al., 2024). Within this context, Flickr and YouTube were most promising, given that they provide high volumes of coordinate-level geolocation IAS-related mentions and are the most open regarding data extraction and processing, which is necessary for setting up automated IAS occurrence validation workflows (Table 2). While Facebook allowed the extraction of a high volume of 'summarized' data metrics, collecting images associated with geolocated posts is currently not permitted. Nonetheless, regarding species groups for which only a few and fragmented databases are available (e.g., pathogens), tracking general internet activity patterns in a more automated fashion presents an underexplored opportunity for more systematic research, despite its many shortcomings.

Table 2: Qualitative characteristics of the most promising low-cost iEcology platforms with geolocation including geolocation detail, temporal resolution, data type(s) and automated occurrence validation potential.

Platform	Geolocation detail	Temporal resolution	Data type	Automated IAS occurrence validation potential
Wikipedia	Country	Daily	Pageview activity	Low
Google Trends	Country/region/city	Hourly	Search activity	Low
YouTube	Full coordinates	Hourly	Thumbnails / videos / comments	High
Facebook	Country	Hourly	Posts / comments	Medium
Flickr	Full coordinates	Hourly	Pictures / comments	High
eBay	Country/region/city	Hourly	Live listings	Medium

Other issues currently also hamper the full automation of species identification workflows. First, not only do most available machine vision models focus on a limited number of taxa (e.g., Pl@ntNet, InsectAI), their accuracy of identification is also model and species-specific. Many high-quality species identification models also lack a low-cost API, requiring users to either combine multiple models per taxon, develop models themselves, run pre-trained models on local hardware (e.g., iNaturalist), or pay high API access fees. All these aspects reduce accessibility to automated species identification significantly, highlighting the need for more generalized and open IAS-focused machine vision models with API capacities (Jakuschona et al., 2022). Second, the amount of AI-generated content online has increased exponentially over the past few years, which was exemplified by e.g., many of our mined YouTube videos on IAS being fully AI generated. As artificial video and image generation techniques improve rapidly, it remains unclear whether AI machine vision models will continue to be able to accurately distinguish between real and artificially generated species images and videos, further complicating fully automating species identification processes (Baraheem & Nguyen, 2023).

4.3. Reconsider the GDPR-related iEcology research context

iEcology data are by definition secondary data extracted from online internet activity of users (i.e., 'legal persons'), often utilized outside of the user-consented context and thus requiring ethical and legal considerations. Within the EU iEcology research context, the General Data Protection Regulation (GDPR) dictates how and when such data (even when publicly available) can be processed and published. In an ideal world, iEcologists would obtain written permission from every user whose data they extract to conduct research. However, given the massive data volumes involved in most iEcology research, this is often unfeasible. Hence, researchers should implement specific measures, based on the context in which the research is conducted, by carefully adhering to the FAIR principle of being as open as possible but as closed as necessary to protect sensitive information. In practice, this means processing information within a closed and secure research environment, minimizing data retention, and anonymizing individual identifiers and post/mention/video/image location (e.g., by adding a noisy radius around coordinates) whenever any data are made publicly available. In some cases (e.g., for research necessary within the context of 'public importance') exceptions to this ruleset may apply, but these are usually institute and research specific, requiring special legal permissions.

These guidelines allow the use of iEcology data in specific contexts, e.g., to create country and European level species checklists. However, they also limit its applicability for many use cases, such as the to coordinate level mining and publishing of IAS occurrences to GBIF and use of coordinate level data for species distribution modelling and mapping. Previous studies also indicate that this ambiguity regarding how to weigh gaining information regarding potential environmental risks vs protection of sensitive user information is often confusing for researchers (Ghermandi et al., 2023). We therefore argue that guidelines regarding the use of iEcology data within the context of investigating and mapping environmental risks to society should be clarified to maximize societal benefits from these data.

5. Conclusions

We explored whether iEcology data from various internet platforms can serve as an automated, low-cost source for IAS monitoring and early warning systems. Our investigation indicates that while general tracking of IAS internet activity is a poor indicator of IAS occurrences and expansions, mining individual observations from platforms that allow automated validation of underlying data shows potential as a complementary data source for

monitoring. However, further increasing the technological readiness level of iEcology and maximizing societal benefits from iEcology data regarding IAS dynamics across the EU requires:

- removing barriers to researcher access to platform data and API's;
- refining automated data validation capacities filtering out bias and low-quality content;
- clarifying the use of iEcology data within a GDPR-safe context given it currently complicates the use and publishing of coordinate-level observations.

6. Acknowledgements

We thank EASIN for providing the first occurrence in EU dataset for all species from the UL.

7. References

Alam, M. A., & Hulme, P. E. (2026). Can Google trends be used to estimate the geographic distribution of alien plants in the United States? *Ecological Informatics*, *95*, 103689. https://doi.org/10.1016/j.ecoinf.2026.103689

Allain, S. (2019). Mining Flickr: A method for expanding the known distribution of invasive species. *Herpetological Bulletin*, (148, Summer 2019), 11–14. https://doi.org/10.33256/hb148.1114

Balestrieri, R., Vento, R., Viviano, A., Mori, E., Gili, C., & Monti, F. (2023). Razorbills Alca torda in Italian Seas: A Massive Irruption of Historical Relevance and Role of Social Network Monitoring. *Animals*, *13*(4), Article 4. https://doi.org/10.3390/ani13040656

Baraheem, S. S., & Nguyen, T. V. (2023). AI vs. AI: Can AI Detect AI-Generated Images? *Journal of Imaging*, *9*(10), 199. https://doi.org/10.3390/jimaging9100199

Bhatt, P., & Pickering, C. M. (2021). Public Perceptions about Nepalese National Parks: A Global Twitter Discourse Analysis. *Society & Natural Resources*, *34*(6), 685–702. https://doi.org/10.1080/08941920.2021.1876193

Bisbee, J., & Munger, K. (2025). The Vibes Are Off: Did Elon Musk Push Academics Off Twitter? *PS: Political Science & Politics*, *58*(1), 139–146. https://doi.org/10.1017/S1049096524000416

Bryce, B. (2025). *The Reddit Instance—PRAW 7.7.1 documentation*. https://praw.readthedocs.io/en/stable/code_overview/reddit_instance.html

Callaghan, C. T., Poore, A. G. B., Mesaglio, T., Moles, A. T., Nakagawa, S., Roberts, C., Rowley, J. J. L., VergÉs, A., Wilshire, J. H., & Cornwell, W. K. (2021). Three Frontiers for the Future of Biodiversity Research Using Citizen Science Data. *BioScience*, *71*(1), 55–63. https://doi.org/10.1093/biosci/biaa131

Canavan, S., Rodríguez, J., Gervazoni, P., Pipek, P., Le Roux, J. J., Castillo, M. L., Lieurance, D., Maříková-Moodley, D., Pyšek, P., & Novoa, A. (2025). iEcology reveals the importance of geography and genetic makeup in the flowering phenology of

invasive Carpobrotus taxa. *Ecological Solutions and Evidence*, *6*(4), e70122. https://doi.org/10.1002/2688-8319.70122

Cardoso, A. S., Malta-Pinto, E., Tabik, S., August, T., Roy, H. E., Correia, R., Vicente, J. R., & Vaz, A. S. (2024). Can citizen science and social media images support the detection of new invasion sites? A deep learning test case with *Cortaderia selloana*. *Ecological Informatics*, *81*, 102602. https://doi.org/10.1016/j.ecoinf.2024.102602

Castro, A., Pinto, J., Reino, L., Pipek, P., & Capinha, C. (2024). Large language models overcome the challenges of unstructured text data in ecology. *Ecological Informatics*, *82*, 102742. https://doi.org/10.1016/j.ecoinf.2024.102742

Cerri, J., Carnevali, L., Monaco, A., Genovesi, P., & Bertolino, S. (2022). Blacklists do not necessarily make people curious about invasive alien species. A case study with Bayesian structural time series and Wikipedia searches about invasive mammals in Italy. *NeoBiota*, *71*, 113–128. https://doi.org/10.3897/neobiota.71.69422

Chowdhury, S., Fuller, R. A., Ahmed, S., Alam, S., Callaghan, C. T., Das, P., Correia, R. A., Di Marco, M., Di Minin, E., Jarić, I., Labi, M. M., Ladle, R. J., Rokonuzzaman, M., Roll, U., Sbragaglia, V., Siddika, A., & Bonn, A. (2024). Using social media records to inform conservation planning. *Conservation Biology*, *38*(1), e14161. https://doi.org/10.1111/cobi.14161

Commission Implementing Regulation (EU) 2022/1203 of 12 July 2022 Amending Implementing Regulation (EU) 2016/1141 to Update the List of Invasive Alien Species of Union Concern, 186 OJ L (2022). http://data.europa.eu/eli/reg_impl/2022/1203/oj/eng

Correia, R. A. (2024). gtrendsAPI: An R wrapper for the Google Trends API. *Software Impacts*, *20*, 100634. https://doi.org/10.1016/j.simpa.2024.100634

Dylewski, Ł., Mikula, P., Tryjanowski, P., Morelli, F., & Yosef, R. (2017). Social media and scientific research are complementary—YouTube and shrikes as a case study. *The Science of Nature*, *104*(5–6), 48. https://doi.org/10.1007/s00114-017-1470-8

eBay. (2026). *API License Agreement | eBay Developers Program*. https://developer.ebay.com/join/api-license-agreement

Edwards, T., Jones, C. B., & Corcoran, P. (2022). Identifying wildlife observations on twitter. *Ecological Informatics*, *67*, 101500. https://doi.org/10.1016/j.ecoinf.2021.101500

Edwards, T., Jones, C. B., Perkins, S. E., & Corcoran, P. (2021). Passive citizen science: The role of social media in wildlife observations. *PLOS ONE*, *16*(8), e0255416. https://doi.org/10.1371/journal.pone.0255416

El bakkali, L. (2023, March 9). Vlaamse overheid blokkeert toegang tot TikTok op computers en smartphones van personeel | VRT NWS: Nieuws. *VRTNWS*. https://www.vrt.be/vrtnws/nl/2023/03/09/vlaamse-overheid-blokkeert-toegang-tot-tiktok-op-computers-en-sm/

EU. (2022). *Article 34, the Digital Services Act (DSA)*. https://www.eu-digital-services-act.com/Digital_Services_Act_Article_34.html

European Comission. (2025). *Commission fines X €120 million under the Digital Services Act* [Text]. European Commission - European Commission. https://ec.europa.eu/commission/presscorner/detail/en/ip_25_2934

Failla, A., & Rossetti, G. (2024). "I'm in the Bluesky Tonight": Insights from a year worth of social data. *PLOS ONE*, *19*(11), e0310330. https://doi.org/10.1371/journal.pone.0310330

Fox, N., Graham, L. J., Eigenbrod, F., Bullock, J. M., & Parks, K. E. (2021). Reddit: A novel data source for cultural ecosystem service studies. *Ecosystem Services*, *50*, 101331. https://doi.org/10.1016/j.ecoser.2021.101331

GBIF. (2025). *iNaturalist Research-grade Observations*. https://www.gbif.org/dataset/50c9509d-22c7-4a22-a47d-8c48425ef4a7

GBIF. (2026). *Pygbif* [Python]. https://github.com/gbif/pygbif (Original work published 2014)

Ghermandi, A., Langemeyer, J., Berkel, D. V., Calcagni, F., Depietri, Y., Vigl, L. E., Fox, N., Havinga, I., Jäger, H., Kaiser, N., Karasov, O., McPhearson, T., Podschun, S.,

Ruiz-Frau, A., Sinclair, M., Venohr, M., & Wood, S. A. (2023). Social media data for environmental sustainability: A critical review of opportunities, threats, and ethical use. *One Earth*, *6*(3), 236–250. https://doi.org/10.1016/j.oneear.2023.02.008

Google. (2025a). *About the search terms report—Google Ads Help*. https://support.google.com/google-ads/answer/2472708?sjid=10754817782390736316-EU

Google. (2025b). *FAQ about Google Trends data—Trends Help*. https://support.google.com/trends/answer/4365533?hl=en

Heinrich, S., Ross, J. V., & Cassey, P. (2019). Of cowboys, fish, and pangolins: US trade in exotic leather. *Conservation Science and Practice*, *1*(8), e75. https://doi.org/10.1111/csp2.75

Henke, T., Novoa, A., Bárðarson, H., & Ólafsdóttir, G. Á. (2024). Let's talk aliens—Stakeholder perceptions of an alien species differ in time and space. *NeoBiota*, *93*, 117–141. https://doi.org/10.3897/neobiota.93.117200

Humair, F., Humair, L., Kuhn, F., & Kueffer, C. (2015). E-commerce trade in invasive plants. *Conservation Biology*, *29*(6), 1658–1665. https://doi.org/10.1111/cobi.12579

Jakuschona, N., Niers, Tom, Stenkamp, Jan, Bartoschek, Thomas, Cardoso, Ana Cristina, Schade, Sven, Cardoso, Ana Cristina, & Schade, Sven. (2022). *Evaluating image-based species recognition models suitable for citizen science application to support European invasive alien species policy*. Publications Office of the European Union. https://data.europa.eu/doi/10.2760/97305

Jarić, I., Bellard, C., Correia, R. A., Courchamp, F., Douda, K., Essl, F., Jeschke, J. M., Kalinkat, G., Kalous, L., Lennox, R. J., Novoa, A., Proulx, R., Pyšek, P., Soriano–Redondo, A., Souza, A. T., Vardi, R., Veríssimo, D., & Roll, U. (2021). Invasion Culturomics and iEcology. *Conservation Biology*, *35*(2), 447–451. https://doi.org/10.1111/cobi.13707

Kalous, L., Nechanská, D., & Petrtýl, M. (2018). Survey of angler's internet posts confirmed the occurrence of freshwater fishes of the genus Ictiobus (Rafinesque, 1819) in natural waters of Czechia. *Knowledge & Management of Aquatic Ecosystems*, (419), Article 419. https://doi.org/10.1051/kmae/2018019

Kourantidou, M., Haubrock, P. J., Cuthbert, R. N., Bodey, T. W., Lenzner, B., Gozlan, R. E., Nuñez, M. A., Salles, J.-M., Diagne, C., & Courchamp, F. (2022). Invasive alien species as simultaneous benefits and burdens: Trends, stakeholder perceptions and management. *Biological Invasions*, *24*(7), 1905–1926. https://doi.org/10.1007/s10530-021-02727-w

López-Guillén, E., Herrera, I., Bensid, B., Gómez-Bellver, C., Ibáñez, N., Jiménez-Mejías, P., Mairal, M., Mena-García, L., Nualart, N., Utjés-Mascó, M., & López-Pujol, J. (2024). Strengths and Challenges of Using iNaturalist in Plant Research with Focus on Data Quality. *Diversity*, *16*(1), Article 1. https://doi.org/10.3390/d16010042

Mittermeier, J. C., Roll, U., Matthews, T. J., & Grenyer, R. (2019). A season for all things: Phenological imprints in Wikipedia usage and their relevance to conservation. *PLOS Biology*, *17*(3), e3000146. https://doi.org/10.1371/journal.pbio.3000146

Morais, P., Afonso, L., & Dias, E. (2021). Harnessing the Power of Social Media to Obtain Biodiversity Data About Cetaceans in a Poorly Monitored Area. *Frontiers in Marine Science*, *8*, 765228. https://doi.org/10.3389/fmars.2021.765228

Nascimento, L. S., Noernberg, M. A., Bleninger, T. B., Lindner, A., & Nogueira Júnior, M. (2024). Not such a rare species, after all? Insights into Drymonema gorgo Müller 1883 (Cnidaria, Scyphozoa), a large and little-known jellyfish from Brazil. *Aquatic Ecology*, *58*(1), 17–30. https://doi.org/10.1007/s10452-023-10074-2

Novoa, A., Canavan, S., Jarić, I., Pipek, P., & Pyšek, P. (2022). Musk's Twitter takeover jeopardizes culturomics. *Nature*, *612*(7939), 211–211. https://doi.org/10.1038/d41586-022-04361-5

Novoa, A., Jarić, I., Pipek, P., & Pyšek, P. (2025). Culturomics and iEcology provide novel opportunities to study human and social dimensions of alien species introductions. *Trends in Ecology & Evolution*, *40*(1), 18–26. https://doi.org/10.1016/j.tree.2024.08.012

Olteanu, A., Castillo, C., Diaz, F., & Kıcıman, E. (2019). Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. *Frontiers in Big Data*, *2*. https://doi.org/10.3389/fdata.2019.00013

Persico, S. (2024). Affective, defective, and infective narratives on social media about nuclear energy and atomic conflict during the 2022 Italian electoral campaign. *Humanities and Social Sciences Communications*, *11*(1), 1–14. https://doi.org/10.1057/s41599-024-02676-4

Pytrends. (2025). https://pypi.org/project/pytrends/

Reddit. (2025). *reddit.com: Api documentation*. https://www.reddit.com/dev/api#GET_subreddits_search

Reynaert, S., Billiet, N., Pipek, P., Novoa, A., Hulme, P., Meeus, S., & Groom, Q. (2026). *Can data mining from various internet platforms systematically accelerate detection of alien species invasions across the EU?* (p. 2026.02.06.704325). bioRxiv. https://doi.org/10.64898/2026.02.06.704325

Schifani, E., & Paolinelli, R. (2018). Los foros y las redes sociales ayudan a descubrir especies exóticas en Europa y monitorear su propagación: El caso de Exaireta spinigera (Wiedemann, 1830) (Diptera, Stratiomyidae) en la península italiana y Sicilia. *Graellsia*, *74*(2), Article 2. https://doi.org/10.3989/graellsia.2018.v74.213

Schuetz, J. G., & Johnston, A. (2019). Characterizing the cultural niches of North American birds. *Proceedings of the National Academy of Sciences*, *116*(22), 10868–10873. https://doi.org/10.1073/pnas.1820670116

Sweet, A. M. (2025). *SOMAR Data Access Application Guide—SOMAR Public Documentation—SOMAR Confluence*.

https://somar.atlassian.net/wiki/spaces/somardocs/pages/249397299/SOMAR+Data+Access+Application+Guide#Meta-Content-Library-and-Content-Library-API

Theophilos, J. A. (2024). Closing the Door to Remain Open: The Politics of Openness and the Practices of Strategic Closure in the Fediverse. *Social Media + Society*, *10*(4), 20563051241308323. https://doi.org/10.1177/20563051241308323

Tomojiri, D., & Takaya, K. (2023). *Aspects of public attention on popular nonindigenous species, as determined by a comprehensive assessment of Japanese social media*. https://doi.org/10.21203/rs.3.rs-3790755/v1

Tomojiri, D., & Takaya, K. (2025). Aspects of public attention on the most mentioned nonindigenous species, as determined by a comprehensive assessment of Japanese social media. *Biological Invasions*, *27*(1), 62. https://doi.org/10.1007/s10530-024-03520-1

Vardi, R., Mittermeier, J. C., & Roll, U. (2021). Combining culturomic sources to uncover trends in popularity and seasonal interest in plants. *Conservation Biology*, *35*(2), 460–471. https://doi.org/10.1111/cobi.13705

Wright, M. (2023). *MW Geofind Wiki*. GitHub. https://github.com/mattwright324/youtube-geofind/wiki/Home

X. (2025a). *About the X API*. X. https://docs.x.com/x-api/getting-started/about-x-api

X. (2025b). *X DSA Researcher Application*. Google Docs. https://docs.google.com/forms/d/e/1FAIpQLSdo0O-D6Kxa3cV4g1JLz2T_0Sk3hdEnTdv8dJmibagCnzJ7kg/viewform?usp=embed_facebook

Youtube. (2025). *Search: List | YouTube Data API | Google for Developers*. https://developers.google.com/youtube/v3/docs/search/list#location

Zachte, E. (2018, October 25). *Wikimedia Traffic Analysis Report*. https://stats.wikimedia.org/wikimedia/squids/SquidReportPageViewsPerLanguageBreakdown.htm

8. Annex

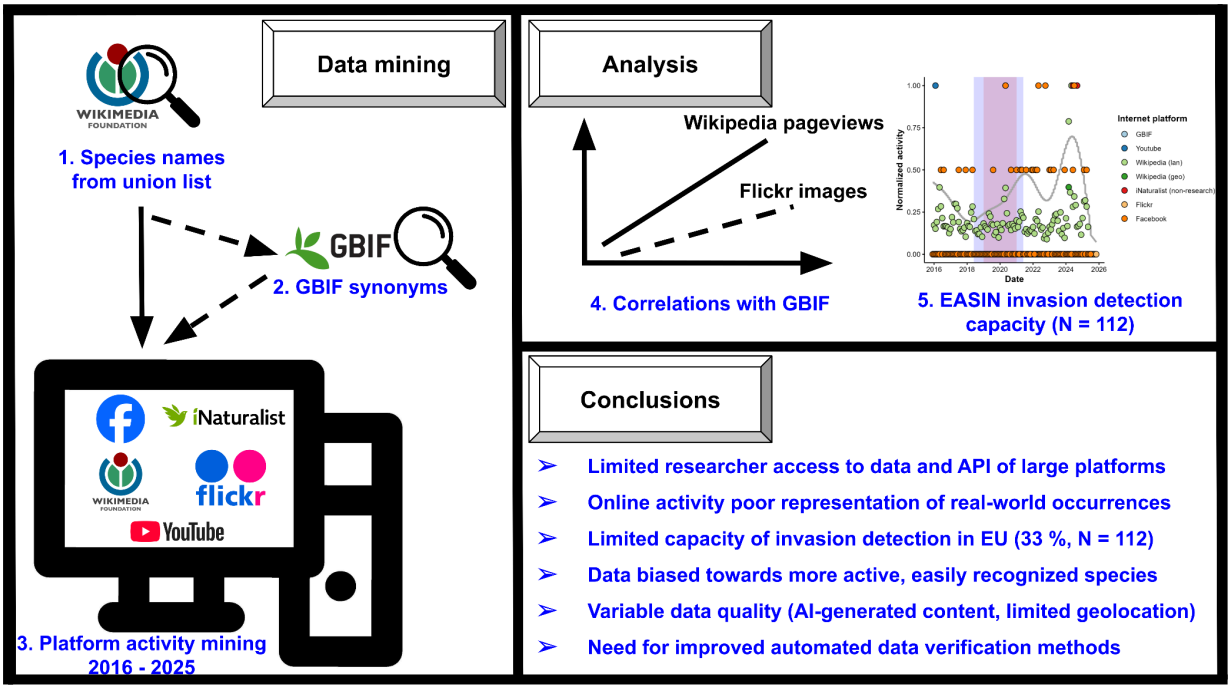

Figure S1: Graphical overview of case study II based on Reynaert et al. (2026). The software package capacities (excluding the Flickr to Pl@ntNet pipeline) are shown in the 'Data mining' panel.