Social media mining as an opportunistic citizen science model in ecological monitoring: a case study using invasive alien species in forest ecosystems.

Dissertation

zur Erlangung des akademischen Grades Doctor of Philosophy (PhD)

der Fakultät für Forstwissenschaften und Waldökologie

der Georg-August-Universität Göttingen

vorgelegt von

Stefan Daume

geboren in Frankenberg (Eder)

Göttingen, 2015

1. Gutachter: Prof. Dr. Dr. h.c. Klaus von Gadow

2. Gutachter: Prof. Dr. Winfried Kurth

3. Gutachter: Prof. Dr. Jürgen Nagel

Tag der Disputation: 27.08.2015

Social media mining as an opportunistic citizen science model in ecological monitoring: a case study using invasive alien species in forest ecosystems.

Stefan Daume

Doctoral Thesis, Faculty of Forest Sciences and Forest Ecology, Georg-August-University Göttingen, Germany 2015

To my family

„We can only see a short distance ahead, but we can see plenty there that needs to be done.“

Alan M. Turing (1950). Computing machinery and intelligence. Mind, 59, 433-460.

Contents

List of papers

Paper I **Daume, S.**, Albert, M., von Gadow, K. (2014). Forest monitoring and social media – Complementary data sources for ecosystem surveillance? Forest Ecology and Management, 316, pp.9–20. doi: 10.1016/j.foreco.2013.09.004

Paper II **Daume, S.**, Albert, M., von Gadow, K. (2014). Assessing citizen science opportunities in forest monitoring using probabilistic topic modelling. Forest Ecosystems, 1(1), p.11. doi:10.1186/s40663-014-0011-6

Paper III **Daume, S.** (Manuscript). Mining Twitter to monitor invasive alien species – An analytical framework and sample information topologies.

Paper IV **Daume, S.**, Galaz, V. (Under review at PLOS One). "Anyone know what species this is?" – Twitter conversations as embryonic citizen science communities.

These four papers constitute the main part of the dissertation and will be referred to as **Paper I** to **Paper IV** in this extended dissertation summary.

Individual reference sections are included with each paper and the extended dissertation summary.

Where applicable, author contributions are listed in the respective articles and manuscripts.

Abstract

Major environmental, social and economic changes threatening the resilience of ecosystems world-wide and new demands on a broad range of forest ecosystem services present new challenges for forest management and monitoring. New risks and threats such as invasive alien species imply fundamental challenges for traditional forest management strategies, which have been based on assumptions of permanent ecosystem stability. Adaptive management and monitoring is called for to detect new threats and changes as early as possible, but this requires large-scale monitoring and monitoring resources remain a limiting factor. Accordingly, forest practitioners and scientists have begun to turn to public support in the form of "citizen science" to react flexibly to specific challenges and gather critical information.

The emergence of ubiquitous mobile and internet technologies provides a new digital source of information in the form of so-called social media that essentially turns users of these media into environmental sensors and provides an immense volume of publicly accessible, ambient environmental information. Mining social media content, such as Facebook, Twitter, Wikis or Blogs, has been shown to make critical contributions to epidemic disease monitoring, emergency management or earthquake detection. Applications in the ecological domain remain anecdotal and a methodical exploration for this domain is lacking.

Using the example of the micro-blogging service Twitter and invasive alien species in forest ecosystems, this study provides a methodical exploration and assessment of social media for forest monitoring. Social media mining is approached as an opportunistic citizen science model and the data, activities and contributors are analyzed in comparison to deliberate ecological citizen science monitoring.

The results show that Twitter is a valuable source of information on invasive alien species and that social media in general could be a supplement to traditional monitoring data. Twitter proves to be a rich source of primary biodiversity observations including those of the selected invasive species. In addition, it is shown that Twitter content provides distinctive thematic profiles that relate closely to key characteristics of the explored invasive alien species and provide valuable insights for invasive species management. Furthermore, the study shows that while there are underutilized opportunities for citizen science in forest monitoring, the contributors of biodiversity observations on Twitter show a more than

casual interest in this subject and represent a large pool of potential contributors to deliberate citizen science monitoring efforts.

In summary, social online media are a valuable source for ecological monitoring information in general and deserve intensified exploration to arrive at operational systems supporting real-time risk assessments.

Keywords: *forest ecosystems, forest monitoring, forest threats, citizen science, invasive alien species, social media, social media mining, Twitter, ecological monitoring, biodiversity monitoring.*

Zusammenfassung

Dramatische ökologische, ökonomische und soziale Veränderungen bedrohen die Stabilität von Ökosystemen weltweit und stellen zusammen mit neuen Ansprüchen an die vielfältigen Ökosystemdienstleistungen von Wäldern neue Herausforderungen für das forstliche Management und Monitoring dar. Neue Risiken und Gefahren, wie zum Beispiel eingebürgerte invasive Arten (Neobiota), werfen grundsätzliche Fragen hinsichtlich etablierter forstlicher Managementstrategien auf, da diese Strategien auf der Annahme stabiler Ökosysteme basieren. Anpassungsfähige Management- und Monitoringstrategien sind deshalb notwendig, um diese neuen Bedrohungen und Veränderungen frühzeitig zu erkennen. Dies erfordert jedoch ein großflächiges und umfassendes Monitoring, was unter Maßgabe begrenzter Ressourcen nur bedingt möglich ist. Angesichts dieser Herausforderungen haben Forstpraktiker und Wissenschaftler begonnen auch auf die Unterstützung von Freiwilligen in Form sogenannter „Citizen Science"-Projekte (Bürgerwissenschaft) zurückzugreifen, um zusätzliche Informationen zu sammeln und flexibel auf spezifische Fragestellungen reagieren zu können.

Mit der allgemeinen Verfügbarkeit des Internets und mobiler Geräte ist in Form sogenannter sozialer Medien zudem eine neue digitale Informationsquelle entstanden. Mittels dieser Technologien übernehmen Nutzer prinzipiell die Funktion von Umweltsensoren und erzeugen indirekt ein ungeheures Volumen allgemein zugänglicher Umgebungs- und Umweltinformationen. Die automatische Analyse von sozialen Medien wie Facebook, Twitter, Wikis oder Blogs, leistet inzwischen wichtige Beiträge zu Bereichen wie dem Monitoring von Infektionskrankheiten, Katastrophenschutz oder der Erkennung von Erdbeben. Anwendungen mit einem ökologischen Bezug existieren jedoch nur vereinzelt, und eine methodische Bearbeitung dieses Anwendungsbereichs fand bisher nicht statt.

Unter Anwendung des Mikroblogging-Dienstes Twitter und des Beispiels eingebürgerter invasiver Arten in Waldökosystemen, verfolgt die vorliegende Arbeit eine solche methodische Bearbeitung und Bewertung sozialer Medien im Monitoring von Wäldern. Die automatische Analyse sozialer Medien wird dabei als opportunistisches „Citizen Science"-Modell betrachtet und die verfügbaren Daten, Aktivitäten und Teilnehmer einer vergleichenden Analyse mit existierenden bewusst geplanten „Citizen Science"-Projekten im Umweltmonitoring unterzogen.

Die vorliegenden Ergebnisse zeigen, dass Twitter eine wertvolle Informationsquelle über invasive Arten darstellt und dass soziale Medien im Allgemeinen traditionelle Umweltinformationen ergänzen könnten. Twitter ist eine reichhaltige Quelle von primären Biodiversitätsbeobachtungen, einschließlich solcher zu eingebürgerten invasiven Arten. Zusätzlich kann gezeigt werden, dass die analysierten Twitterinhalte für die untersuchten Arten markante Themen- und Informationsprofile aufweisen, die wichtige Beiträge im Management invasiver Arten leisten können. Allgemein zeigt die Studie, dass einerseits das Potential von „Citizen Science" im forstlichen Monitoring derzeit nicht ausgeschöpft wird, aber andererseits mit denjenigen Nutzern, die Biodiversitätsbeobachtungen auf Twitter teilen, eine große Zahl von Individuen mit einem Interesse an Umweltbeobachtungen zur Verfügung steht, die auf der Basis ihres dokumentierten Interesses unter Umständen für bewusst geplante „Citizen Science"-Projekte mobilisiert werden könnten.

Zusammenfassend dokumentiert diese Studie, dass soziale Medien eine wertvolle Quelle für Umweltinformationen allgemein sind und eine verstärkte Untersuchung verdienen, letztlich mit dem Ziel, operative Systeme zur Unterstützung von Risikobewertungen in Echtzeit zu entwickeln.

Schlagwörter: *Waldökosysteme, forstliches Monitoring, Waldbedrohungen, Bürgerwissenschaft, eingebürgerte invasive Arten, Neobiota, soziale Medien, Analyse sozialer Medien, Twitter, Umweltmonitoring, Biodiversitätsmonitoring.*

Introduction

Motivation and background

Some of the oldest practices of deliberate, methodical and regular ecosystem monitoring were developed by forest scientists. Data on managed and unmanaged forest ecosystems is collected in different types of inventories, experiments or observational studies, primarily directed at resource assessments and varying with regard to the coverage of temporal scales as well as environmental conditions (Zhao et al., 2014). These data may be complemented by monitoring programs driven by specific properties or aspects of ecosystem functions, such as biodiversity, conservation or recreational value (Noss, 1999).

Typically, forest monitoring is implemented within institutional frameworks, applying tested and established methods that guarantee high-levels of data quality, replicability and reuse (Kleinn et al., 2010). However, amateur contributions have also always been part of ecological monitoring (Miller-Rushing et al., 2012; Silvertown, 2009) including forest ecosystems. One of the oldest continuous recordings on tree leafing times for example was started by Robert Marsham - the “father of phenology” - nearly 300 years ago, continued as an amateur effort and now provides insights into the likely effects of climate change on tree community compositions (Roberts et al., 2015).

Forest monitoring programs are the pillars of sustainable management strategies which historically are based on the assumption that ecosystems remain in a stable state. They have become even more important given the magnitude of global environmental changes (Rockström et al., 2009; Zalasiewicz et al., 2010) which can lead to unsuspected surprises and irreversible shifts (Scheffer et al., 2001). Such changes demand adaptive management strategies (Bolte et al., 2009), complex risk assessments (Albert et al., 2015), but also flexible monitoring responses (Lindenmayer and Likens, 2009) (**Paper I**). At the same time societal requirements on forest ecosystems are changing more rapidly (Gadow, 2013; Gadow et al., 2007), a broad range of ecosystem services provided by managed and unmanaged forests deserves consideration (Nasi et al., 2002) and in the context of global trends such as urbanization (UN-CBD, 2012) forests in urban areas will require more attention in the development of suitable forest management (Bolund and Hunhammar, 1999; Gadow, 2002) and monitoring (Kleinn et al., 2010) strategies.

In addition, holistic views and global strategies for the sustainable provisioning of critical ecosystem services (Millenium Ecosystem Assessment, 2005) as well as mounting national commitments to international frameworks like for example the Convention on Biological Biodiversity (UN-CBD, 2012) present new challenges in the allocation of limited monitoring resources in general (Wintle et al., 2010).

"Citizen science"

It is thus no coincidence that in particular with regard to new emerging threats forest scientists and practitioners turn increasingly to public support as a flexible means to supplement traditional monitoring programs. This includes the detection of forest pests (Rutledge et al., 2013), carbon stock estimates (Butt et al., 2013), urban tree monitoring (Roman et al., 2013) or long-term studies on pest resistance (Clark, 2013; Ingwell and Preisser, 2011) to name a few examples of volunteer contributions known as "citizen science" (Dickinson et al., 2012).

It is precisely the threat of sudden, often unexpected and possibly irreversible changes in ecosystems as well as the increasing number and significance of "citizen science" projects that motivated this study which explores the potential of social online media as alternative sources of ecological information.

Social online media

Social online media denote a type of web-based applications and information sources that exhibit features of social networks, where content is created by users of the media, often collaboratively within virtual communities (**Paper I**). Prominent examples include the social network *Facebook*, the micro-blogging service *Twitter* or the image sharing service *Instagram*, but the notion of social media extends to so-called *Blogs*, *Wikis* or content communities like *Youtube*.

Social media mining has received significant research attention in recent years and has proven to provide valuable contributions to critical fields such as public health monitoring (Mykhalovskiy and Weir, 2006), specifically early warnings (Achrekar et al., 2011) and prediction of trends (Culotta, 2010; Gomide et al., 2011). Social media are also utilized as a real-time data source in emergency response (Qu et al., 2011; Vieweg et al., 2010), the detection of earth quakes (Crooks et al., 2013; Earle et al., 2010), typhoons (Sakaki et al., 2010) or even to predict criminal activities (Wang et al., 2012).

With reference to these successful applications, comparable studies in the environmental domain have been called for (Galaz et al., 2010), but remain as yet anecdotal (Barve, 2014; Malcevschi et al., 2012; Stafford et al., 2010) (**Paper III & IV**). Furthermore, environmental applications lack a clear theoretical and methodological framework to establish this type of information as a recognized source in ecosystem surveillance, both with regard to research and practical use.

This study approaches social media mining as an opportunistic form of citizen science, thus linking it to a conceptual framework within which this data can be assessed, explored and evaluated further.

Main objectives of this study

While applications of social media mining in critical domains are now more frequent, applications with relevance for environmental monitoring are rare and specific examples with relevance for forest ecosystems are yet unknown. This study thus aims to develop a basis for future uses of this type of information source in the monitoring of forests and addresses the following major research questions:

1. Do social media contain useful information with relevance to the management and monitoring of forest ecosystems?
2. How can this information be obtained, how abundant is it, how can it be characterized, and what are its quality and relevance?
3. How does this information relate to existing utilized ecological monitoring sources, specifically to informal monitoring efforts such as citizen science?
4. What are the conceptual and practical limitations of this data source?

Structure of the thesis

The main part of this thesis consists of four papers which are referred to as **Paper I** to **Paper IV**.

Paper I provides both an introduction into the subject of social online media as well as an outlook on possible applications of social media mining for forest monitoring. The paper discusses social media in the context of citizen science, elaborates on the choice of Twitter and invasive alien species (IAS) as examples to frame the pursued research, and introduces

results for one species (*Oak processionary moth*), which is revisited in more depth in **Paper III**.

Acknowledging that examples of citizen science efforts exist which benefit forest monitoring but are sparse, **Paper II** attempts a high-level assessment of principal citizen science opportunities in forest monitoring through an analysis of the topical overlaps in 20 years of published research literature on the two subjects. Probabilistic topic modelling is employed for an analysis of 1015 documents to extract the prevailing themes in both areas, identify topical overlaps and assess the utilization of citizen science in forest monitoring.

Paper III summarizes nearly three years of Twitter data collected on 11 IAS, that directly or indirectly impact forest ecosystems. Detailed results of a manual classification of Twitter messages for three sample species are presented. **Paper III** suggests a conceptual and methodical framework for a structured analysis of Twitter data in the context of ecological monitoring. The results of the analysis of observational and non-observational information on IAS, sourced from Twitter, are presented in the form of an information topology profile, which is proposed as a generic model to compare future results for other monitoring subjects drawn from Twitter or alternative social media channels. The paper discusses the possible contributions of Twitter mining to ecological monitoring in general and addresses practical hurdles that need to be overcome in developing operational systems.

Paper IV elaborates the concept of social media mining as an opportunistic citizen science model and explores potential active contributions of Twitter communities to citizen science efforts. The paper is motivated by the recognition that biodiversity observations posted on Twitter often lead to responses from other users, offering taxonomic determinations of an observed species. The paper explores if the resulting data is on par with comparable deliberate citizen science efforts, and what potential these ad-hoc communities hold in advancing from passive, ad-hoc contributions to active engagements with citizen science projects.

Summary of material and methods

The notion of "social media for ecosystem monitoring" exceeds the scope of a preliminary exploration. Available social media channels extend significantly beyond well-known examples like Facebook or Twitter, and the predominantly used social media also vary geographically. Similarly the type of social media content that could be of relevance to ecological monitoring ranges from information capturing indirect effects on ecosystems (such as technological or commercial trends) to direct biodiversity observations (**Paper I**). The research presented in this study was thus framed to a specific type of social media and a monitoring subject with significant impact on forest ecosystems: Twitter and invasive alien species.

Invasive alien species

Paper III deals with the choice of invasive alien species (IAS) as a representative example for this study, specifically with regard to forest ecosystems. Approximately half of the nearly 900 invasive species currently listed in the *Global Invasive Species Database* maintained by the *Invasive Species Specialist Group (ISSG)* at the *IUCN* (IUCN-ISSG, 2015) impact different types of forest ecosystems. IAS are known drivers and indicators for ecosystem change (Crowl et al., 2008). They are a global concern with significant ecological and economic impacts (Pejchar and Mooney, 2009; Pimentel et al., 2005) and receive heightened attention by policy makers world-wide (European Commission, 2011; U.S. Government, 2010). IAS thus offer a broad range of perspectives to explore and assess information with environmental relevance in social media.

A list of 11 sample invasive alien species with direct or indirect impacts on forests was compiled in collaboration with IAS experts on the Aliens-L mailing list[1] of the ISSG (**Paper III**). The selection was based on a set of criteria ensuring broad coverage with regard to for example geographic coverage, organism type, introduction vector, invasion extent and type of impact (see Appendix of **Paper III** for details).

Social media content directly or (potentially) indirectly referencing these 11 IAS was collected since May 2013 (for one species since May 2012). **Paper I** builds on early results for

[1] https://list.auckland.ac.nz/sympa/info/aliens-l

one of these species (Oak processionary moth), and results for three species (Oak processionary moth, Emerald ash borer, Eastern grey squirrel) are explored in detail in **Paper III**.

Twitter

Paper I provides an overview of the different types of social media, such as Wikis, Blogs or social networks, that may be explored in research or be included in future operational systems. The micro-blogging service Twitter was identified as a suitable social media source for this study due to properties elaborated in **Paper I** and **Paper III**. These include the large volume of data (300 million actives users (The Verge, 2015) posting more than 500 million messages daily (Krikorian, 2013)) as well as the usage of textual content enabling the application of standard text-analysis methods. Moreover, Twitter is a proven information hub to other online and social media sources (De Longueville et al., 2009) and the size limitation (at most 140 characters) of Twitter messages ("Tweets") implies a low contribution threshold increasing the likelihood of casual observations being reported. Finally, Twitter content is typically public and can be accessed programmatically via two public Application Programming Interfaces (APIs).

Many social media channels may offer similar volume, but content is private by default (e.g. Facebook), emphasize non-textual content (e.g. Instagram) or may not necessarily be real-time and lend itself to short casual statements (e.g. Blogs) as might be expected for informal ecological observations.

A key limitation of both the so-called Twitter Streaming API and Twitter Search API is their partial coverage of the complete Twitter data stream. According to informal estimates the Twitter Streaming API provides access to approximately 1% of all messages posted on Twitter (Huet, 2014), whereas for the Twitter Search API the sample coverage depends on a combination of frequency and popularity of search keywords (**Paper III & IV**). The obtained data can thus be assumed to provide a potentially significant underestimate of all available relevant Twitter information matching specific search keywords (**Paper III & IV**).

The Ecoveillance platform

In order to facilitate a targeted large-scale data collection from Twitter and support the analysis of the obtained Twitter messages, a web-based platform denoted 'Ecoveillance' was implemented. The Ecoveillance platform utilizes the Twitter Search API to continuously query Twitter for Tweets matching a predefined set of keywords for each selected invasive alien species.

Search terms ranged from direct references to a species ("emerald ash borer", "Agrilus planipennis") to descriptive references ("green beetle"). The data collection approach and the choice of keywords are covered in detail in **Paper III**, which also includes a complete list of all keywords. Finally, the need for a continuous data collection system highlights another limitation of the Twitter API, namely that Tweets cannot be retrieved based on keyword searches if they are older than 7-9 days.

The Twitter API returns the Tweet content together with a wealth of meta-data (author, timestamp, geo-coordinates, used devices, linked resources) in JSON format (**Paper I**), which were stored and incorporated in the analysis.

The detailed manual analysis of Tweets as described in **Paper III & IV** was supported by a categorization module in the Ecoveillance platform. Via user-configurable filters, sample Tweet sets can be obtained and analyzed using flexible category sets. Categories range from basic decisions on topical relevance of a Tweet ("on-topic", "off-topic") to observation types or covered IAS subjects. A complete list of the applied classification system is provided in **Paper III**.

The Ecoveillance platform was developed as a modular, extensible software with the intent to provide a basis for future operational use and thus represents one of the practical results of this study.

Probabilistic topic modelling

Probabilistic topic modelling refers to a suite of algorithms applied to identify distinct latent topics in large document collections (Blei, 2012; Steyvers and Griffiths, 2007). This method is applied in **Paper II** to analyze possible topical overlaps in the published literature on forest monitoring and citizen science.

Summary of results

The summary of results based on **Papers I-IV** is presented with reference to the main research questions of this study listed in the Introduction.

Do social media contain useful information with relevance to the management and monitoring of forest ecosystems?

The outlook and initial results presented in **Paper I** as well as the detailed analysis in **Paper III & IV** clearly demonstrate that social media represent a relevant data source for information on invasive alien species in forest ecosystems and for ecological monitoring in general. **Paper III & IV** show that this covers the whole range of possible contributions suggested in **Paper I**: detection of events (observations), public perceptions of natural surroundings, stakeholder information.

The latter two are covered by Twitter messages with direct references to the analyzed invasive species, which are largely of non-observational type and cover a broad range of themes. In line with the characteristics of the respective species (recognizability, type of impact, invasion history) distinctive thematic profiles emerged (**Paper III**). While those are thematic snapshots, it is likely that the temporal patterns of these themes show equally distinctive profiles that might give an indication of trends and typical developments.

The results are not sufficient to judge the assessment of perceived values of ecosystem services, but the type of themes covered (IAS impacts, critical statements on IAS management methods, location mentions) suggests that such assessments may be possible when collecting messages with additional or different keywords.

Twitter messages of observational character with direct or potential descriptive references to the analyzed IAS are also found. The share of primary observations of the targeted species is small, but holds potential with regard to the ability to contribute to the early detection of IAS infestations. Examples include Oak processionary sightings in private gardens (**Paper III**) which are not typically covered in standard monitoring programs (**Paper I**) and can thus supplement the routine monitoring of such forest threats (FVA-BW, 2012; NW-FVA, 2012).

Generally, Tweets matching descriptive terms of the targeted species (such as "green insect") proved to be a rich source of biodiversity observations (**Paper III**). However, only for specific keywords falling short of the actual species name, observations of the targeted IAS could be found. This indicates potential applications for general biodiversity monitoring using social media like Twitter. At the same time it highlights the fact that the pursued data collection strategies need to find a suitable combination of 1) search terms that capture a maximum of all relevant messages and 2) effective filtering mechanisms to identify the relevant content. This will vary depending on monitored species or ecological subject, thus requiring a good understanding of the monitored subject and clearly defined objectives for the collected information (**Paper III**). Here the vast knowledge and experience gathered in traditional forest monitoring programs could thus be employed to inform data gathering from social media and at the same time provide a natural starting point of integrating these different data sources.

A key characteristic of social media such as Twitter is that they represent a continuous data stream. Traditional forest monitoring and inventories are characterized by sampling in regular intervals of several years, which are recognized as a challenge and addressed methodically in terms of data procurement, modeling and data analysis, or by adjusting sampling intervals (Kleinn et al., 2010). The FAO for example changed their Global Forest Resource Assessment from a 10 to a 5-year interval in order to provide more adequate assessments (Kleinn et al., 2010). While social media data will not supply the fine-grained data collected in inventories, it has the potential to highlight unforeseen human impacts or natural hazards, thus contributing to a better understanding of forest developments or even providing specific monitoring triggers (**Paper I**).

How can this information be obtained, how abundant is it, how can it be characterized, and what are its quality and relevance?

Collecting social media content with potential relevance does not represent a major technical challenge. However, given the data volume and assuming a broader coverage (i.e. more IAS or additional subjects), tens of millions of Tweets would have to be processed in real-time and partly stored. While not a major technical hurdle, it requires a non-trivial amount of computational resources in order to arrive at operational systems.

Generally, the information is abundant, has a high topical relevance, and observed message numbers are likely to underestimate all available relevant content (**Paper III & IV**), in particular when including other social media channels. Furthermore, the topical relevance can largely be decided on the basis of textual content thus suggesting the feasibility of automatic filtering routines (**Paper III**).

A high-level characterization of the analyzed content is best approached on the basis of an information grid, organized along two dimensions: information relevance and information completeness (**Paper III**). Broadly, it divides the available information according to observational and non-observational content, the former representing ecological observations (here sightings of a species), the latter representing reflections on a topic (here invasive alien species). This in turn prompted the development of the information topology profile proposed in **Paper III** as a useful approach to compare different ecological subjects or the results from different social media channels.

With regard to non-observational content the assigned thematic categories may vary for each subject, but for observational data the focus on observation type, verification resources, quality of the verification and the availability of geo-information seems generally applicable.

The majority of analyzed observational content came with attached verification resources (typically images), that were of sufficient quality to verify the observations, i.e. determine the observed species (**Paper III**). Furthermore, in cases where an observation triggered conversations, taxonomic determinations were contributed by other users that proved largely correct (**Paper IV**).

A clear shortcoming with regard to completeness of observational content is the shortage of exact geo-information. Only between 1-4% of relevant Tweets come with attached geo-coordinates, other geo-location information exists, but has to be judged as less reliable (**Paper III & IV**).

The practical relevance of the attainable information thus varies. A good understanding of public perceptions is a prerequisite for successful IAS management (Bremner and Park, 2007) and non-observational content provides such information (**Paper III**). Primary species observations can of course be of immediate practical value and the rich pool of general

biodiversity observations found in the explored examples may hold applications beyond the monitoring of IAS.

How does this information relate to existing utilized ecological monitoring sources, specifically to informal monitoring efforts such as citizen science?

Social media content is unstructured, contributed opportunistically and may be ambiguous and variable in the extent of meta-data required in ecological monitoring. It thus differs significantly from the structured, comprehensive data collected in planned forest monitoring programs, but shows similarities to data collected in citizen science efforts which already supplement traditional monitoring efforts.

Paper II reveals that there are underutilized opportunities for citizen science in forest monitoring. **Paper IV** reflects on the similarities between citizen science and social media content in an ecological monitoring context and concludes that biodiversity data and ensuing activities observed on Twitter fit standard citizen typologies (Bonney et al., 2009; Newman et al., 2012; Shirk et al., 2012; Wiggins and Crowston, 2012), and with the exception of geo-location information is comparable to examples of deliberate citizen science projects for biodiversity monitoring. Moreover, even though the contributions via Twitter are ad-hoc, those contributing show apparently a more than casual interest in the reported observations (**Paper IV**).

An analysis of Twitter users contributing to biodiversity observations also revealed that these are predominantly participants with no previously documented interest or education in the biological domain (**Paper IV**). There is thus a huge potential of alert crowds that are passively and often unknowingly contributing to environmental monitoring, may show a more than casual interest in the subject and could possibly be mobilized for active deliberate monitoring, for example in citizen science projects (**Paper IV**).

In summary, the results in **Paper III & IV** suggest that both with regard to the mined data and the contributors of this data, Twitter and other social media channels could supplement traditional forest monitoring efforts. Either by providing additional observational data, assisting to direct intensified monitoring to areas with surprise observations or simply helping to raise public awareness on critical issues and gather public support.

What are the conceptual and practical limitations of this data source?

Paper III concludes that there are practical but no principal conceptual hurdles in utilizing Twitter content for ecological monitoring.

Operational systems have to address both the data volume and the need for real-time processing, especially if additional data sources should be incorporated. Thus sufficient bandwidth, processing power and storage is required. Generally, automation will be crucial to deal with this information in an efficient way. The results (**Paper III & IV**) indicate that textual content largely suffices to decide on the topical relevance of Tweets, and specifically thematic trends should be extractable using automatic text analysis approaches (including the aforementioned probabilistic topic modelling).

With regard to messages comprising species observations automatic approaches would have to encompass image recognition or even species recognition in images. However, even with recent advances in this area (see for example (Kaya et al., 2015; Kumar et al., 2012)), the quality range and variation in the observed images suggests that manual approaches will be required. Operational systems would thus have to be modelled after examples like the *Global Public Health Intelligence Network* (Mykhalovskiy and Weir, 2006) were a combination of automatic text processing and evaluations by domain experts form a successful approach in monitoring disease trends (**Paper III**). A suitable approach would be to reach out to citizen science and "crowd-process" this "crowd-sourced" data.

A practical limitation with regard to applicable monitoring subjects is indicated in **Paper III & IV**. The results here showed that the abundance and type of information is related to recognizability of a species. Common, notable and easy to recognize species produce more observations, rare or difficult to spot species very little. This eliminates certain monitoring subjects or requires a focus on indirect effects such as possibly easier to observe damages caused by a species.

Finally, a generic conceptual challenge, although not impediment, in using this information, can be seen in the representation of both data and meta-information on this data in standardized formats. The issue of data quality is frequently raised when addressing informal sources such as citizen science data (Butt et al., 2013; Crall et al., 2011; See et al., 2013) and will extend to information mined from social media when attempting to integrate it with traditional professionally collected monitoring data. Increasingly, a representation of provenance and quality as meta-data emerges (Reichman et al., 2011; Sheppard et al., 2014).

As indicated in **Paper IV** this should however be viewed as a generic challenge that extends to and would equally benefit traditional ecological data sources including forest monitoring data.

Discussion and conclusions

Starting with a very broad question concerning the usefulness of "social media in ecological monitoring" this study aimed to provide a methodical assessment of social media as a data source for environmental information focusing on the micro-blogging service Twitter and the example of invasive alien species in forest ecosystems.

The results presented in **Paper I-IV** clearly indicate that this informal data source deserves consideration in forest monitoring and beyond.

A rigorous conceptual framework and theoretical grounding seems however essential to lift this data source from the level of anecdotal application to reusable method in the monitoring toolset, and enable integration with other monitoring data sources.

Conceptually, digital information sources could be approached as a form of *digital local knowledge*. Local or indigenous knowledge always had an important place in resource management practices (Berkes et al., 2000). In forest monitoring the integration of this type of knowledge is commonly denoted participatory or community-based monitoring (Evans and Guariguata, 2008). More generally the term "citizen science" is now used when describing contributions of the general public to scientific research, with the most common activities involving monitoring in the form of data collection or evaluation (Wiggins and Crowston, 2012).

Research into the epistemology of this information is required to firmly place it into the canon of ecological information sources, for forest monitoring and ecological monitoring in general. As the cited research into typologies of citizen science indicates, there is a very practical motivation in focusing on the theoretical aspects of this information type, namely that practitioners and researchers alike benefit from generic models and guidelines that contribute to the mobilization of volunteers and ensure high data quality and successful monitoring outcomes of these volunteer activities (Shirk et al., 2012).

Practical applications in forest monitoring

A broad range of applications of social media mining in forest monitoring can be envisaged based on the results of this study.

Firstly, the approach presented in this study could be broadened to include other invasive species or forest pests in general. In parallel other social media sources could be incorporated to extend the volume of information. In specific examples the usage of this information can extend the detection of isolated infestations and can contribute to distribution maps of certain species. The large volume of messages and images posted with reference to "rhododendron" (**Paper III**) hints at this opportunity.

Since social media data uniquely provides information about the sensed environmental information and the "sensors" themselves, social media could also be applied for a stakeholder analysis (**Paper I**). This may range from obtaining public preferences for forests as recreational sites to the public discourse about organizational stakeholders in forest management scenarios. Existing survey-driven approaches for these examples (Edwards et al., 2012; Kearney et al., 1999) could be supplemented and extended via information mined from social media.

Furthermore, focusing on geo-tagged social media content with images showing trees in urban areas, tree health assessments could be pursued on a large scale with the help of volunteers. Image classification is a common and successful task in several citizen science domains (Fritz et al., 2009; Hill et al., 2012; Smith et al., 2011). Citizen science projects focusing on tree health assessments already exist (UK Forestry Commission, 2013), and could thus be expanded using geo-tagged, high-quality images from social media resulting in data assemblies that approach the level of real-time inventories.

The most promising applications would however involve the integration of multiple different data sources, for example to provide risk assessments or early warnings. One such example would be forest fire risk assessments in highly visited areas with recreational value: (Wood et al., 2013) show for example that photos taken at recreational sites such as national parks and posted on the image sharing service Flickr allow a quantification of visitation rates, while (Cortez and Morais, 2007) present a model for forest fire prediction using meteorological data. Together with traditional inventories that implicitly provide fire fuel estimates, these different data sources and models could be combined into integrated approaches that enable a real-time risk assessment of forest fires.

Finally, applications in forestry could rely on the primary nature of social media as communication channels. Having identified individuals that contribute to or are interested in specific forest-related subjects via social media channels such as Twitter, those individuals could be addressed directly via those same social media, to communicate information, warnings or request help in monitoring local threats.

Future research directions

Three major research streams emerge to continue this work on social online media as sources in support of ecological monitoring: 1) use-case and data assessments, 2) theoretical foundations and information models, 3) method development and practical applications.

Specific use-case and data assessments, resulting in information topology profiles for other IAS and social media channels, will help to obtain an even better understanding of the abundance, representativeness and quality of this data source. Throughout, other subjects and monitoring examples have been indicated and some of the collected data still awaits further exploration. These assessments should also address the taxonomic coverage of observations, the timeliness of the information and the provision of reliable geo-location information. The representativeness of social media contributors with regard to stakeholders in a particular domain and the alignment of social media communications with the "real world", i.e. whether the social media dynamics are aligned with, follow or precede "real-world events", is another research angle with important practical implications.

The theoretical foundations refer to the development of information models and typologies similar to or aligned with comparable models in citizen science. As elaborated earlier this has very practical motivations, since a better understanding of the involved communities and participation models can directly contribute to the mobilization of volunteers in monitoring efforts.

Finally, practical applications as well as testing and development of automation routines need intensified efforts in order to arrive at operational systems. This includes data standards, ontologies and formal data processing workflows to integrate this informal data source with structured data sources that are available in forestry and ecological monitoring in general.

References

Achrekar, H., Gandhe, A., Lazarus, R., Yu, S.-H., Liu, B., 2011. Predicting Flu Trends using Twitter data, in: Computer Communications Workshops (INFOCOM WKSHPS), 2011 IEEE Conference on. pp. 702–707.

Albert, M., Hansen, J., Nagel, J., Schmidt, M., Spellmann, H., 2015. Assessing risks and uncertainties in forest dynamics under different management scenarios and climate change. For. Ecosyst. 2, 14. doi:10.1186/s40663-015-0036-5

Barve, V., 2014. Discovering and developing primary biodiversity data from social networking sites: A novel approach. Ecol. Inform. 24, 194–199. doi:10.1016/j.ecoinf.2014.08.008

Berkes, F., Colding, J., Folke, C., 2000. REDISCOVERY OF TRADITIONAL ECOLOGICAL KNOWLEDGE AS ADAPTIVE MANAGEMENT. Ecol. Appl. 10, 1251–1262. doi:10.1890/1051-0761(2000)010[1251:ROTEKA]2.0.CO;2

Blei, D., 2012. Probabilistic Topic Models. Commun. ACM 55, 77–84. doi:10.1109/MSP.2010.938079

Bolte, A., Ammer, C., Löf, M., Madsen, P., Nabuurs, G.-J., Schall, P., Spathelf, P., Rock, J., 2009. Adaptive forest management in central Europe: Climate change impacts, strategies and integrative concept. Scand. J. For. Res. 24, 473–482. doi:10.1080/02827580903418224

Bolund, P., Hunhammar, S., 1999. Ecosystem services in urban areas. Ecol. Econ. 29, 293–301. doi:10.1016/S0921-8009(99)00013-0

Bonney, R., Cooper, C.B., Dickinson, J., Kelling, S., Phillips, T., Rosenberg, K. V., Shirk, J., 2009. Citizen Science: A Developing Tool for Expanding Science Knowledge and Scientific Literacy. Bioscience 59, 977–984. doi:10.1525/bio.2009.59.11.9

Bremner, A., Park, K., 2007. Public attitudes to the management of invasive non-native species in Scotland. Biol. Conserv. 139, 306–314. doi:10.1016/j.biocon.2007.07.005

Butt, N., Slade, E., Thompson, J., Malhi, Y., Riutta, T., 2013. Quantifying the sampling error in tree census measurements by volunteers and its effect on carbon stock estimates. Ecol. Appl. 23, 936–943. doi:10.1890/11-2059.1

Clark, J., 2013. Living ash project [WWW Document]. URL http://livingashproject.org.uk/pdfs/Clark. Living Ash Project.pdf

Cortez, P., Morais, A., 2007. A Data Mining Approach to Predict Forest Fires using Meteorological Data, in: Neves, J., Santos, M.F., Machado, J. (Eds.), New Trends in Artificial Intelligence : Proceedings of the 13th Portuguese Conference on Artificial Intelligence (EPIA 2007), Guimarães, Portugal, 2007, Proceedings of the 13th EPIA 2007 - Portuguese Conference on Artificial Intelligence. APPIA, pp. 512–523.

Crall, A.W., Newman, G.J., Stohlgren, T.J., Holfelder, K.A., Graham, J., Waller, D.M., 2011. Assessing citizen science data quality: an invasive species case study. Conserv. Lett. 4, 433–442. doi:10.1111/j.1755-263X.2011.00196.x

Crooks, A., Croitoru, A., Stefanidis, A., Radzikowski, J., 2013. #Earthquake: Twitter as a Distributed Sensor System. Trans. GIS 17, 124–147. doi:10.1111/j.1467-9671.2012.01359.x

Crowl, T.A., Crist, T.O., Parmenter, R.R., Belovsky, G., Lugo, A.E., 2008. The spread of invasive species and infectious disease as drivers of ecosystem change. Front. Ecol. Environ. 6, 238–246. doi:10.1890/070151

Culotta, A., 2010. Detecting influenza outbreaks by analyzing Twitter messages.

De Longueville, B., Smith, R.S., Luraschi, G., 2009. "OMG, from here, I can see the flames!": a use case mining Location Based Social Networks to acquire spatio-temporal data on forest fires, in: Proceedings of the 2009 International Workshop on Location Based Social Networks - LBSN '09. ACM Press, New York, New York, USA, pp. 73–80. doi:10.1145/1629890.1629907

Dickinson, J.L., Shirk, J., Bonter, D., Bonney, R., Crain, R.L., Martin, J., Phillips, T., Purcell, K., 2012. The current state of citizen science as a tool for ecological research and public engagement. Front. Ecol. Environ. 10, 291–297. doi:10.1890/110236

Earle, P., Guy, M., Buckmaster, R., Ostrum, C., Horvath, S., Vaughan, A., 2010. OMG Earthquake! Can Twitter Improve Earthquake Response? Seismol. Res. Lett. 81, 246–251. doi:10.1785/gssrl.81.2.246

Edwards, D.M., Jay, M., Jensen, F.S., Lucas, B., Marzano, M., Montagne, C., Peace, A., Weiss, G., 2012. Public Preferences Across Europe for Different Forest Stand Types as Sites for Recreation. Ecol. Soc. 17. doi:http://dx.doi.org/10.5751/ES-04520-170127

European Commission, 2011. European Commission - Environment - Nature & Biodiversity - Invasive Alien Species (http://ec.europa.eu/environment/nature/invasivealien) [WWW Document]. URL http://ec.europa.eu/environment/nature/invasivealien (accessed 4.26.11).

Evans, K., Guariguata, M.R., 2008. Participatory Monitoring in tropical forest management - a review of tools, concepts and lessons learned. CIFOR.

Fritz, S., McCallum, I., Schill, C., Perger, C., Grillmayer, R., Achard, F., Kraxner, F., Obersteiner, M., 2009. Geo-Wiki.Org: The Use of Crowdsourcing to Improve Global Land Cover. Remote Sens. 1, 345–354. doi:10.3390/rs1030345

FVA-BW, 2012. Aktueller Hinweis zum Eichenprozessionsspinner. Stand: 17.09.2012, Forstliche Versuchs- und Forschungsanstalt Baden-Württemberg (Germany).

Gadow, K. v., 2013. Messung und Modellforschung - Grundlagen der Forsteinrichtung (Measurement and Modeling - Basis of Forest Design). Allg. Forst und Jagdzeitung 184, 143–158.

Gadow, K. v., 2002. Adapting silvicultural management systems to urban forests. Urban For. Urban Green. 1, 107–113. doi:10.1078/1618-8667-00011

Gadow, K. v., Kurttila, M., Leskinen, P., Leskinen, L., Nuutinen, T., Pukkala, T., 2007. Designing forested landscapes to provide multiple services. CAB Rev. Perspect. Agric. Vet. Sci. Nutr. Nat. Resour. 2. doi:10.1079/PAVSNNR20072038

Galaz, V., Crona, B., Daw, T., Bodin, Ö., Nyström, M., Olsson, P., 2010. Can web crawlers revolutionize ecological monitoring? Front. Ecol. Environ. 8, 99–104. doi:10.1890/070204

Gomide, J., Veloso, A., Meira, W., Almeida, V., Benevenuto, F., Ferraz, F., Teixeira, M., 2011. Dengue surveillance based on a computational model of spatio-temporal locality of Twitter, in: Proceedings of the ACM WebSci'11, June 14-17 2011, Koblenz, Germany. pp. 1–8. doi:doi:10.1145/2527031.2527049

Hill, A., Guralnick, R., Smith, A., Sallans, A., Rosemary Gillespie, Denslow, M., Gross, J., Murrell, Z., Tim Conyers, Oboyski, P., Ball, J., Thomer, A., Prys-Jones, R., de Torre, J., Kociolek, P., Fortson, L., 2012. The notes from nature tool for unlocking biodiversity records from museum records through citizen science. Zookeys 219–33. doi:10.3897/zookeys.209.3472

Huet, R., 2014. Connecting to the pulse of the planet with Twitter APIs [WWW Document]. Twitter Off. blog. URL https://blog.twitter.com/2014/connecting-to-the-pulse-of-the-planet-with-twitter-apis

Ingwell, L.L., Preisser, E.L., 2011. Using citizen science programs to identify host resistance in pest-invaded forests. Conserv. Biol. 25, 182–188. doi:10.1111/j.1523-1739.2010.01567.x

IUCN-ISSG, 2015. Global Invasive Species Database [WWW Document]. URL http://www.issg.org (accessed 5.30.15).

Kaya, Y., Kayci, L., Uyar, M., 2015. Automatic identification of butterfly species based on local binary patterns and artificial neural network. Appl. Soft Comput. 28, 132–137. doi:10.1016/j.asoc.2014.11.046

Kearney, A.R., Bradley, G., Kaplan, R., Kaplan, S., 1999. Stakeholder perspectives on appropriate forest management in the Pacific Northwest. For. Sci. 45, 62–73.

Kleinn, C., Ståhl, G., Fehrmann, L., Kangas, a., 2010. Issues in forest inventories as an input to planning and decision processes. Allg. Forst- und Jagdzeitung 181, 205–210.

Krikorian, R., 2013. New Tweets per second record, and how! [WWW Document]. Twitter Off. blog. URL https://blog.twitter.com/2013/new-tweets-per-second-record-and-how

Kumar, N., Belhumeur, P.N., Biswas, A., Jacobs, D.W., Kress, W.J., Lopez, I.C., Soares, J.V.B., 2012. Leafsnap: a computer vision system for automatic plant species identification, in: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (Eds.), ECCV'12 Proceedings of the 12th European Conference on Computer Vision - Volume Part II, Lecture Notes in Computer Science. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 502–516. doi:10.1007/978-3-642-33709-3

Lindenmayer, D.B., Likens, G.E., 2009. Adaptive monitoring: a new paradigm for long-term research and monitoring. Trends Ecol. Evol. 24, 482–6. doi:10.1016/j.tree.2009.03.005

Malcevschi, S., Marchini, A., Savini, D., Facchinetti, T., 2012. Opportunities for Web-Based Indicators in Environmental Sciences. PLoS One 7, e42128. doi:10.1371/journal.pone.0042128

Millenium Ecosystem Assessment, 2005. Ecosystems and Human Well-Being: Synthesis (Millennium Ecosystem Assessment Series). Island Press.

Miller-Rushing, A., Primack, R., Bonney, R., 2012. The history of public participation in ecological research. Front. Ecol. Environ. 10, 285–290. doi:10.1890/110278

Mykhalovskiy, E., Weir, L., 2006. The Global Public Health Intelligence Network and early warning outbreak detection: a Canadian contribution to global public health. Can. J. public Heal. Rev. Can. santé publique 97, 42–4.

Nasi, R., Wunder, S., Campos A., J.J., 2002. Forest ecosystem services: can they pay our way out of deforestation?

Newman, G., Wiggins, A., Crall, A., Graham, E., Newman, S., Crowston, K., 2012. The future of citizen science: emerging technologies and shifting paradigms. Front. Ecol. Environ. 10, 298–304. doi:10.1890/110294

Noss, R.F., 1999. Assessing and monitoring forest biodiversity: A suggested framework and indicators. For. Ecol. Manage. 115, 135–146. doi:10.1016/S0378-1127(98)00394-6

NW-FVA, 2012. Nordwestdeutsche Forstliche Versuchsanstalt, Abteilung Waldschutz: Hinweise zur Überwachung und Bekämpfung des Eichenprozessionsspinners im Waldschutz (25.10.2012).

Pejchar, L., Mooney, H.A., 2009. Invasive species, ecosystem services and human well-being. Trends Ecol. Evol. (Personal Ed. 24, 497–504. doi:10.1016/j.tree.2009.03.016

Pimentel, D., Zuniga, R., Morrison, D., 2005. Update on the environmental and economic costs associated with alien-invasive species in the United States. Ecol. Econ. 52, 273–288. doi:10.1016/j.ecolecon.2004.10.002

Qu, Y., Huang, C., Zhang, P., Zhang, J., 2011. Microblogging after a major disaster in China: a case study of the 2010 Yushu earthquake, in: Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work - CSCW '11. ACM Press, New York, New York, USA, pp. 25–34. doi:10.1145/1958824.1958830

Reichman, O.J., Jones, M.B., Schildhauer, M.P., 2011. Challenges and Opportunities of Open Data in Ecology. Science (80-.). 331. doi:10.1126/science.1197962

Roberts, A.M.I., Tansey, C., Smithers, R.J., Phillimore, A.B., 2015. Predicting a change in the order of spring phenology in temperate forests. Glob. Chang. Biol. doi:10.1111/gcb.12896

Rockström, J., Steffen, W., Noone, K., Persson, A., Chapin, F.S., Lambin, E.F., Lenton, T.M., Scheffer, M., Folke, C., Schellnhuber, H.J., Nykvist, B., de Wit, C.A., Hughes, T., van der Leeuw, S., Rodhe, H., Sörlin, S., Snyder, P.K., Costanza, R., Svedin, U., Falkenmark, M., Karlberg, L., Corell, R.W., Fabry, V.J., Hansen, J., Walker, B., Liverman, D., Richardson, K., Crutzen, P., Foley, J.A., 2009. A safe operating space for humanity. Nature 461, 472–5. doi:10.1038/461472a

Roman, L.A., Mcpherson, E.G., Scharenbroch, B.C., Bartens, J., 2013. Identifying Common Practices and Challenges for Local Urban Tree Monitoring Programs Across the United States 39, 292–299.

Rutledge, C., Fierke, M., Careless, P., Worthley, T., 2013. First detection of Agrilus planipennis in Connecticut made by monitoring Cerceris fumipennis (Crabronidae) colonies. J. Hymenopt. Res. 32, 75–81. doi:10.3897/jhr.32.4865

Sakaki, T., Okazaki, M., Matsuo, Y., 2010. Earthquake Shakes Twitter Users : Real-time Event Detection by Social Sensors, in: WWW '10 Proceedings of the 19th International Conference on World Wide Web. ACM, pp. 851–860. doi:10.1145/1772690.1772777

Scheffer, M., Carpenter, S., Foley, J.A., Folke, C., Walker, B., 2001. Catastrophic shifts in ecosystems. Nature 413, 591–6. doi:10.1038/35098000

See, L., Comber, A., Salk, C., Fritz, S., van der Velde, M., Perger, C., Schill, C., McCallum, I., Kraxner, F., Obersteiner, M., 2013. Comparing the quality of crowdsourced data contributed by expert and non-experts. PLoS One 8, e69958. doi:10.1371/journal.pone.0069958

Sheppard, S.A., Wiggins, A., Terveen, L., 2014. Capturing quality: retaining provenance for curated volunteer monitoring data, in: Proceedings of the 17th ACM Conference on Computer

Supported Cooperative Work & Social Computing - CSCW '14. ACM Press, pp. 1234–1245. doi:10.1145/2531602.2531689

Shirk, J.L., Ballard, H.L., Wilderman, C.C., Phillips, T., Wiggins, A., Jordan, R., McCallie, E., Minarchek, M., Lewenstein, B. V., Krasny, M.E., Bonney, R., 2012. Public Participation in Scientific Research: a Framework for Deliberate Design. Ecol. Soc. 17, art29. doi:10.5751/ES-04705-170229

Silvertown, J., 2009. A new dawn for citizen science. Trends Ecol. Evol. 24, 467–471. doi:10.1016/j.tree.2009.03.017

Smith, A.M., Lynn, S., Sullivan, M., Lintott, C.J., Nugent, P.E., Botyanszki, J., Kasliwal, M., Quimby, R., Bamford, S.P., Fortson, L.F., Schawinski, K., Hook, I., Blake, S., Podsiadlowski, P., Jönsson, J., Gal-Yam, A., Arcavi, I., Howell, D.A., Bloom, J.S., Jacobsen, J., Kulkarni, S.R., Law, N.M., Ofek, E.O., Walters, R., 2011. Galaxy Zoo Supernovae. Mon. Not. R. Astron. Soc. 412, 1309–1319. doi:10.1111/j.1365-2966.2010.17994.x

Stafford, R., Hart, A.G., Collins, L., Kirkhope, C.L., Williams, R.L., Rees, S.G., Lloyd, J.R., Goodenough, A.E., 2010. Eu-social science: the role of internet social networks in the collection of bee biodiversity data. PLoS One 5, e14381. doi:10.1371/journal.pone.0014381

Steyvers, M., Griffiths, T., 2007. Probabilistic topic models, in: Handbook of Latent Semantic Analysis.

The Verge, 2015. Twitter reaches 300 million active users, but the stock crashes after earnings leak early [WWW Document]. URL http://www.theverge.com/2015/4/28/8509855/twitter-earnings-q1-2015-leak-selerity

U.S. Government, 2010. Obama Administration Releases 2011 Asian Carp Control Strategy Framework; Press Release 16 Dec 2010 (http://www.whitehouse.gov) [WWW Document]. URL http://www.whitehouse.gov/administration/eop/ceq/Press_Releases/December_16_2010 (accessed 4.26.11).

UK Forestry Commission, 2013. OPAL Tree Health Survey (Forest Research) [WWW Document]. URL http://www.forestry.gov.uk/fr/INFD-97NLES (accessed 7.2.15).

UN-CBD, 2012. Secretariat of the Convention on Biological Diversity (2012) Cities and Biodiversity Outlook. Montreal, 64 pages. Montreal.

Vieweg, S., Hughes, A.L., Starbird, K., Palen, L., 2010. Microblogging during two natural hazards events: what twitter may contribute to situational awareness, in: Proceedings of the 28th International Conference on Human Factors in Computing Systems - CHI '10. ACM Press, New York, New York, USA, p. 1079. doi:10.1145/1753326.1753486

Wang, X., Brown, D.E., Gerber, M.S., 2012. Spatio-temporal modeling of criminal incidents using geographic, demographic, and twitter-derived information, in: 2012 IEEE International Conference on Intelligence and Security Informatics (ISI). pp. 36–41. doi:10.1109/ISI.2012.6284088

Wiggins, A., Crowston, K., 2012. Goals and Tasks: Two Typologies of Citizen Science Projects, in: 2012 45th Hawaii International Conference on System Sciences. IEEE, pp. 3426–3435. doi:10.1109/HICSS.2012.295

Wintle, B.A., Runge, M.C., Bekessy, S.A., 2010. Allocating monitoring effort in the face of unknown unknowns. Ecol. Lett. 13, 1325–1337. doi:10.1111/j.1461-0248.2010.01514.x

Wood, S.A., Guerry, A.D., Silver, J.M., Lacayo, M., 2013. Using social media to quantify nature-based tourism and recreation. Sci. Rep. 3, 2976. doi:10.1038/srep02976

Zalasiewicz, J., Williams, M., Steffen, W., Crutzen, P., 2010. The new world of the Anthropocene. Environ. Sci. Technol. 44, 2228–31. doi:10.1021/es903118j

Zhao, X., Corral-Rivas, J., Zhang, C., Temesgen, H., Gadow, K. v, 2014. Forest observational studies-an essential infrastructure for sustainable use of natural resources. For. Ecosyst. 1, 8. doi:10.1186/2197-5620-1-8

Acknowledgements

First of all, I would like to thank my supervisor Prof. Dr. Dr. h.c. Klaus von Gadow for giving me the opportunity to return to the much missed academic work and helping me to complete this project. I still vividly remember one of his first lectures I attended as a graduate student, in which he challenged us with fractals and chaos theory as way to understand the complexity of forest dynamics. As much as it may have been puzzling for many, it left me with a lasting fascination, and is representative of his way to always encourage his students to think out of the box and cross the boundaries of scientific disciplines. I have found that to be a helpful and rewarding way of approaching problem-solving ever since, and it had no small part in arriving at this project either.

This project would not have been possible without the support of my local supervisor Victor Galaz at the Stockholm Resilience Centre. He invited me to stay as a guest researcher at the Stockholm Resilience Centre, helped shape the project, provided advice and encouragement all the way, connected me to a large crowd of researchers and inspired with his innovative ideas. I am extremely grateful for his support.

I am also forever indebted to Matthias Albert who supported me with this project from day one. Not only was he always available to provide support, but his comments were always spot on and exactly what was needed. He has been an invaluable advisor, but most of all is a great friend.

Furthermore, I would like to thank Prof. Dr. Jürgen Nagel for his interest in this project and his support and advice along the way. I am very grateful for the opportunity to participate remotely in his PhD seminars, present my work at the Northwest German Forest Research Institute and discuss ideas for practical applications of this research.

I am particularly grateful to the Stockholm Resilience Centre for the opportunity to stay as a guest researcher for such an extended period and make use of the excellent research infrastructure. This is an inspiring place with inspiring people. My special thanks go to Emma Sundström, Ingo Fetzer, Juan Rocha and Örjan Bodin, for many helpful conversations and seminars.

This research was partly done in parallel to an employment with the Swedish GBIF node and the Biodiversity Informatics group at the Swedish Museum of Natural History. I am very grateful to Fredrik Ronquist and Anders Telenius for giving me this opportunity to explore

the field of historic biodiversity information. Our many conversations on the challenges of mining and managing this vast amount of biodiversity data did not only complement my research but also provided me with many new perspectives and ideas waiting to be explored.

This work depended a lot on input from invasive alien species experts. My first thanks goes to Melanie Josefsson at the Swedish Environmental Protection Agency who provided me with essential insights on the importance of this subject and connected me to researchers and projects providing much needed expertise. This includes the many experts on the Aliens-L mailing list at the ISSG which helped to select suitable examples for this research. Their support is gratefully acknowledged.

I am also very grateful to SESYNC, the National Socio-Environmental Synthesis Center at the University of Maryland, which hosted the brilliant workshop on *"How can social media be used to explore coupled socio-environmental systems?"*. The organizers Nick Magliocca and Andrew Crooks managed to assemble a group that covered a fascinating range of perspectives on this new area of research and will influence my thinking on this subject for some time to come. I am grateful for the invitation and the generous funding from SESYNC.

Finally, I would like to thank my extended family for their support over the years. From my childhood days I have always been encouraged to follow my curiosity and desire to learn new things for the pure joy of discovery, and endless hours spend with my parents and grandparents exploring the local forests instilled a lasting love and concern for nature in me.

The long hours spent on this work have at times demanded a lot of patience from my family, and I thank my wife Antje and our sons Lukas and Jonas for their understanding and support. Now will be the time to concentrate completely on playing football, hiking, cycling, swimming and fishing again. Promised!

Stockholm, July 2015

Paper I

Forest Ecology and Management 316 (2014) 9–20

Contents lists available at ScienceDirect

Forest Ecology and Management

journal homepage: www.elsevier.com/locate/foreco

Forest monitoring and social media – Complementary data sources for ecosystem surveillance?

Stefan Daume [a,b,*], Matthias Albert [c], Klaus von Gadow [a,d]

[a] Faculty of Forest Sciences and Forest Ecology, Georg-August-University Göttingen, Büsgenweg 5, 37077 Göttingen, Germany
[b] Stockholm Resilience Centre, Stockholm University, Kräftriket 2B, 114 19 Stockholm, Sweden
[c] Northwest German Forest Research Institute, Grätzelstraße 2, 37079 Göttingen, Germany
[d] Dept. of Forest and Wood Science, University of Stellenbosch, South Africa

ARTICLE INFO

Article history:
Available online 29 September 2013

Keywords:
Forest monitoring
Forest observational studies
Social media mining
Forest ecosystem services
Anthropocene
Societal context

ABSTRACT

Forest monitoring captures human impacts and other biotic and abiotic influences on forests and is a prerequisite for the sustainable use and protection of forest ecosystems. Forest inventories for example are a key tool to plan sustainable harvesting, whereas Forest Observational Studies provide the empirical basis for an improved understanding and long-term evaluation of forest ecosystem dynamics. To that end detailed data is collected at stand level, often integrated in larger forest observational networks, which feeds into forest ecosystem models. Forests exist however in a constantly changing societal context and the direct or indirect impact of human activity has become a crucial driver on all types of ecosystems. The Millenium Ecosystem Assessment underlines the linkage between social and ecological systems, highlighting the centrality of ecosystem services to human well-being and the requirement for ecosystem monitoring in the "anthropocene" to provide a holistic view of ecosystems as social-ecological systems.

Framing information about the social context of a forest ecosystem, gaining the expertise and providing resources to collect this type of information is usually outside the scope of data collection for forest inventories and monitoring. Studies in other domains faced a similar challenge and turned to data mining informal online information sources to supplement traditional monitoring and data collection strategies.

This paper explores how forest monitoring approaches especially Forest Observational Studies with their long-term and large-scale focus may be complemented by social media mining. We outline (a) how social media mining methods from other domains could be applied to forest monitoring, (b) discuss identification of stakeholders, events and demands on forest ecosystems as examples of social contextual information that could be obtained via this route and (c) explain how this information could be automatically mined from social media, online news and other similar online information sources. The proposed approach is discussed on the basis of examples from a broad set of other domains.

1. Introduction

1.1. Scope and challenges of forest monitoring

Forest monitoring aims to capture the dynamics of forest ecosystems in response to a broad range of biotic and abiotic influences. Monitoring programmes vary with regard to their primary focus and objective; forest inventories for example have historically been a key instrument in planning sustainable harvests, whereas long-term and large-scale Forest Observational Studies, as defined comprehensively in the introductory contribution to this Special Issue, provide the empirical data basis for analysing ecosystem structure and dynamics. The development of a forest ecosystem is not only the result of "natural" processes, but is largely influenced by human activity. Field data – often integrated in bigger observational networks – allows modelling of forest dynamics, including tree growth, mortality, recruitment and abiotic and biotic risks, in response to site conditions, harvest events and other silvicultural operations.

Despite an assumed long-term perspective, even silviculture, as a direct and planned influence, is characterised by frequent policy changes rather than constancy (Heyder, 1984) and Gadow et al. (2007) conclude that dynamics of managed (or exploited) forest ecosystems is thus predominantly a cultural rather than ecological issue (Gadow et al., 2007).

Forests exist however in an even broader societal context and the Millenium Ecosystem Assessment (2005) underlines the linkage between social and ecological systems. More and more forest management approaches acknowledge this linkage and provide

* Corresponding author at: Faculty of Forest Sciences and Forest Ecology, Georg-August-University Göttingen, Büsgenweg 5, 37077 Göttingen, Germany.
E-mail addresses: stefan.daume@ecoveillance.org (S. Daume), matthias.albert@nw-fva.de (M. Albert), kgadow@gwdg.de (K. von Gadow).

0378-1127/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.foreco.2013.09.004

a form of adaptive forest management that aims to identify and prioritise available management options for a forested landscape resulting in a desirable mix of forest ecosystem services (Gadow et al., 2007; Heinimann, 2010; Millar et al., 2007). According to Gadow et al. (2007) the valuation and prioritisation of different forest ecosystem services is thus a typical problem of *Public Choice* that is influenced by and has to accommodate current demands of a broad range of stakeholders.

Forest planning and management thus faces both the challenge of providing an adaptive management regime that can incorporate changing demands as well as the need to obtain the necessary data to identify a desirable mix of services. An evaluation of these demands does however require to assess the perspective of social systems which is usually outside the scope of data collection for forest planning. Adaptive forest management practices will thus also have to bridge the disconnect between the available data and its suitability to assess societal demands on forest ecosystem services.

In addition to frequently changing silvicultural practices and demands, forest management has to address the uncertainty and surprises that result from the broader human impacts on the Earth's ecosystems (Albert and Schmidt, 2012; Lindner et al., 2010; Schneider and Root, 1996; Spellmann et al., 2011), which extend beyond the harvesting of resources. These impacts are so significant that the term *"Anthropocene"* was coined (Zalasiewicz et al., 2010) to highlight that humans have become the major driver of global change. The Anthropocene is characterised by large-scale, and often uncontrolled human disturbances, and the relatively scattered forest observational networks face the same known challenges as any ecological monitoring programme in picking up and anticipating these disturbances (Wintle et al., 2010). Such disturbances may occur at any time and therefore even large scale forest inventories which have time intervals of 5–10 years between successive assessments, are usually only able to capture such events long after they have occurred.

Even tightly-knit forest monitoring programmes with shorter time intervals in densely populated areas such as Europe face new challenges. Invasive alien species for example have been highlighted as a growing concern in general (EEA, 2012) and forestry has seen devastating impacts through for example *Ash dieback* (Pautasso et al., 2013) or the faster emerging *Pine wilt disease* (Vicente et al., 2012).

Native forest threats like the *Pine* or *Oak processionary* present new challenges as well; there are indications that these species, possibly due to changing climate conditions, are extending their distribution range thus potentially becoming a threat in new areas (Netherer and Schopf, 2010; Petercord et al., 2008; van Oudenhoven et al., 2008). Monitoring programmes for both native and alien invasive species exist but are resource-intensive and regionally divided responsibilities may complicate adequate responses.

In addition, the assessment of the impact is not limited to tangible goods like timber production and will vary with the predominant societal role of the forest. The *Oak processionary* for example is causing major damages to oak forests (Habermann, 2012) but also presents a significant health risk (Gottschling and Meyer, 2006). The latter is of greater importance in e.g. densely populated areas or where forests have a major recreational function. This in turn will influence the actions taken and the most suitable monitoring approach and effort.

Given the multitude of threats and challenges more resources could be committed to even closer monitoring and observational networks. However, we propose that, alternatively or in addition, existing monitoring efforts may be informed by additional, previously unused informal information sources in order to identify changing demands or unanticipated threats and thus guide data collection in existing monitoring networks. Examples of such supplementary information includes identifying stakeholders or societal demands on ecosystems at a local level which may also help to guide adaptations in the type and scope of data collected in forest inventories.

In consideration of this, we propose to mine informal online information sources as an efficient and flexible way to supplement traditional forest monitoring and data collection strategies. We will show with examples how social contextual information and indications of notable events can be mined from social media, online news and other similar online information sources.

1.2. Social online media as monitors for social systems

Social online media are a class of web-based applications and information sources, that are typically characterised by collaborative content creation driven by explicit or implicit social networks that represent virtual communities of shared interest.

The terms "Social media", "Web 2.0" or "User-generated content" are often used interchangeably to describe the characteristics of these information sources. Despite a certain fuzziness surrounding the term "Web 2.0" (DeveloperWorks, 2006) it can be best described as a set of technologies (i.e. AJAX, RSS) and tools (i.e. Blogs, Wikis, social online networks) which Kaplan and Haenlein (2010) define as "the platform for the evolution of Social Media" where users, to varying degrees, contribute to the creation of content, thus becoming "prosumers" – producers and consumers of content (Ritzer and Jurgenson, 2010; Wikipedia, 2012).

Social media classes vary with regard to the level of "personalisation" or "self-disclosure" and the "richness of the media" employed (Kaplan and Haenlein, 2010); blogs and micro-blogs for example are highly personalised as the author(s) provide personalised content and information about themselves, whereas collaborative projects like Wikipedia have a low degree of personalisation – content is not personalised and author information not as prominent. Both of these have in common that they are employing text as a main medium, in contrast, "content communities" like YouTube focus on video content as the main medium and "Virtual Worlds" like Second Life belong to social media classes that employ even richer media. Table 1 provides a definition and examples of different social media classes.

Another discriminator between the different types of social media focuses on two patterns of information flow. "Information-pull" media (Marques et al., 2012) are applications where a reader or content consumer has to actively visit the information source to obtain content; blogs belong to this class of applications. In "information-push" models (Marques et al., 2012) on the other hand the information is delivered to the content consumer; micro-blogs and social networks like Facebook are typical examples. Information-push models guarantee a broader distribution of information and facilitate recursive dissemination of this information in the networks of readers.

The potential value of informal online information sources in the ecological domain in general was recently advocated by Galaz et al. (2010). In this contribution we discuss whether informal online information sources – specifically social online media – could act as an efficient and representative source for the societal context of forest ecosystems. We will provide an introduction to social media mining, present relevant examples of social media analysis from other domains and discuss three potential areas – (1) identification of stakeholders, (2) detection of events with an impact on forests and (3) identification of demands on forest ecosystems – for which social media analysis could provide insights with regard to forest ecosystem dynamics, early warnings or management options and potentially augment and guide forest observational studies.

Table 1
Social online media classes.

Social media class	Description
Wikis	Collaborative tools facilitating the creation and organisation of predominantly text-based content; earliest examples of tools for user-generated content and precursors of other social media tools. Examples: Wikipedia online encyclopedia
Blogs	Blogs, also called weblogs or online diaries, are publishing tools with a high level of personalisation. Authors' profile information, personalised and often opiniated content is common. Commenting functions facilitate discussions around blog articles; both content and comments can be consumed as so-called RSS feeds. Examples: Wordpress, Blogspot
Micro-blogs	Micro-blogs combine features of publishing and conversational tools as well as social networks and are characterised by their real-time nature and the shortage of messages posted (maximum 140 characters). Micro-blogs are typically non-reciprocal social networks – any user interested in messages from another user can follow this user without this connection having to be reciprocated. Examples: Twitter, Sina-Weibo
Social online networks	Social online networks are reciprocal networks of inviduals with a common interest. In professional networks like LinkedIn the network itself is the main content, build up to retain connections or establish new links into neighbouring networks. With social networks like Facebook the focus shifts to information exchange, they combine the characteristics of conversational and publishing tools without the constraint on message shortage as in micro-blogs. Examples: Facebook, LinkedIn
Content communities	Content communities typically focus on richer types of media such as videos or photos. Content is usually public and networks are established only implicitly. Functions that allow to provide public comments or rate content are common features in content communities. Examples: YouTube, Flickr
Virtual worlds	Virtual social worlds are complex software applications that replicate a broad type of real world social interactions; users interact through virtual identities (avatars). Applications include virtual situation rooms facilitating information exchange in e.g. emergency situtions. Examples: Second Life, World of Warcraft
Online news	Online news can be considered as social media due to their information aggregation function of social processes, but also if they provide channels for readers to interact – this includes functions to tag, rate or comment on articles and push content into other social media applications
Online forums	Online discussion forums are an early class of social media applications. Most forums require registration and content is usually created with reference to specific subjects. Social networks are not supported explicitly, but virtual communities of shared interest may emerge around certain subjects

2. Methods – data mining online sources

Social online media have become an important information source in domains as diverse as public health (Eysenbach, 2011b), finance (De Choudhury et al., 2008) or emergency management (Vieweg et al., 2010). Depending on their predominant characteristics they can have different functions: collaborative projects like Wikipedia will aid iterative information collection, organisation and management that may even allow to extract formal representations of domain knowledge (Milne et al., 2006), whereas blogs, micro-blogs like Twitter and social online networks such as Facebook with their real-time nature, conversational characteristics and often short messages are more suitable to capture important events, emerging trends and explicit or implicit networks of stakeholders.

Social media content and social networks thus provide a direct or indirect reflection of primary information sources (real-life events), secondary sources (scientific articles, news, etc) as well as discussions and reflections on both of those. Therefore, this content is not only a source of domain information but also of the societal context and perception of the domain and issues surrounding it.

While the primary use of social media is the direct communication and dissemination of information items, the aggregation and mining of this data has become an important indirect information source in many domains. The emphasis is here on the analysis of textual, real-time content obtained from blogs, micro-blogs or online news. The best researched domain is in the area of epidemic disease monitoring which resulted in several operational systems. One of the first – employed by the World Health Organisation – is the *Global Public Health Intelligence Network* (GPHIN), which automatically monitors various online sources for signs of infectious disease epidemics (Mykhalovskiy and Weir, 2006). The collected content is automatically pre-filtered and then further evaluated by health experts to obtain early disease warnings or follow the development of epidemics.

Fig. 1 illustrates the general approach and the basic steps of a social media mining application.

The majority of social online media provide publicly accessible content, which can be collected by standard web crawlers, tools that recursively retrieve and index the textual content of web sites. In addition, most social media applications also provide public APIs (Application Programming Interfaces) which have in common that they return content based on a reliable contract and in a structured form thus easing the automatic analysis or mining of content. The details of this process are described in the following sections.

2.1. Step 1: Content retrieval, transformation and storage

All online content mining systems share the same main processing steps independent of the analysed sources (e.g. Twitter, Blogs, online news). In a first step content needs to be retrieved from the source, which is usually achieved through a programmatic routine that secures content automatically and continuously in order to incorporate new content as it becomes available.

Whereas certain types of data collectors (such as common web crawlers) retrieve "raw unstructured content" (web pages in HTML format), many sources provide APIs through which content can be obtained with at least a basic structure imposed; RSS feeds are a common and basic type of API typically available for blogs, online news and many other online sources. Sites like Twitter offer more sophisticated APIs through which content can be retrieved in the form of simple queries constraining returned information items by e.g. keywords, date, author, geo-location, etc.

In addition to the imposed structure, all APIs add semantics to the data allowing to identify authors, posted content or publishing dates. More advanced APIs add geo-location data, identify embedded links to other sites or names of places or persons. Fig. 2 illustrates how raw Twitter site content is represented in the JSON data format; the transformation being facilitated via the Twitter API.

Data collection concludes with the storage of the obtained content in a structure and format suitable for further processing.

2.2. Step 2: Content filtering and categorisation

Prior to a detailed analysis the content has to be filtered for further processing. The objective of this step is to separate relevant from irrelevant items and is best explained with the example of an online news site which publishes articles covering any number of subjects including articles in the forest domain – the latter will

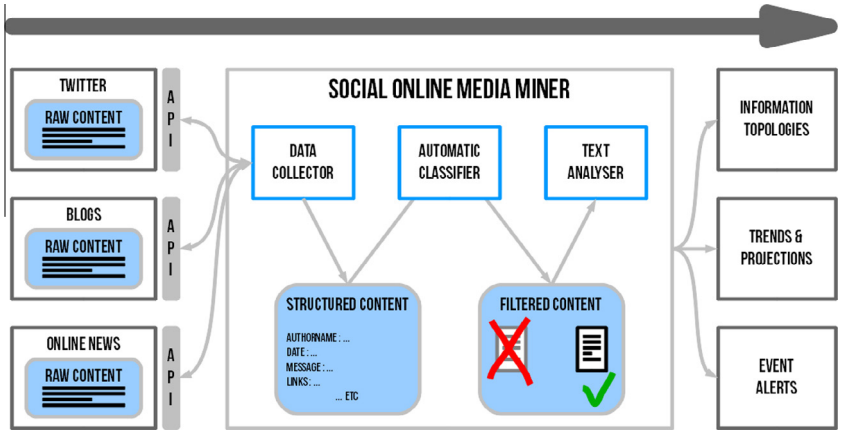

Fig. 1. General illustration of social media mining processes. Data collected via APIs from a variety of online sources is stored, filtered, categorised and analysed to extract new aggregated information.

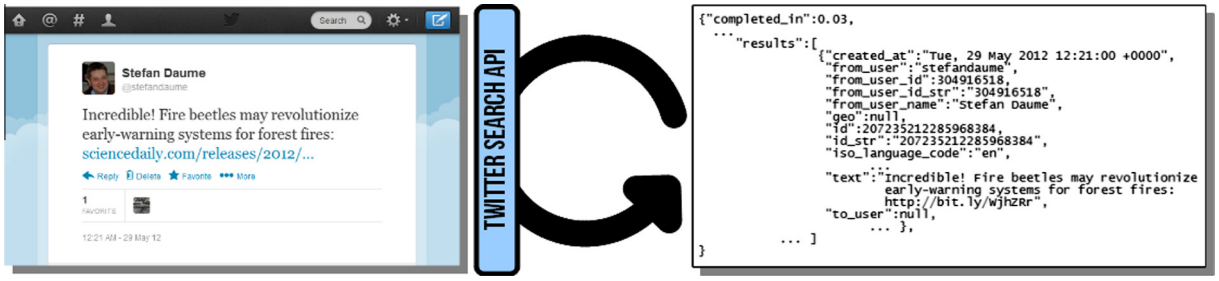

Fig. 2. Exemplary Twitter website snapshot transformed into the JSON data format retrieved via the Twitter API.

have to be identified for a system that intends to analyse content relating to forest ecosystems.

With large, continuously collected data sets automatic classification routines are employed. In order to assign a given text or document to a predefined *class* ("relevant" or "irrelevant"; "text covers forests" or "text does not cover forests") *supervised learning techniques* such as *Naïve Bayesian Classification*, *Support Vector Machines* or *k-nearest Neighbour Learning* have proven to be successful methods (Liu, 2009). For all of the above a sample of retrieved content has to be classified by a domain expert in order to create a *training set*, which will be used to train one of the above classification routines and obtain a classifier that can be applied to new content.

A classifier assigns a class to a given text based on certain features of this text (such as word frequencies); measures such as *precision*, *recall* and *F-score* indicate the accuracy of the trained classifier (Liu, 2009).

It is important to note that due to semantic ambiguities this classification step has to be applied even if the initial data set consists of pre-filtered content matching certain keywords – the word "tree" could for example be used in a genealogical ("family tree") or mathematical ("decision tree") context and is thus insufficient to indicate content about the forestry domain.

2.3. Step 3: Text and content meta-data analysis

Further analysis of the filtered/classified content depends on the purpose of the data mining exercise. An initial basic output are the frequencies over time with which documents containing certain keywords are published. They can indicate trends or may be used for projections – examples include discovery of disease epidemics (Schmidt, 2012) or predictions of stock market activity (De Choudhury et al., 2008).

In addition, a more fine-grained application of the mentioned classification routines can help to identify specific types of content from a large pool of items. This may be a precursor to further analysis, for example by distinguishing between primary (e.g. references to original observations) and secondary (e.g. comments on news articles) information, or to deliver specific types of content items to specific audiences (e.g. official warnings in an emergency situation (Palen et al., 2010)).

Next, a detailed (statistical) text analysis can help to identify important topics through most frequently used words or to find correlated issues by looking at words commonly used together. Those can again be analysed along a temporal dimension in order to explore how major topics or correlated themes change over time. Other resources may be included in the analysis; an analysis of web links embedded in a text is for example a common application (Yu et al., 2010).

Another advanced type of text analysis is *sentiment and opinion analysis* (Liu, 2010), which attempts to determine whether a given text makes a positive or negative statement about a covered topic. Sentiment analysis allows to describe how certain issues are perceived by the authors of the analysed content.

Finally, author and geo-location information can be used to name the most important information providers, stakeholders in general and in certain cases infer social networks established with regard to a specific subject. Geo-location information may enable views into local stakeholder networks or identify geographical trends. Given the rich meta-data usually available for collected content the boundaries between text mining and data mining in general are often blurred in this case.

Section 3 explores this process on the basis of an example analysis of Twitter content related to a specific forest threat: *Oak processionary*. While all social media sources listed in Table 1 qualify as potential information sources the initial exploration of social

media as a complementary source in forest monitoring as well as the example presented here focuses exclusively on messages posted on the micro-blogging service Twitter. The following properties suggest that Twitter is, at least initially, a suitable social media source to explore:

- *Low contribution threshold*. It requires less effort to write a message limited to 140 characters then for example a Blog or Wiki article and hence Twitter seems more likely to deliver new observations (i.e. sightings of a species like oak processionary). Compared to online news Twitter is also more likely to deliver primary observations.
- *Public content and API*. Messages posted on Twitter are typically public and a public web interface is available to retrieve content programmatically. Facebook may be a valuable source as well, but its content is not public.
- *Non-reciprocal network*. In contrast to Facebook where connections need to be confirmed by both users connecting with each other, a Twitter user can follow any other user without the need for reciprocal confirmation. Hence, networks grow faster and information spreads further.
- *Textual content*. For an initial exploration, textual content is more accessible and offers a broader and established set of semi-automated analytical tools, than for example video content posted on YouTube or pictures published on Flickr.
- *Information hub*. Related Twitter mining studies point out that Twitter is an information hub that links to other types of social media such as Facebook, Blogs, Wikis, etc (De Longueville et al., 2009). Thus Twitter provides access to other social media sources and allows an at least indirect assessment of the importance of these additional sources.

For the monitoring of a forest threat like *Oak processionary*, Twitter with its real-time nature seems a particularly good choice, however – as will be discussed later – depending on the explored subject other social media choices may be more relevant.

3. Example

Our research explores a diverse set of targets with relevance to forestry in order to assess the potential of social media as a supplementary information source. These include *Ash dieback* (Chalara disease) which can only be recognised by symptoms on the host species, *Emerald ash borer* which is having dramatic impacts in urban environments in the United States or the *Horse-chestnut leaf miner* which is slowly expanding its range in Europe.

Other examples are the *Grey squirrel* – invasive in several parts of Europe – where social media mining can not only deliver sightings but also interesting insights in the public perception and response to the control and eradication of a species that is perceived as fairly harmless or "cute".*Rhododendron* is yet another example monitored on Twitter which is not only tackled as an invasive species by the UK Forestry Commission but is also recognised as a vector of *Sudden oak death*. Geo-tagged pictures posted via Twitter with *Rhododendron* in full bloom are not uncommon and could thus provide additional information for two different perspectives on this species.

However, in the following we will focus on the example of *Oak processionary (Thaumetopoea processionea),* a type of moth indigenous in Central European oak forests which has several distinctive features and impacts that make it a good candidate for an exploratory social media mining exercise.

The caterpillars build large colonies and have long defensive poisonous setae (hairs) which break when touched, then releasing a toxin that can cause skin irritation, respiratory problems and anaphylactic shocks in humans. Apart from their impact on forests they represent thus also a major health issue.

In recent years more northerly sightings indicate that the species is extending its distribution range, which has been discussed as a possible sign of climate change by related studies (van Oudenhoven et al., 2008).

In Germany as in other central European countries the spatial and temporal distribution and the population density of *Oak processionary* is routinely monitored (e.g. FVA-BW, 2012; NW-FVA, 2012). The monitoring methods sometimes vary between federal states and even county districts and different institutions are responsible to collect and analyse the data. For example, the Departments of Forest Protection within the German Forest Research Institutes process the notification about the appearance of *Oak processionary* sent by local forest practitioners (Habermann, 2012; Möller, 2012). Based on these analyses pest control is coordinated to prevent damage to oak forests or warnings to the public are given to make aware of health risks. Public health departments on the other hand collect data on the distribution and frequency of allergic reactions due to *Oak processionary* (Scherbaum, 2012). Altogether the monitoring is diverse in method, intensity and information gained. Thus, it seems promising to explore how social media mining can support or complement an improved monitoring system.

The *Ecoveillance* system (http://www.ecoveillance.org), a social media mining platform under development to obtain early warnings for ecological changes (with an initial focus on alien invasive species) was used to collect Twitter messages explicitly mentioning "*oak processionary*". In addition, variations on the common German name "*Eichenprozessionsspinner*" and the scientific name "*Thaumetopoea processionea*" where used to obtain potentially relevant tweets on this subject. In order to capture tweets by users who do not recognise the species, a second set of keywords – such as "hairy caterpillars" – containing descriptions of distinctive features of *Oak processionary* caterpillars are used.

Table 2 represents a small selection from more than 2000 tweets mentioning this subject which were posted on Twitter between May 2012 and June 2013.

The majority of the above tweets refer to the subject directly. We find references to geo-locations (*"Steinfurt"*, *"London"*, *"Berkshire"*, *"Regents Park"*, *"Lainzer Wald"*), oak processionary are referenced as a public health hazard (T1, T6), a related species – "*Horse-Chestnut Leaf Miner*" – is mentioned in the context of changing climate conditions (T2), the cause of an infestation is discussed (T4) and other users are specifically addressed by name (i.e. *"@JusJane53"*).

Several tweets contain links to external resources such as news sites (T1, T7), a privately maintained site with health information (T6), an information and reporting page by the UK Forestry Commission (T3) and blogs that provide background information and bibliographies on *Oak processionary* (T5) or private reports by a birdwatcher (T8).

Three tweets include or link to pictures (T8-10) of sufficient quality to identify the species. T9 triggered responses by several other users that helped to identify the species and while this turns out not to be an *Oak processionary* sighting it is an example of a tweet with attached geo-coordinates that allows to place the identified species at the location where it was observed.

Finally, several tweets report indirectly (T6) or directly (T8, T10) sightings of *Oak processionary* – one in London's Regents Park (T8) and the other in a private garden (T10). In the latter case the location can be inferred as "Holland" from the user's profile. However, if we envisage these types of messages as part of a monitoring system, the advantages of Twitter or other social media tools is that they allow follow-up with the message's author to verify the sighting or gather additional information.

Table 2
Sample Tweets referring to Oak processionary or related subjects.

T1	"Eichenprozessionsspinner – ein Nachtfalter erreicht Steinfurt, Gefahr durch giftige Raupenhaare http://t.co/Hidf9bpU" (Steinfurt Tweet (@steinfurt_tweet), 2012)
T2	"Horse-Chestnut Leaf Miner (Cameraria ohridella) & Oak Processionary (*Thaumetopoea processionea*) Med' natives, but 'the climate is a changin'" (Russel G. Sharp (@Rusty_Sharp), 2012)
T3	"Stay alert for the nests which oak processionary moth caterpillars are now building in London and Berkshire oak trees. http://t.co/6cu6ow38" (FC Tree Pest News (@treepestnews), 2012)
T4	@Lynnibinny @JusJane53 "Infestations of oak processionary moth have been tracked back to large specimen oaks brought in from Holland" (Marco (@vBelz), 2012)
T5	New blog post up. Oak processionary moth a new pest to UK oak trees http://wp.me/p3pKvg-3f #plantscience #botany #moth (Sarah Shailes (@SarahShailes), 2013)
T6	KiGa Ausflug: viele Kinder und Betreuerinnen haben Ausschlag wg Kontakt mit Prozessionsspinnerraupe im Lainzer Wald. http://www.med4you.at/derma/allerg_intol/eichenprozessionsspinner.htm … (Bettina Schimak (@BettinaSchimak), 2012)
T7	Eichenprozessionsspinner: Wo man die Raupennester melden sollte http://bit.ly/Mvoef8 (Berlin aktuell (@Berlin_de_News), 2012)
T8	More oak processionary caterpillars, and my neck needs a massage. #wildlife#moths#LNHS http://regentsparkbirds.blogspot.co.uk/2012/06/12th-june.html?m=1 … (Regents Park Birds (@parkbirdslondon), 2012)
T9	Keeping my children entertained finding loads of hairy caterpillars. Do you know what they are? http://ow.ly/i/2cLfj #recordwildlife (Record Wildlife (@RecordWildlife), 2013)
T10	Not very happy about this in our backyard, a processionary caterpillar train on our oak! My dad has removed them http://twitpic.com/9u2q61 (Ildikó (@nlduranie), 2012)

These samples illustrate Twitter messages as a rich information source and it is important to note that additional information is available which is not visible here, for example author profiles, geo-coordinates attached to the tweets, etc.

The samples also hint at the potential of other linked social media sources, but equally illustrate the challenges associated with the analysis of this information, whether it is filtering out the small amounts of novel sightings from the majority of more general references to the *Oak processionary*, the variable quality of location information or the identification of false positives. A future system based on this prototype will probably have to incorporate additional content from e.g. Blogs and Wikis, but will also have to build on a combination of extensive manual categorisations and automatic classification techniques in order to be employed as a supplementary monitoring tool.

The Twitter messages were collected via the Twitter API and then processed following the steps outlined in the previous section. Among all the possible results the following basic outputs should illustrate the value this aggregated information promises.

Firstly, the frequency with which messages mentioning "*oak processionary*" or one of the other keywords have been posted (Fig. 3) may offer initial insights.

Expectedly, Fig. 3 shows that the bulk of the communication occurs between May and July (after the larvea hedge). Intermittent messaging bursts could relate to certain events (new oak processionary outbreaks) or indicate a generally hightened attention the topic receives including for example conversations on the significance of oak processionary as a climate change indicator.

Next, a look at the 20 most frequently used "hashtags" (words in a tweet preceded by a "#" symbol to indicate a topic) allows an assessment of the most important themes in the messages (Fig. 4). One of the keywords used to retrieve tweets – "*eichenprozessionsspinner*" – is the most frequent. However, the hashtags "*gesundheit*" (health) and "*allergie*" (allergy) indicate that the monitored subject is connected to health topics; place names ("*potsdam*", "*berlin*", "*bromley*") indicate locations potentially affected by oak processionary outbreaks.

Aggregated social media content could be correlated with spatial and temporal patterns obtained through already operating forest monitoring networks, but may also indicate correlated topics, include information not covered by forest monitoring (such as observations in private gardens), highlight new geographic areas that deserve closer monitoring and represents a cost-effective and real-time information aggregation source.

Finally, a combination of a message's author and mentions of users (for example "*@JusJane53*" in T4 (see above)) allows to infer stakeholders about the topic (those communicating on it) as well as connections between them. Fig. 5 illustrates a social network inferred from the collected tweets.

Nodes in Fig. 5 represent Twitter users. If a user mentions another user in a message this is interpreted as a link between them; multiple links between two nodes are indicated by thicker lines

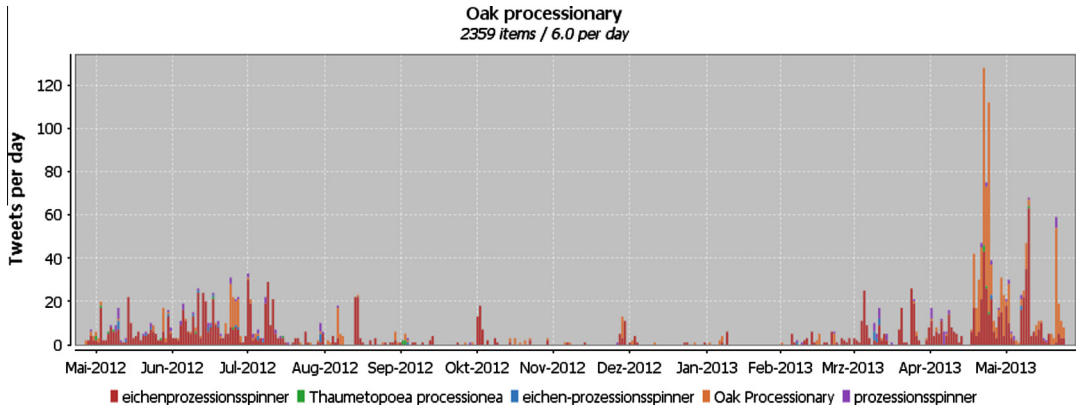

Fig. 3. Frequency of tweets posted between May 2012 and June 2013 that specifically mention "oak processionary" or similar terms. Screenshot of a chart produced with the Ecoveillance platform.

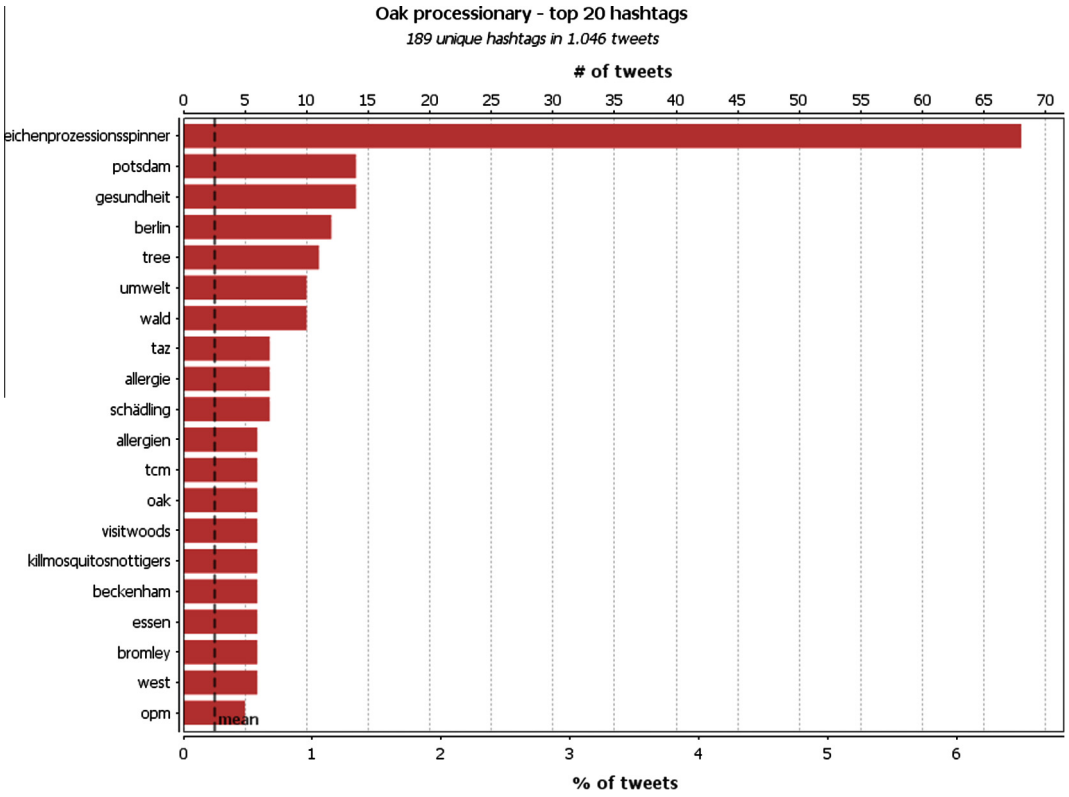

Fig. 4. Top 20 hashtags used in tweets posted between May 2012 and January 2013 that mention "oak processionary" or related keywords. Screenshot of a chart produced with the Ecoveillance platform.

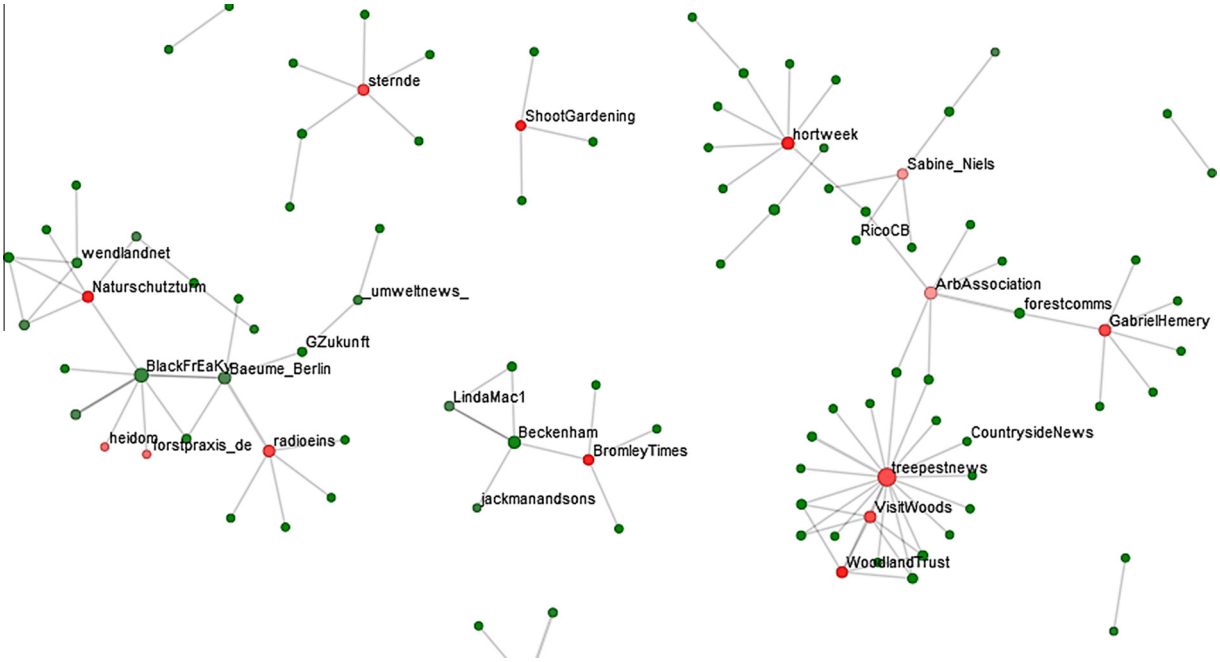

Fig. 5. Social network inferred from author information and mentions of Twitter users in posted Tweets. Screenshot of a chart produced with the Ecoveillance platform.

and more connections influence the node size. Finally, the node colours indicate whether a user is mentioned more often in messages then (s)he is posting self (red/light-grey) or whether a user is predominantly posting messages rather than being mentioned (green/dark-grey).

Geo-location information and the incorporation of the temporal dimension could be included in a more a detailed analysis. The simple network visualisation does however immediately highlight information hubs and potential information distribution pathways. In the case of *Oak processionary* as a health risk this represents one

valuable social media mining result, for example when applying these information hubs to warn about threats identified through other traditional monitoring networks.

Finally, the connectedness of such a network could indicate the level of attention a topic receives or the maturity of a conversation (how well organised it is with regard to its participants); this in turn might be correlated with how advanced a threat is or at least as how severe it is perceived and it could thus present a generic threat level indicator. Social network parameters as generic threat level indicators are at this stage hypothetical. This idea is however motivated by findings in the area of early warnings for catastrophic transitions (Scheffer et al., 2009) and suggests valuable future directions for research into our proposed approach.

4. Discussion – social online media as a supplementary data source for forest ecosystem surveillance

One of the strongest indicators that informal online information sources and the described data mining techniques could deliver useful information on forest ecosystems and their societal role and perception, is the success of citizen science projects in the environmental and ecological domain.

Projects such as the Christmas Bird Count, OPAL (Open Air Laboratories) or the Invasive Species Survey (Silvertown, 2009) rely on contributions from amateur scientist contributing observations through digital channels. The common trait of these projects is that they are examples of "participatory sensing" – they rely on an active contribution of information through specific channels requiring pre-defined data structures. Social media mining on the other hand aggregates potentially large volumes of information, contributed through generic channels, by users who will probably often be unaware that they are contributing useful information to a specific domain, such as forest ecosystem monitoring.

Suitable supplementary information on forests and the role of forests in societal contexts could of course be obtained through traditional data collection approaches such as interviews or the mining of primary data sources i.e. demographic or economic data. This is however outside the scope of forest inventories and requires substantial additional resources. Given the availability and accessibility of social online media content as well as advances in the processing of these large data volumes, social media may thus offer a suitable source and an efficient approach to obtain information about the societal context of forests.

We argue that similarly to domains such as health monitoring social online media could provide valuable information on forest ecosystems, specifically on the role and perception of forest ecosystems and could thus contribute to traditional approaches of forest monitoring.

Information obtained via informal sources could help to identify threats and long-term trends between inventory intervals, could provide background information regarding the use of forests, perceptions or expectations and may help to identify information needs. With the help of matching examples from other domains, we will discuss three potential contributions in the following sections:

- Stakeholder identification. – Who is interested in/talking about forests and forest ecosystem services?
- Event identification. – What is happening in for example a specific geo-region?
- Demands on and use of forest ecosystems. – Which ecosystem services are considered most important by different stakeholders?

4.1. Stakeholder identification

A broad set of stakeholders – aside from forest managers and researchers – are interested in or are affected by forest ecosystems or forest ecosystem services. Examples are researchers in non-forestry domains, resource managers, local or special interest groups, politicians or the general public which are not only consumers but also providers of additional information relevant to forest management.

Studies on stakeholder perspectives of forest management highlight the differences and similarities between stakeholder groups (Kearney et al., 1999), the diversity of the dimensions with which forests are assessed and the different importance placed on these dimensions by distinct stakeholder groups (Kearney and Bradley, 1998), but also discrepancies in the perception of stakeholder groups of each other (Kearney et al., 1999).

The consideration of different stakeholder perspectives plays a major role in models of adaptive governance of forest ecosystems (Elbakidze et al., 2010) and (Elbakidze et al., 2010) state that the objectives of sustainable forest management formulated through national as well as international policies demand a broader inclusion of stakeholder perspectives. This does however present specific requirements on the data collected in regular forest inventories in order to address different stakeholder perspectives. Examples include information that allows to assess biodiversity perspectives or the recreational value of forests. In order to assess the need for this kind of information and decide on the information needs and data collection strategies, the relevance of different stakeholder perspectives has to be addressed first. Monitoring of social media content related to forest ecosystems or forest ecosystem services could be a feasible and cost-effective approach in order to identify both stakeholders as well as stakeholder perspectives.

Examples from various domains show how stakeholder information can be collected via mining of online information sources in general and social media specifically. The meaning of the term *stakeholder* has to be interpreted differently depending on the domain but here generally encompasses *producers and consumers of domain-specific information*, that is anyone creating relevant online content, as well as referencing or commenting on it.

A standard output of studies in social media mining is a characterisation of the "information actors" or stakeholders in a domain's information landscape. A popular example is the identification of political opinions (Sobkowicz et al., 2012) or political developments like the recent uprisings in the middle east (Wang, 2011). While many studies focus on the development of sentiments about political issues and politicians, or the patterns of information propagation they can also help to identify information actors with regard to a particular subject and thus a stakeholder in a given domain.

Social media mining studies in the area of emergency management identify stakeholders relevant during natural disasters such as floods (Starbird et al., 2010), forest fires (De Longueville et al., 2009) or earthquakes (Qu et al., 2011). Apart from the identification of different stakeholder groups (general public, emergency services, government agencies, news providers) information mined from social media can help to identify good, reliable or relevant sources of information, prioritise emergency operations or identify communication channels (Starbird et al., 2010).

With the help of implicit or explicit geo-location information attached to social media content spatial trends can be identified (Sakaki et al., 2010) and stakeholders can be tied to a geographic context.

Applications in ecology or forest management could begin by focusing on specific issues that are relevant topics at local or national levels such as conservation, biodiversity, tourism, recreation

or water management and identify information actors around these topics. Examples include local companies or organisations, environmental interest groups, news providers or individuals in the general public. The collected information may confirm known stakeholders and concerns or bring up new connections, for example potential conflicts on resources such as water between different land use groups. With regard to participatory sensing this information will also identify potential groups of people that could act as information providers in citizen science projects thus contributing to the data collection efforts in standard forest inventories. In addition, identified stakeholders will also be targets for information distribution.

However, in general the ease and cost-effectiveness with which we can get hold of this additional information will come at the price of a lower information resolution and reliability. This applies particularly to sources that are characterised by the shortage of posted content (Twitter, Facebook). Furthermore, while it seems reasonable to frame an initial exploration of this new approach to one source, with Twitter offering several advantages outlined in Section 2, further research needs to ensure coverage of all types of social media.

Specifically, an advanced assessment of the potential contributions of social media mining to stakeholder identification in forest monitoring will require case studies that need to analyse how well the social media landscape and content reflects actors, events and demands in the "real world". Twitter for example has known demographic and geographic biases (Infographics Labs, 2012; Smith and Brenner, 2012) that will apply to other social media tools as well. These biases may not necessarily preclude the usefulness of the obtained information but have to be taken into account particularly when envisioning the potential practical use in forest monitoring or even as a qualitative data source considered by policy makers.

Hence, while social media offer a large volume of data that will likely contain relevant and novel information for the forestry domain, considerable efforts will be required to address its potential practical impact. Next steps in this research effort must thus not only address the technical challenges in mining this information, but should also include comparisons between different types of social media tools and their representativeness of stakeholders and their primary issues – it is for example imaginable that blogs may be a better source to obtain information about the demands on forest ecosystems, online news may be a better reflection of stakeholders while Twitter may be more suited to event detection.

4.2. Event detection

We outlined in the introduction that the human impact on forest ecosystems goes beyond harvest events and can happen at any time. These impacts as well as other natural hazards may thus be missed between forest inventory intervals even if the collected data would support the detection of such events.

The term "events" includes long-term developments such as a shift in water regimes (possibly induced by other land uses in a forest's proximity) or discrete catastrophic events such as forest fires, pests or natural disasters. The likelihood of such events and thus the ability to assess the resilience of a forest ecosystem will be crucial information for forest planning as well, since the stability of a managed forest's state is a prerequisite for the realisation of any sustainable management scenario (Gadow and Pukkala, 2008).

Information about these type of events or indicators of pending events exist and are covered by appropriate monitoring efforts, but the coverage is limited and cannot possibly be extended to the same level as forest inventories and clearly not to cover all possible unexpected impacts on forest ecosystems that are relevant for forest modelling and planning. As an example we may consider infestations of *Oak processionary* or invasive alien species that start in private gardens; here early observations may initially be limited to individuals who share this observation informally but cannot assess the significance in the context of a forest monitoring programme.

Other domains facing similar challenges successfully turned to social online media monitoring to augment and direct existing monitoring structures. The prime example is the monitoring for epidemic diseases. The *Global Public Health Intelligence Network* (GPHIN) (Mykhalovskiy and Weir, 2006) for example utilises a broad range of informal online information sources to detect signs for the outbreak and spread of infectious diseases. The WHO uses the GPHIN system together with traditional health reports to get early warnings for disease outbreaks – in the case of influenza outbreaks 1–2 weeks earlier than with traditional health monitoring approaches. Similar systems utilising only Twitter as a data source were successfully applied to the detection of influenza (Achrekar et al., 2011) or Dengue fever (Gomide et al., 2011) epidemics.

The potential of this approach for the detection of ecological changes is currently explored in the *Ecoveillance* project (http://www.ecoveillance.org) focusing on the example of alien invasive species. If information about species invasions can be extracted from social media this may be applied to other relevant events as well.

The value of this information lies – similar to examples from the health domain – in its capability to augment existing inventory efforts. It could highlight hotspots that merit the gathering of data or indicate the need to gather additional data to assess specific threats. The assessment of the abundance and reliability of this data is part of the aforementioned project and will be crucial to assess the influence this information will have on forest data collection. However, it can in any case be expected to provide contextual information that was otherwise unavailable.

The main contribution of social online media mining with regard to event detection thus lies in utilising previously unavailable or costly to obtain information that will at least help to reflect on the relevance of collected forest inventory data or suggest additional data needs.

However, while the relevance of social media content relating to discrete events such as forest threats may seem easier to verify, it remains to be seen how abundant this information is and how it can be applied. New, primary observations reported for example via Twitter can be found but long-term studies will have to explore if the collected data can trigger early warnings, and comparisons with other (traditional) monitoring approaches need to clarify whether a sample taken via social media are representative enough to allow extrapolations – an increased number of Twitter sightings of e.g. *Oak processionaries* may reflect the spread of the species, but could equally be triggered by increased awareness, increased media coverage or simply a growing number of social media users.

Furthermore, the geographic biases mentioned in the previous section have to be taken into account to calibrate the sampled data; urban users may for example have broader access to internet and mobile computing technologies and this could thus lead to a biased representation of sightings. In this context an assessment of the reliability of geo-location information attached to social media content is crucial. Geo-information for messages posted on Twitter is for example often limited to the static location a user provided in her profile, which must not necessarily be the location of an observation. But even geo-coordinates attached automatically to a Tweet sent from a mobile device may not be guaranteed to be the place of the actual observation. In general, the demographic, geographic and behavioural properties of the "sensors" providing the information has to be incorporated into the analysis and thus has to be collected in addition to the basic social media content.

Finally, it has to be acknowledged that social media mining as a supplementary monitoring approach will not work for all types of events. Events or observations that have special, unusual and recognisable features are more likely to be reported – large colonies of long-haired caterpillars will attract more attention than small black beetles and thus be more likely to feature in social media content.

4.3. Understanding demands on and use of forest ecosystems

A motivation for the establishment and long-term upkeep of forest observational studies is the potential for a comprehensive analysis of ecosystem structure and dynamics, including the effects of disturbances, such as harvest events, on the provision of all types of ecosystem services (Kareiva et al., 2011). However, forest observational studies cannot assess social demands reflected by an increasing number of studies focusing on the identification and valuation of non-timber forest ecosystem services such as recreational value and scenic beauty (Edwards et al., 2012; Pukkala et al., 1988), berry yields (Ihalainen et al., 2002), pollination services for coffee plantations (Ricketts et al., 2004) or biodiversity and carbon sequestration (Nelson et al., 2009; Pagiola et al., 2002).

Other ecological services assigned to forest ecosystems include water and soil protection, climate regulation, seed dispersal, natural pest control, cultural and recreational services or tourism (Nasi et al., 2002). Especially in urban settings alternative forest ecosystem services such as air filtration and micro-climate regulation play a more important role than classical forest goods (Bolund and Hunhammar, 1999). Given the current rate of urbanisation and the demands on urban systems (UN-CBD, 2012) these alternative forest ecosystem services are likely to become even more central to forest management. Standard forest inventories could thus benefit from an understanding of the demands on forest ecosystems both at a local and national level. Social online media as one easily accessible expression of the societal reflection on social-ecological systems could again provide an efficient supplementary source of information in this context.

Relevant social media mining examples can be found in the general application of recommender systems, specifically with regard to the identification of emerging news topics, political trends or long-term predictions of general technological trends. Typical targets for recommendations harvested from social online media include books, movies or TV programmes (Park et al., 2012). They rely both on the pervasiveness of information in the online domain as well as the speed of information propagation, where the propagation speed, frequency and information reach together can be interpreted as a social filter that allows an interpretation of the significance of certain topics within the social group commenting on it.

This has been successfully applied to the identification of emerging news topics or trends (Lee et al., 2010; Phelan et al., 2009; Phuvipadawat and Murata, 2010), political trends emerging in the blogosphere (Demartini et al., 2011) or social media in general (Sobkowicz et al., 2012), and mentions of scientific articles may even be correlated with impact factors of scientific journals (Eysenbach, 2011a).

An example in the area of ecology and environmental sciences was proposed by (Malcevschi et al., 2012) who analyse the frequency of specific terms in search results in order to identify a "set of web-based indicators for quantifying and ranking the relevance of terms related to key-issues in Ecology and Sustainability Science" (Malcevschi et al., 2012). Similar approaches incorporating social online media sources and focusing on the forestry domain may deliver relevant indicators to identify key issues and demands on forest ecosystem services.

The value of this information with regard to forest monitoring is in the potential alignment of data collection efforts and an assessment of the significance of the collected data. Any monitoring effort – including forest inventories as well as forest observational studies – must be driven by clear questions in order to be successful (Lindenmayer and Likens, 2010). If however forest observations are driven by the evaluation of forest ecosystem services that have decreased importance, whereas others cannot be assessed with the data collected, those data collection efforts may no longer meet the requirements of successful monitoring programmes (Lindenmayer and Likens, 2010).

However, it is important to point out that this additional information will probably primarily function as background or supplementary information. Monitoring of social online media is unlikely to deliver the specific data required for an assessment of other ecosystem services, it may however provide a pointer to the relevant services and as such help to formulate data needs, indicate required additional monitoring efforts or help to adapt existing inventory efforts. Examples include the collection of deadwood inventories where biodiversity has been identified as a major ecosystem service or the selection of alternative plot locations in areas where the micro-climate regulation service of forests in urban areas is perceived as the major ecosystem service.

The impact of potential geographic and demographic biases of social online media discussed in the previous sections applies equally when attempting to mine information about social demands on forest ecosystems. In general, the aggregation of large volumes of content is accompanied by the risk of losing important information; minority stakeholders or issues may for example be drowned by actors that have a stronger affinity to social media, and the ease of information propagation (i.e. a "Like" on Facebook or a "retweet" on Twitter) may not be a true reflection of the importance of certain subjects. Again, this does not preclude the usefulness or applicability of the collected and aggregated information, but it underlines that the properties and interaction patterns of the "social sensors" publishing the content must be part of the meta-data that feeds into the analysis of this data.

Finally, especially when dealing with insights into demands on forest ecosystems, the aggregation of opinions and sentiments requires special techniques to deal for example with ironic comments. Generic sentiment mining techniques are available but will often not be adequate for specific domains (Liu, 2010). At a minimum domain- or subject-specific samples are required to train existing algorithms. Further customizations may however be necessary and future case studies will have to explore whether a specific sentiment word vocabulary for the forest domain will suffice or if this will have to be adapted for each monitored subject.

5. Conclusions

The identification of stakeholders, events or perceptions and demands on forest ecosystems is only a small selection of areas that can be cited to characterise the societal context in which forest ecosystems exist – they may be in a continuous thematic, spatial and temporal flux and thus build a strong contrast to forest inventories assembling data that can be integrated over long timescales and potentially broad data collection intervals. These characteristics are actually a prerequisite in order to capture ecosystem dynamics. However, given the scale of the human impact on ecosystems in general, we argued that the societal context of forest ecosystems should find consideration in forest ecosystem surveillance as it offers additional support or guidance for the implementation of adaptive forest management regimes.

Numerous examples for participatory sensing ("citizen science") models with relevance to forest management exist; social

online media mining as an opportunistic sensing model has proven successful in many other domains but is still in its infancy for the ecological domain in general.

The *Ecoveillance* project is an early example which provided a simple data snapshot illustrating social media mining with a focus on an ecological subject.

Apart from technical challenges, this and other potential projects in the same domain will have to consider several issues which influence the potential application and interpretation of results, including the applications proposed in this article.

For a start, access to internet technologies and a broad use of social online media are a pre-requisite for obtaining data for the scenario we proposed. Despite a broad and accelerating uptake of ubiquitous computing technologies in all parts of the world, this limits the data coverage to regions with a broad internet and mobile device coverage. This does not preclude useful applications, but may lead to geographic biases that result in an overrepresentation of certain regions, limit the type of data available to non-local reflections or exclude coverage altogether – possibly for especially vulnerable regions in particular.

A second caveat that must be considered is that social media usage has a demographic bias which must be accounted for when interpreting collected data – especially in order to identify stakeholders or perceptions and demands on forest ecosystems. In general, this highlights that a focus on the social media content itself is not sufficient. The analysis must incorporate meta-data on the properties and usage patterns of the "social sensors" providing the content.

Moreover, given the general nature of social media interactions it can be assumed the majority of information will be secondary information (e.g. newspaper articles, etc) or reflections on secondary information rather than first-hand descriptions of ecosystem observations. For some types of analysis this will be desirable or sufficient (e.g. to analyse perceptions of forest ecosystems), for others (event detection) it may limit the applicability or usefulness of the results.

Finally, advanced interpretations of the results suggested in Section 3 (i.e. the interpretation of social connectedness as a generic measure for the advancement of a threat) are at this stage hypothetical and need a larger number of more detailed examples, but they hint at the potential value of a detailed exploration of social media mining in the context of forest observational studies or forest monitoring in general.

Further research in this area should thus first focus on charting the available information landscape and identify the type of media covered (online news, blogs, etc), the type of information (primary, secondary, etc), the type of authors, the general volume of information, the proportion of relevant information, etc. Furthermore, a sufficient level of automation, especially with regard to identifying relevant content, needs to be achieved in order to manage the expected large volume of data, before specific case studies prepare a broader application of this approach. An extensive data collection effort is underway with the *Ecoveillance* system, which will hopefully help to answer some of the above questions and guide future efforts in this interdisciplinary research subject.

Acknowledgements

At the time of writing the first author is staying as a guest researcher at the Stockholm Resilience Centre (SRC), which has generously supported the author with access to an excellent research infrastructure. Many conversations with colleagues at the SRC have contributed to this research. This support is gratefully acknowledged. Finally, the authors would like to thank the anonymous reviewers for their valuable feedback which helped to enhance this paper.

References

Achrekar, H., Gandhe, A., Lazarus, R., Yu, S.-H., Liu, B., 2011. Predicting Flu Trends using Twitter data. In: 'Computer Communications Workshops (INFOCOM WKSHPS), 2011 IEEE Conference on', pp. 702–707.

Albert, M., Schmidt, M., 2012. Standort-Leistungs-Modelle für die Entwicklung von waldbaulichen Anpassungsstrategien unter Klimawandel. Archiv f. Forstwesen u. Landschafts.oekol. 83(2), 57–71.

Berlin aktuell (@Berlin_de_News). 2012. "Eichenprozessionsspinner: Wo man die Raupennester melden sollte http://bit.ly/Mvoef8", 21 June 2012, 3:27 pm. Tweet. <https://twitter.com/berlin_de_news/status/215798026906771456>.

Bettina Schimak (@BettinaSchimak). 2012. "KiGa Ausflug: viele Kinder und Betreuerinnen haben Ausschlag wg Kontakt mit Prozessionsspinnerraupe im Lainzer Wald. http://www.med4you.at/derma/allerg_intol/eichenprozessionsspinner.htm", 21 June 2012, 1:48 pm. Tweet. <https://twitter.com/bettinaschimak/status/215773242311442432>.

Bolund, P., Hunhammar, S., 1999. Ecosystem services in urban areas. Ecol. Econ. 29 (2), 293–301.

De Choudhury, M., Sundaram, H., John, A., Seligmann, D.D., 2008. Can blog communication dynamics be correlated with stock market activity? In: Proceedings of the Nineteenth ACM Conference on Hypertext and Hypermedia – HT '08. ACM Press, New York, New York, USA, pp. 55–60.

De Longueville, B., Smith, R.S., Luraschi, G., 2009. "OMG, from here, I can see the flames!": a use case mining Location Based Social Networks to acquire spatio-temporal data on forest fires. In: Proceedings of the 2009 International Workshop on Location Based Social Networks – LBSN '09. ACM Press, New York, New York, USA, pp. 73–80.

Demartini, G., Siersdorfer, S., Chelaru, S., Nejdl, W., 2011. Analyzing Political Trends in the Blogosphere. In: 'Fifth International AAAI Conference on Weblogs and Social Media – ICWSM 2011', pp. 466–469.

DeveloperWorks. 2006. developerWorks Interviews: Tim Berners-Lee. (accessed 11.07.13). <http://www.ibm.com/developerworks/podcast/dwi/cm-int082206txt.html>

Edwards, D.M., Jay, M., Jensen, F.S., Lucas, B., Marzano, M., Montagne, C., Peace, A., Weiss, G., 2012. Public Preferences Across Europe for Different Forest Stand Types as Sites for Recreation. Ecol. Soc. 17(1).

EEA. 2012. The Impacts of Invasive Alien Species in Europe. EEA Technical Report, No. 16/2012, Technical Report 16, European Environment Agency.

Elbakidze, M., Angelstam, P., Sandström, C., Axelsson, R., 2010. Multi-stakeholder collaboration in Russian and Swedish model forest initiatives: adaptive governance toward sustainable forest management. Ecol. Soc. 15(2).

Eysenbach, G., 2011a. Can tweets predict citations? Metrics of social impact based on Twitter and correlation with traditional metrics of scientific impact. J. Med. Internet Res. 13 (4), e123.

Eysenbach, G., 2011b. Infodemiology and infoveillance tracking online health information and cyberbehavior for public health. Am. J. Prev. Med. 40 (5 Suppl. 2), S154–8.

FC Tree Pest News (@treepestnews). 2012. "Stay alert for the nests which oak processionary moth caterpillars are now building in London and Berkshire oak trees. http://bit.ly/Klrj2a". 12 June 2012, 2:43 pm. Tweet. <https://twitter.com/treepestnews/status/212525448603766784>.

FVA-BW. 2012. Aktueller Hinweis zum Eichenprozessionsspinner. Stand: 17.09.2012, Forstliche Versuchs- und Forschungsanstalt Baden-Württemberg (Germany). (accessed 14.02.13). <http://www.fva-bw.de/publikationen/sonstiges/120917eps_aktuell_07.pdf>

Gadow, K.v., Pukkala, T., (Eds.), 2008. Designing Green Landscapes (Managing Forest Ecosystems). Springer.

Gadow, K.v., Kurttila, M., Leskinen, P., Leskinen, L., Nuutinen, T., Pukkala, T., 2007. Designing forested landscapes to provide multiple services. CAB Reviews: Perspectives in Agriculture, Veterinary Science, Nutrition and Natural, Resources 2(038).

Galaz, V., Crona, B., Daw, T., Bodin, O., Nyström, M., Olsson, P., 2010. Can web crawlers revolutionize ecological monitoring? Front. Ecol. Environ. 8 (2), 99–104.

Gomide, J., Veloso, A., Meira, W., Almeida, V., Benevenuto, F., Ferraz, F., Teixeira, M., 2011. Dengue surveillance based on a computational model of spatio-temporal locality of Twitter. In: 'Proceedings of the ACM WebSci'11, June 14–17, 2011, Koblenz, Germany. pp. 1–8.

Gottschling, S., Meyer, S., 2006. An epidemic airborne disease caused by the oak processionary caterpillar. Pediatr. Dermatol. 23 (1), 64–66.

Habermann, M., 2012. Abschätzung von Schad- und Bekämpfungsschwellen beim Eichenprozessionsspinner. AFZ – Der Wald 22, 30–31.

Heinimann, H.R., 2010. A concept in adaptive ecosystem management – an engineering perspective. For. Ecol. Manage. 259 (4), 848–856.

Heyder, J., 1984. Waldbau im Wandel. PhD thesis, Georg-August-Universität Göttingen.

Ihalainen, M., Alho, J., Kolehmainen, O., Pukkala, T., 2002. Expert models for bilberry and cowberry yields in Finnish forests. For. Ecol. Manage. 157 (1–3), 15–22.

Ildikó (@nlduranie). 2012. "Not very happy about this in our backyard, a processionary caterpillar train on our oak! My dad has removed them http://twitpic.com/9u2q61", 8 June, 2012, 1:41 pm. Tweet. <https://twitter.com/nlduranie/status/211060466666061824>.

Kaplan, A.M., Haenlein, M., 2010. Users of the world, unite! The challenges and opportunities of Social Media. Bus. Horiz. 53 (1), 59–68.

Kareiva, P., Tallis, H., Ricketts, T.H., Daily, G.C., Polasky, S., 2011. Natural Capital: Theory and Practice of Mapping Ecosystem Services (Oxford Biology). Oxford University Press, USA.

Kearney, A.R., Bradley, G., 1998. Human dimensions of forest management: an empirical study of stakeholder perspectives. Urban Ecosyst. 2 (1), 5–16.

Kearney, A.R., Bradley, G., Kaplan, R., Kaplan, S., 1999. Stakeholder perspectives on appropriate forest management in the Pacific Northwest. Forest Sci. 45 (1), 62–73.

Infographics Labs. 2012. Twitter 2012. (accessed 11.07.13). <http://infographiclabs.com/news/twitter-2012/>.

Lee, Y., Jung, H.-Y., Song, W., Lee, J.-H., 2010. Mining the blogosphere for top news stories identification. In: 'Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval – SIGIR '10', SIGIR '10. ACM Press, New York, New York, USA, pp. 395–402.

Lindenmayer, D.B., Likens, G.E., 2010. The science and application of ecological monitoring. Biol. Conserv. 143 (6), 1317–1328.

Lindner, M., Maroschek, M., Netherer, S., Kremer, A., Barbati, A., Garcia-Gonzalo, J., Seidl, R., Delzon, S., Corona, P., Kolström, M., Lexer, M.J., Marchetti, M., 2010. Climate change impacts, adaptive capacity, and vulnerability of European forest ecosystems. For. Ecol. Manage. 259 (4), 698–709.

Liu, B., 2009. Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data (Data-Centric Systems and Applications). Springer.

Liu, B., 2010. Sentiment Analysis and Subjectivity. In: Indurkhya, N., Damerau, F.J. (Eds.), Handbook of Natural Language Processing, second ed., Chapman & Hall/CRC Machine Learning & Pattern Recognition, p. 704.

Malcevschi, S., Marchini, A., Savini, D., Facchinetti, T., 2012. Opportunities for web-based indicators in environmental sciences. PLoS ONE 7 (8), e42128.

Marco (@vBelz). 2012. "@Lynnibinny @JusJane53 "Infestations of oak processionary moth have been tracked back to large specimen oaks brought in from Holland". 2 November 2012, 7:58 pm. Tweet. <https://twitter.com/vbelz/status/264441447716114432>.

Marques, A.M., Krejci, R., Siqueira, S.W., Pimentel, M., Braz, M.H.L., 2012. Structuring the discourse on social networks for learning: Case studies on blogs and microblogs. Comput. Human. Behav. 29(2).

Millar, C.I., Stephenson, N.L., Stephens, S.L., 2007. Climate change and forests of the future: managing in the face of uncertainty. Ecol. Appl. 17 (8), 2145–2151.

Millenium Ecosystem Assessment. 2005. Ecosystems and Human Well-Being: Synthesis (Millennium Ecosystem Assessment Series), Technical Report.

Milne, D., Medelyan, O., Witten, I., 2006. Mining Domain-Specific Thesauri from Wikipedia: A Case Study. In: 2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2006 Main Conference Proceedings). IEEE, pp. 442–448.

Möller, K., 2012. Zum Schadpotenzial des Eichenprozessionsspinners in den Wäldern Brandenburgs. (Presentation), In: 'Fachgespräch Prozessionsspinner 2012 - Fakten, Folgen, Strategien" – 06. bis 07. März 2012 am Julius Kühn-Institut Berlin-Dahlem', Julius Kühn-Institut, Berlin-Dahlem.

Mykhalovskiy, E., Weir, L., 2006. The global public health intelligence network and early warning outbreak detection: a Canadian contribution to global public health. C. J. Public. Health 97 (1), 42–44.

Nasi, R., Wunder, S., Campos A.J., 2002. Forest ecosystem services: can they pay our way out of deforestation?. (accessed 11.07.13). <http://cgspace.cgiar.org/handle/10568/18673>.

Nelson, E., Mendoza, G., Regetz, J., Polasky, S., Tallis, H., Cameron, D., Chan, K.M., Daily, G.C., Goldstein, J., Kareiva, P.M., Lonsdorf, E., Naidoo, R., Ricketts, T.H., Shaw, M., 2009. Modeling multiple ecosystem services, biodiversity conservation, commodity production, and tradeoffs at landscape scales. Front. Ecol. Environ. 7 (1), 4–11.

Netherer, S., Schopf, A., 2010. Potential effects of climate change on insect herbivores in European forests – General aspects and the pine processionary moth as specific example. For. Ecol. Manage. 259 (4), 831–838.

NW-FVA. 2012. Nordwestdeutsche Forstliche Versuchsanstalt, Abteilung Waldschutz: Hinweise zur Überwachung und Bekämpfung des Eichenprozessionsspinners im Waldschutz (25.10.2012). (accessed 11.07.13). <http://www.nw-fva.de/fileadmin/user_upload/Sachgebiet/Schmetterlinge_Saeugetiere/EPS/AAnw_Ueberwachung_des_Eichenprozessionsspinners_25-10-2012.pdf>.

Pagiola, S., Bishop, J., Landel-Mills, N., 2002. Selling Forest Environmental Services: Market-Based Mechanisms for Conservation and Development. Routledge.

Palen, L., Anderson, K.M., Mark, G., Martin, J., Sicker, D., Palmer, M., Grunwald, D., 2010. A vision for technology-mediated support for public participation & assistance in mass emergencies & disasters. In: 'ACM-BCS '10 Proceedings of the 2010 ACM-BCS Visions of Computer Science Conference', p. 8.

Park, D.H., Kim, H.K., Choi, I.Y., Kim, J.K., 2012. A literature review and classification of recommender systems research. Expert Syst. Appl. 39 (11), 10059–10072.

Pautasso, M., Aas, G., Queloz, V., Holdenrieder, O., 2013. European ash (Fraxinus excelsior) dieback – A conservation biology challenge. Biol. Conserv. 158, 37–49.

Petercord, R., Veit, H., Delb, H., Schröter, H., 2008. Forstinsekten im Klimawandel – alte Bekannte mit neuem Potenzial?. FVA-einblick+Wald und Klima 12(1), 36–39.

Phelan, O., McCarthy, K., Smyth, B., 2009. Using twitter to recommend real-time topical news. In: Proceedings of the Third ACM Conference on Recommender Systems – RecSys '09. ACM Press, New York, New York, USA, pp. 385–388.

Phuvipadawat, S., Murata, T., 2010. Breaking news detection and tracking in Twitter. In: '2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology', IEEE, pp. 120–123.

Pukkala, T., Kellomaki, S., Mustonen, E., 1988. Prediction of the amenity of a tree stand. Scand. J. For. Res. 3 (1), 533–544.

Qu, Y., Huang, C., Zhang, P., Zhang, J., 2011. Microblogging after a major disaster in China: a case study of the 2010 Yushu earthquake. In: Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work – CSCW '11. ACM Press, New York, New York, USA, pp. 25–34.

Record Wildlife (@RecordWildlife). 2013. "Keeping my children entertained finding loads of hairy caterpillars. Do you know what they are? http://ow.ly/i/2cLfj #recordwildlife", 26 May. 2013, 1:23 pm. Tweet. <https://twitter.com/recordwildlife/status/338616284528930816>.

Regents Park Birds (@parkbirdslondon). 2012. "More oak processionary caterpillars, and my neck needs a massage. #wildlife#moths#LNHS http://regentsparkbirds.blogspot.co.uk/2012/06/12th-june.html?m=1 ", 12 June 2012, 9:11 pm. Tweet. <https://twitter.com/parkbirdslondon/status/212623135990743040>.

Ricketts, T.H., Daily, G.C., Ehrlich, P.R., Michener, C.D., 2004. Economic value of tropical forest to coffee production. Proc. Natl. Acad. Sci. USA 101 (34), 12579–12582.

Ritzer, G., Jurgenson, N., 2010. Production, consumption, prosumption: the nature of capitalism in the age of the digital 'prosumer'. J. Consum. Culture 10 (1), 13–36.

Russel G. Sharp (@Rusty_Sharp). 2012. "Horse-Chestnut Leaf Miner (Cameraria ohridella) & Oak Processionary (*Thaumetopoea processionea*) Med' natives, but 'the climate is a changin'", 25 June 2012, 6:24 pm. Tweet. <https://twitter.com/Rusty_Sharp/status/217292195873562624>.

Sakaki, T., Okazaki, M., Matsuo, Y., 2010. Earthquake shakes Twitter users: real-time event detection by social sensors. In: 'WWW '10 Proceedings of the 19th International Conference on World Wide Web'. ACM, pp. 851–860.

Sarah Shailes (@SarahShailes). 2013. "New blog post up. Oak processionary moth a new pest to UK oak trees http://wp.me/p3pKvg-3f #plantscience #botany #moth". 9 June. 2013, 1:31 pm. Tweet. <https://twitter.com/SarahShailes/status/343691797156986880>.

Scheffer, M., Bascompte, J., Brock, W.A., Brovkin, V., Carpenter, S.R., Dakos, V., Held, H., van Nes, E.H., Rietkerk, M., Sugihara, G., 2009. Early-warning signals for critical transitions. Nature 461 (7260), 53–59.

Scherbaum, M., 2012. Umweltmedizinische Bedeutung des Eichenprozessionsspinners. Retrospektive Analyse von EPS-Erkrankungsfällen in den Jahren 2004 und 2005 im Kreis Kleve. (Presentation). (accessed 11.07.13). <http://www.bfr.bund.de/cm/343/umweltmedizinische-bedeutung-des-eichenprozessionsspinners.pdf>.

Schmidt, C.W., 2012. Trending now: using social media to predict and track disease outbreaks. Environ. Health. Persp. 120 (1), A30–3.

Schneider, S.H., Root, T.L., 1996. Ecological implications of climate change will include surprises. Biodivers. Conserv. 5 (9), 1109–1119.

Silvertown, J., 2009. A new dawn for citizen science. Trends Ecol. Evol. 24 (9), 467–471.

Smith, A., Brenner, J., 2012. Twitter Use 2012. Technical report, Pew Research Center's Internet & American Life Project. (accessed 11.07.13). <http://pewinternet.org/Reports/2012/Twitter-Use-2012.aspx>.

Sobkowicz, P., Kaschesky, M., Bouchard, G., 2012. Opinion mining in social media: modeling, simulating, and forecasting political opinions in the web. Gov. Inf. Q. 29 (4), 470–479.

Spellmann, H., Albert, M., Schmidt, M., Sutmöller, J., Overbeck, M., 2011. Waldbauliche Anpassungsstrategien für veränderte Klimaverhältnisse. AFZ – Der Wald 11, 19–23.

Starbird, K., Palen, L., Hughes, A.L., Vieweg, S., 2010. Chatter on the red: what hazards threat reveals about the social life of microblogged information. In: Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work – CSCW '10. ACM Press, New York, New York, USA, pp. 241–250.

Steinfurt Tweet (@steinfurt_tweet). 2012. "Eichenprozessionsspinner – ein Nachtfalter erreicht Steinfurt, Gefahr durch giftige Raupenhaare http://bit.ly/LFbenw". 29 June. 2012, 11:35 am. Tweet. <https://twitter.com/steinfurt_tweet/status/218638787192958976>.

UN-CBD. 2012. Secretariat of the Convention on Biological Diversity (2012) Cities and Biodiversity Outlook. Montreal, 64 pages., Technical report, Montreal. (accessed 11.07.13). <http://www.cbd.int/en/subnational/partners-and-initiatives/cbo>.

van Oudenhoven, A., van Vliet, A., Moraal, L., 2008. Climate change exacerbates the oak processionary caterpillar problem in The Netherlands. In: KNPV Symposium Pests and Climate Change. 3 December, 2008, Wageningen, The Netherlands.

Vicente, C., Espada, M., Vieira, P., Mota, M., 2012. Pine Wilt Disease: a threat to European forestry. Eur. J. Plant Pathol. 133 (1), 89–99.

Vieweg, S., Hughes, A.L., Starbird, K., Palen, L., 2010. Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In: 'Proceedings of the 28th International Conference on Human Factors in Computing Systems – CHI '10'. ACM Press, New York, New York, USA, p. 1079.

Wang, F.-Y., 2011. Social Media and the Jasmine Revolution. IEEE Intell. Syst. 26 (2), 2–4.

Wikipedia. 2012. Web 2.0 (Wikipedia article). (accessed 11.07.13). <http://en.wikipedia.org/wiki/Web_2.0>.

Wintle, B.A., Runge, M.C., Bekessy, S.A., 2010. Allocating monitoring effort in the face of unknown unknowns. Ecol. Lett. 13 (11), 1325–1337.

Yu, P.S., Han, J., Faloutsos, C., 2010. Link Mining: Models, Algorithms, and Applications. Springer.

Zalasiewicz, J., Williams, M., Steffen, W., Crutzen, P., 2010. The new world of the Anthropocene. Environ. Sci. Technol. 44 (7), 2228–2231.

Paper II

Forest Ecosystems
a SpringerOpen Journal

RESEARCH ARTICLE Open Access

Assessing citizen science opportunities in forest monitoring using probabilistic topic modelling

Stefan Daume[1,2*], Matthias Albert[3] and Klaus von Gadow[1]

Abstract

Background: With mounting global environmental, social and economic pressures the resilience and stability of forests and thus the provisioning of vital ecosystem services is increasingly threatened. Intensified monitoring can help to detect ecological threats and changes earlier, but monitoring resources are limited. Participatory forest monitoring with the help of "citizen scientists" can provide additional resources for forest monitoring and at the same time help to communicate with stakeholders and the general public. Examples for citizen science projects in the forestry domain can be found but a solid, applicable larger framework to utilise public participation in the area of forest monitoring seems to be lacking. We propose that a better understanding of shared and related topics in citizen science and forest monitoring might be a first step towards such a framework.

Methods: We conduct a systematic meta-analysis of 1015 publication abstracts addressing "forest monitoring" and "citizen science" in order to explore the combined topical landscape of these subjects. We employ 'topic modelling', an unsupervised probabilistic machine learning method, to identify latent shared topics in the analysed publications.

Results: We find that large shared topics exist, but that these are primarily topics that would be expected in scientific publications in general. Common domain-specific topics are under-represented and indicate a topical separation of the two document sets on "forest monitoring" and "citizen science" and thus the represented domains. While topic modelling as a method proves to be a scalable and useful analytical tool, we propose that our approach could deliver even more useful data if a larger document set and full-text publications would be available for analysis.

Conclusions: We propose that these results, together with the observation of non-shared but related topics, point at under-utilised opportunities for public participation in forest monitoring. Citizen science could be applied as a versatile tool in forest ecosystems monitoring, complementing traditional forest monitoring programmes, assisting early threat recognition and helping to connect forest management with the general public. We conclude that our presented approach should be pursued further as it may aid the understanding and setup of citizen science efforts in the forest monitoring domain.

Keywords: Forest monitoring; Citizen science; Participatory forest monitoring; Probabilistic topic modelling; Text analysis

* Correspondence: stefan.daume@ecoveillance.org
[1]Faculty of Forest Sciences and Forest Ecology, Georg-August-University Göttingen, Büsgenweg 5, 37077 Göttingen, Germany
[2]Department of Biodiversity Informatics, Swedish Museum of Natural History, Box 50007, 104 05 Stockholm, Sweden
Full list of author information is available at the end of the article

Background

The ability of ecosystems worldwide to provide essential products and services is being threatened by major environmental, social and economic changes (Millenium Ecosystem Assessment 2005), and there is a rising demand for intensive monitoring to detect threats and potentially catastrophic changes earlier (Biggs et al. 2009). Forests provide many vital ecosystem services, but with increasing ecological and economic pressures their resilience and stability are under threat. Forest managers and scientists are thus required to constantly re-evaluate and communicate strategies for intensified monitoring; this includes general environmental monitoring for emerging threats but also "traditional" forest monitoring using field plots and remote sensing for forest management purposes. In addition, there is an urgent need to inform and educate the general public on the value of forest ecosystems and the direct and indirect anthropogenic influences on forests (European Environment Agency 2011a; European Environment Agency 2011b).

Participatory forest monitoring – involving local communities and stakeholders in forest monitoring activities - plays an increasingly important role in delivering useful information, especially in areas where communities are relying heavily on forests for their livelihood and where a community's forest use can have massive impacts on the ecosystem (Evans and Guariguata 2008). Participatory monitoring is thus one avenue to provide additional resources to intensify forest monitoring.

In research generally, *"citizen science"* – the volunteer participation of members of the public in scientific projects - has emerged as a valuable tool in data collection, processing and dissemination, and offers effective channels for educating the general public on research (Bonney et al. 2009). Many citizen science projects cover subjects in the environmental domain (Silvertown 2009; Bonney et al. 2009), but citizen science extends over a broad set of application areas (such as astronomy, cancer research, etc.) utilising a wide range of skills, interests and motivations.

Citizen science biodiversity monitoring projects in general (Silvertown 2009) can potentially deliver information relevant to forest monitoring programmes. In fact, volunteers are already contributing to specific forest monitoring challenges. The *Living Ash Project* (http://livingashproject.org.uk/) for example aims to counter the effects of *Ash dieback* disease by calling for members of the public to tag and regularly monitor ash trees with the long-term objective to identify pest-resistant trees. Mobile and web technologies in particular help to facilitate these contributions: Ferster and Coops (2014) report that citizen scientists can use smartphone applications to collect data on forest fuel loading to identify wildfire hazards, and the *Forest Watchers* web application (http://forestwatchers.net) calls on volunteers to identify remote deforested areas in aerial images.

While these projects can make a potentially dramatic difference to existing monitoring efforts, they still represent singular and often localized efforts. A solid, generic and applicable framework or toolset for utilising the true potential of citizen science projects in the forestry domain still seems to be lacking. We propose that a better understanding of shared and related topics in citizen science and forest monitoring can be a first step towards opening up citizen science as an additional resource in the forest monitoring toolset.

Accordingly, this contribution explores the potential of citizen science initiatives in forest monitoring from a high-level perspective through an assessment of topical overlaps in the published literature on "citizen science" and "forest monitoring". Specifically, we are interested in a fine-grained analysis and the discovery of latent topics that may point to opportunities in employing citizen science for the benefit of forest science. Such a meta-analysis could be a first step in encouraging new developments and specific designs of citizen science initiatives in the forest monitoring domain.

Data and methods

For the proposed meta-analysis we employ an approach known as "topic modelling", an unsupervised probabilistic machine learning method for the automatic analysis of large text collections (Blei 2012). Topic modelling has seen a rising number of applications in recent years with an emphasis on applications in the digital humanities (Blevins 2010; Templeton et al. 2011; Yang et al. 2011), but also for bibliometric analysis of publications in the natural sciences (Griffiths and Steyvers 2004; Blei and Lafferty 2007). The technique has been employed both to discover topics in text collections and to structure document sets for advanced searching.

In this study, we apply probabilistic topic modelling to analyse a combined collection of scientific articles on the subjects of "forest monitoring" and "citizen science". We aim to provide a description of the combined topical landscape of these two broad thematic sets of publications, explore to what extent shared topics exist, which topics are clearly separated but potentially related and discuss the potential of this approach in providing new insights and opportunities for citizen science applications in the forest monitoring domain.

Data

We applied topic modelling to a set of documents obtained through a search in the literature database Scopus for documents published from 1994 to 2013, explicitly mentioning the terms "forest monitoring" or "citizen science" in the

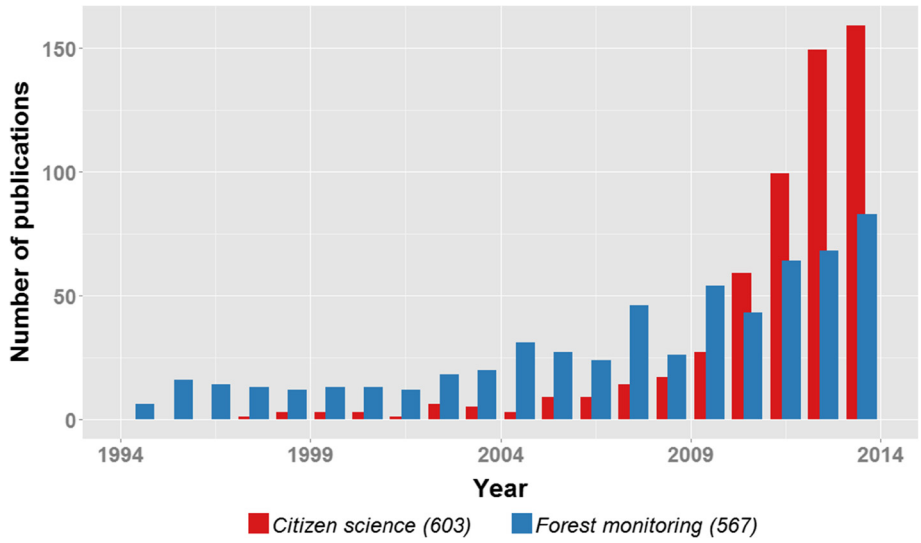

Figure 1 Distribution of "citizen science" and "forest monitoring" publications from 1994 to 2013 according to Scopus. The topic analysis included publications explicitly mentioning "citizen science" or "forest monitoring" in the title, abstract or keywords (based on a search in Scopus).

title, abstract or keywords. Figure 1 shows the development of the number of publications for the two document sets. The increase of the citizen science material since 2005 is rather dramatic. Only two articles (Roman et al. 2013; Butt et al. 2013), both published in 2013, contained both search terms; for our analysis we assigned Roman et al. (2013) to the "forest monitoring" and Butt et al. (2013) to the "citizen science" document set.

We obtained the abstracts for each matching publication for analysis, but excluded all documents not published in English as well as documents with abstracts of less than 100 words, which left 477 documents on "citizen science" and 538 documents on "forest monitoring". Many of the "forest monitoring" publications present a global coverage, though with an apparent bias towards studies focusing on Europe and North America. Our "citizen science" publications refer almost exclusively to projects in North America and Europe. This bias is also reflected by the geographical distribution of the corresponding authors of the two sets of publications.

Prior to running the topic modeller the text corpus is split into tokens and so-called stop-words (e.g. "the", "and", "if") are removed. The quality of the topic analysis can often be further improved by removing additional domain specific stop-words; we added "citizen", "science", "forest" and "monitoring" to the stop-word list since one of either combination would have occurred in every document which effectively turned them into stop-words. In addition, all words occurring only once were removed from the text corpus. This left us with a vocabulary of 6.181 unique terms, occurring a total of 100.274 times in the 1.015 abstracts.

Probabilistic topic modelling

Probabilistic topic models represent a suite of algorithms for analysing large document collections and identifying the distinct latent topics in these documents (Steyvers and Griffiths 2007; Blei 2012). Topic models are based on the assumption that documents are typically composed of multiple topics. Each topic in turn may be viewed as a distinct set of unique words that frequently occur together. All documents in a set share multiple topics, but individual documents will exhibit only a subset of all available topics to a certain degree. More formally let:

- W be the unique set of words in
- D a set of documents containing
- T topics, where
- each topic t is a discrete probability distribution Φ_t over all words w and
- each document d has a specific distribution Θ_d over all topics T.

Topic modelling is based on the assumption that each document d is the result of a generative process by which iteratively a topic t is first drawn from Θ_d and then a word w is drawn from Φ_t until the document is complete. Topic modelling algorithms reverse this assumed document creation process in order to infer topics and topic compositions that best explain a set of observed variables, here represented by the word occurrences in a given set of documents.

Figure 2 illustrates informally the intuition behind topic modelling: assuming that the topic composition of a document and the frequencies with which words appear

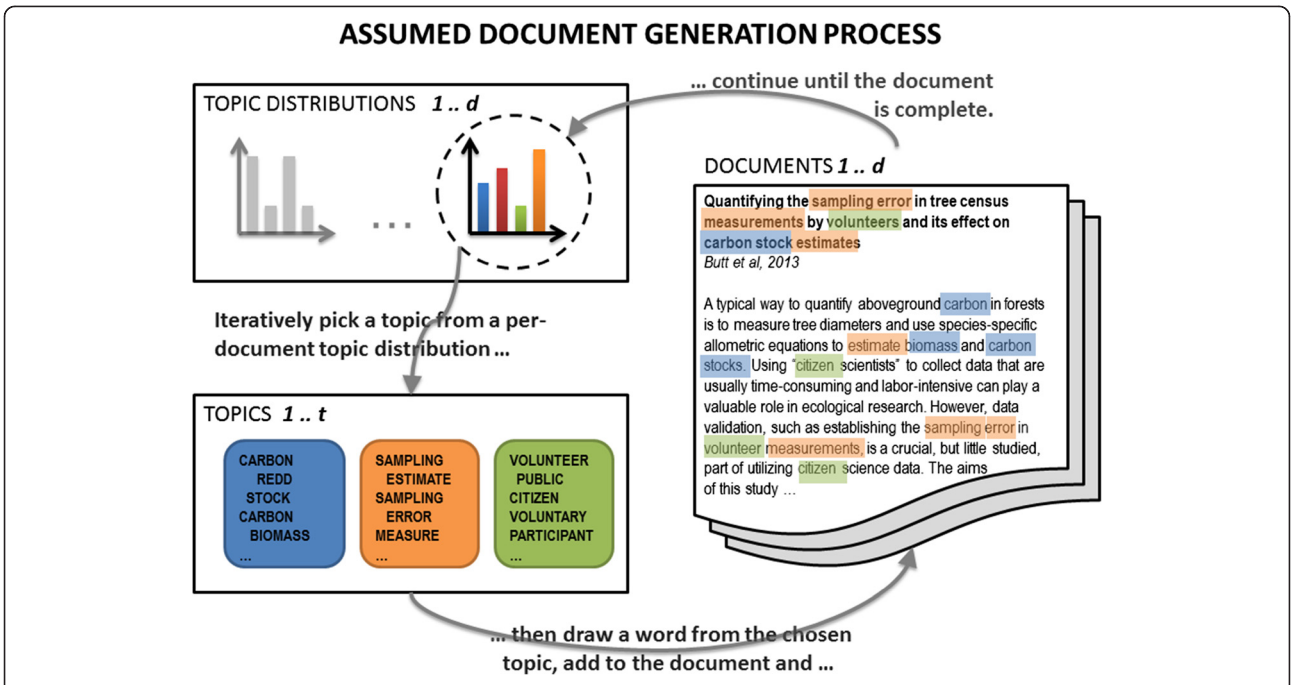

Figure 2 Informal explanation of the intuition behind topic modelling (adapted from (Blei 2012)). An excerpt of a sample abstract (Butt et al. 2013) is quoted for illustration purposes. A document is assumed to be the result of an iterative process that selects topics from a document-specific probabilistic topic distribution and words from a topic-specific probabilistic word distribution. Topic modelling algorithms reverse this process in order to find assignments for the two distributions that best explain a set of documents.

in a topic are known, a document can be generated by iteratively choosing words from a topic according to the frequencies of the topics. A topic modelling algorithm then reverses this process by, simply put, assigning the words in a given "observed" set of documents to topics, and topic distributions to documents, such that a set of documents generated on the basis of these distributions best fits the set of "observed" documents.

Topic models typically employ variational inference (Asuncion et al. 2009) to estimate the best topic-word and topic-document assignments. We use a topic model called *Latent Dirichlet Allocation (LDA)* first described by Blei et al. (Blei et al. 2003). In LDA the assumed prior distributions for Θ_d and Φ_t are Dirichlet distributions with concentration parameters α and β respectively. The choice of these so-called hyper-parameters determines the sparsity of the distributions and thus the variability in likelihood with which words will be assigned to topics and topics to documents. LDA has emerged as a reliable and popular topic modelling approach successfully applied in many different domains. Furthermore, it offers several freely available implementations. We use the MALLET machine learning package (McCallum 2002) which provides an open source implementation of LDA.

LDA configuration – choosing the number of topics

A key choice in running topic modelling algorithms is the number of distinct topics that are expected to be covered by the document corpus. The number of topics T and the priors α and β are the only required input parameter for LDA, but they have a significant impact on the resulting topic assignments. Choosing larger topic numbers may result in a fragmentation of topics which may not always be easy to interpret semantically. However, MALLET offers a feature called *hyper-parameter optimisation* which alleviates the impact of the chosen topic number (Wallach et al. 2009a) and allows to safely work with larger topic numbers.

It can be argued that the choice of T is ultimately an arbitrary one driven by the research questions and the intended use of the resulting topic model; small topic numbers will result in semantically broad topics, with increasing topic numbers, those broader topics will be split in semantically more refined topics. Several evaluation methods allow however a quantitative assessment of the optimal number of topics (Wallach et al. 2009b). We followed an approach chosen by Griffiths and Steyvers (2004) and compared the converged log-likelihood (LL) per token (returned by the LDA algorithm) as a measure of best model fit for topic numbers T ranging from 10 to 300. We repeated 10 topic analyses for each T in this range and measured the final LL/token which suggested 100 topics as a suitable topic number for our analysis (see Figure 3).

We thus ran MALLET's implementation of LDA with 100 topics. The algorithm was run for 2000 iterations

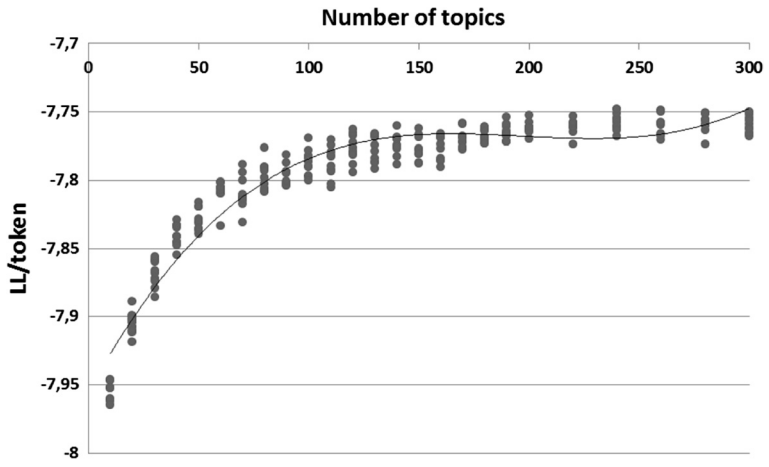

Figure 3 Evaluation of topic model fit with different topic numbers. The relationship between the number of topics and the log-likelihood (LL) per token as a measure of best model fit for topic numbers is shown for 10 sample MALLET LDA modelling runs for topic numbers in the range from 10 to 300.

with the hyper-parameter optimization feature enabled, producing a topic model that will be presented in the next section.

Results

The MALLET LDA topic modelling implementation produces two main outputs that will be referred to further analysis:

1. **Topic word sets for each topic:** the collection of terms with associated occurrence frequencies that characterise a topic.
2. **Topic composition for each analysed document:** the share of each topic in a given document.

Table 1 lists a sample selection of the 100 topics returned by the topic modelling algorithm for our document corpus. For each topic the 10 most frequent terms and relative word frequencies in the topic are provided. It is important to note that the topic modelling algorithm returns purely a distribution of topic terms that do not come with a semantic interpretation. Suitable labels can however often be inferred from the word frequencies. In the following we will either refer to a topic by its ID (0–99) assigned by the topic modelling algorithm or labels that we assigned on inspection of the most frequent terms in a topic.

A term is not necessarily exclusive to one topic. We find the term "results" for example as a top term in both topics 69 and 38 (see Table 1). In both cases it co-occurs with terms that are characteristic for scientific publications in general and would thus be expected in publications on citizen science as well as forest monitoring; both topics were accordingly labelled "science study". The term "change" can be found in topic 38 ("science study") and topic 32 ("climate change"). For generic terms like this the most frequent co-occurring terms as well as the term's specificity to a topic can help to infer suitable labels. This may also clarify topic semantics in case of ambiguous term combinations. Topic 85 for example combines astronomy terms like "galaxy" and "supernovae" with "dna" and "genetic". Figure 4 plots terms according to their frequency in and specificity to a topic for three sample topics. Considering these two dimensions suitable topic labels - here "galaxies", "risk perceptions" and "birds" - can usually be suggested even for heterogeneous or ambiguous word combinations.

For each document in the analysed corpus the resulting topic model will include a topic composition distribution which specifies the shares of each topic in a given document. Figure 5 shows a sample topic composition for one (Butt et al. 2013) of the two publications that matched both the search term "citizen science" and "forest monitoring". This example illustrates that only a small number of topics are active in this document. A comparison with the publication abstract confirms that the topic composition shown in Figure 5 appears to closely reflect the content of the analysed abstract.

For our analysis we were furthermore interested in the distribution of topics between the "citizen science" and "forest monitoring" document corpora. Figure 6 combines the topic composition of all analysed documents and shows the distribution of topics for the combined document corpora.

The cumulative topic distribution in Figure 6 includes only topic proportions greater than 0.02. The topic modelling algorithm attempts to assign a share of each of the chosen 100 topics for every document, but as the sample topic composition in Figure 5 illustrated, this will result in a large number of very small and negligible proportions.

Table 1 Most frequent words and relative word frequencies by topic for a sample set of topics

Topic 69 "science study"		**Topic 38 "science study"**		**Topic 67 "information systems"**		**Topic 32 "climate change"**	
Results	0.047	Study	0.029	Information	0.049	Climate	0.180
Methods	0.046	Results	0.027	Systems	0.036	Change	0.129
Method	0.044	Change	0.024	Development	0.035	Effects	0.037
Study	0.030	Time	0.022	Paper	0.033	Water	0.026
Accuracy	0.030	Analysis	0.019	Support	0.024	Ecosystems	0.024
Based	0.028	Studies	0.018	Developing	0.019	Response	0.022
Compared	0.023	Large	0.017	Process	0.019	Management	0.021
High	0.021	Significant	0.016	Framework	0.017	Integrated	0.016
Evaluated	0.014	Years	0.016	Key	0.016	Ground	0.016
Developed	0.014	Found	0.015	Based	0.013	Impacts	0.015
Topic 0 "volunteer surveys"		**Topic 30 "education"**		**Topic 24 "plant phenology"**		**Topic 85 "galaxies"**	
Volunteers	0.161	Students	0.112	Plant	0.082	Galaxy	0.065
Volunteer	0.118	Learning	0.056	Phenology	0.078	Galaxies	0.054
Collected	0.041	Education	0.056	Plants	0.066	Zoo	0.044
Scientists	0.030	Classroom	0.024	Species	0.059	Project	0.026
Groups	0.025	Teaching	0.020	Phenological	0.043	dna	0.026
Recording	0.021	School	0.019	Interactions	0.033	Morphological	0.026
Professional	0.020	Literacy	0.018	Network	0.032	Spiral	0.023
Surveying	0.020	Teachers	0.018	Networks	0.032	Supernovae	0.021
Environment	0.019	Educational	0.017	Observations	0.023	Genetic	0.021
Motivations	0.018	Experiences	0.013	Timing	0.023	Classifications	0.016
Topic 53 "forest growth"		**Topic 87 "SAR"**		**Topic 91 "ozone"**		**Topic 20 "carbon stocks"**	
Tree	0.126	Sar	0.068	Ozone	0.103	Carbon	0.091
Trees	0.082	Coherence	0.038	Concentrations	0.051	Redd	0.052
Growth	0.064	Radar	0.034	Sites	0.048	Countries	0.049
Species	0.032	Backscatter	0.032	Measured	0.036	National	0.044
Structure	0.032	l-band	0.023	Site	0.029	Change	0.032
Area	0.023	ers	0.022	Passive	0.027	Deforestation	0.031
Composition	0.021	Stands	0.022	Symptoms	0.023	Stocks	0.024
Plots	0.020	Biomass	0.022	Measurements	0.021	Climate	0.022
Diameter	0.020	Areas	0.020	Critical	0.019	Inventory	0.019
Conditions	0.017	Images	0.020	Sampling	0.019	Reporting	0.017

The table shows sample generic topics (top row), typical citizen science topics (middle row) and typical forest monitoring topics (bottom row). For each topic the topic ID and a representative label is provided.

The threshold of 0.02 was chosen, because the average number of words per analysed abstract after removal of stop-words was approximately 100 - a topic proportion of 0.02 thus corresponds to two words, which we propose is the absolute minimum for a semantic interpretation of a topic assignment. On average only 9 topic proportions per document are greater than 0.02, which however cumulatively explain approximately 90 % of the document.

Figure 6 illustrates that several topics have a large contribution from either corpus, thus occurring with high frequency and large proportions in both "citizen science" and "forest monitoring" publications - examples include topics 6, 38, 69 (see Table 1 for topic words). These topics combine keywords which are typical for scientific studies in general, thus topics which can be expected to be shared between the two corpora.

More specific large topics shared between the two corpora exist as well: topics 67 (labelled "information systems") and 88 ("large-scale analysis") are examples that seem to fit data intensive research fields. Given that these topics are characteristic for both domains the question arises whether shared topics in general point to synergies that could guide intensified citizen science contributions in forest monitoring. Similarly, examples

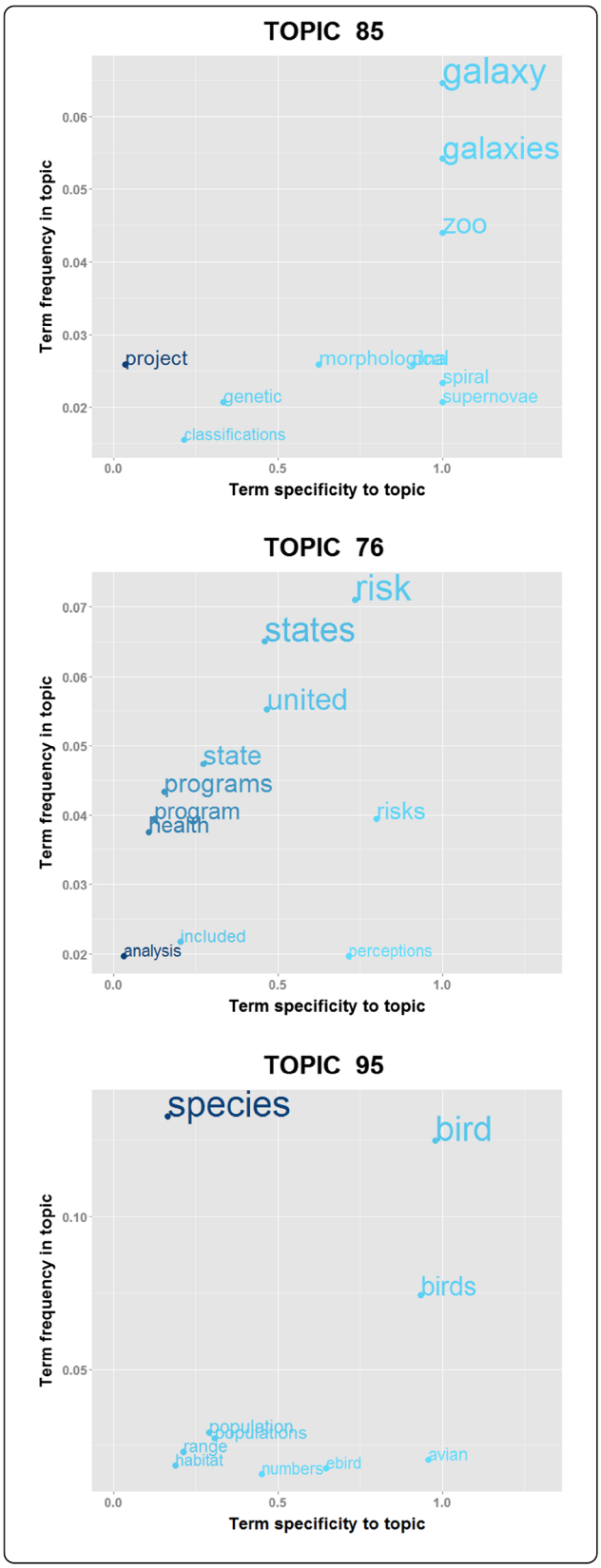

Figure 4 Term/topic frequency and specificity for three sample topics: "galaxies" (T85), "risk perceptions" (T76), "birds" (T95). A term that is exclusive to one topic has a specificity of 1. The relative size of the plotted words is proportional to their frequency in the topic. The colour gradient from light to dark blue indicates larger frequencies of a word in the complete document set.

of specific and shared but less frequent topics - for example 47 ("local/community-based"), 76 ("risk perceptions"), 96 ("urban environments") or 97 ("natural resource management") - have to be evaluated from the same angle and we will refer to those in more detail in the discussion section.

Typical or exclusive topics in either the "citizen science" or "forest monitoring" publications are also of particular interest. Here the question arises whether these are niche topics, truly un-related domains or potential examples of non-utilised citizen science opportunities in forest monitoring. Examples that we can discuss in "citizen science" are topics such as 0 ("volunteer surveys"), 24 ("phenology"), 27 ("citizen science"), 35 ("social media") and 85 ("galaxies"); dominant topics in the "forest monitoring" corpus include 1 ("crown studies"), 16 ("clearcuts"), 53 ("forest growth") and 87 ("SAR/remote sensing").

We conclude the result section with a network representation of the topical landscape of the analysed documents (Figure 7). Each document and topic is represented by a node in the network graph. An arc between a document and a topic represents a share of this topic in the connected document. The size of the topic nodes reflects their overall share in the analysed corpus. All topic proportions less than 0.02 in an individual document were excluded from this representation.

The network representation in Figure 7 provides a comprehensive visual summary of the topical structure of the combined document corpus, and confirms and extends the results in Figure 6. While the two document corpora have an intersection around major generic shared topics such as 6/38/69 ("science study"), 67 ("information systems") or 88 ("large-scale analysis"), they are visually clearly separated in the network layout. Corpus-specific topics such as 39 ("remote sensing") or 27 ("citizen science") are located in the centre of the respective document cloud, confirming that they are largely exclusive to these document corpora.

Discussion

Given the global pressures on forest ecosystems and the resulting challenges for forest managers and researchers, forest monitoring can benefit from additional and intensified efforts through citizen science projects. However, the topical landscape obtained through our analysis suggests that this opportunity is not yet pursued to a large extent. Although shared topics exist, the obtained topic

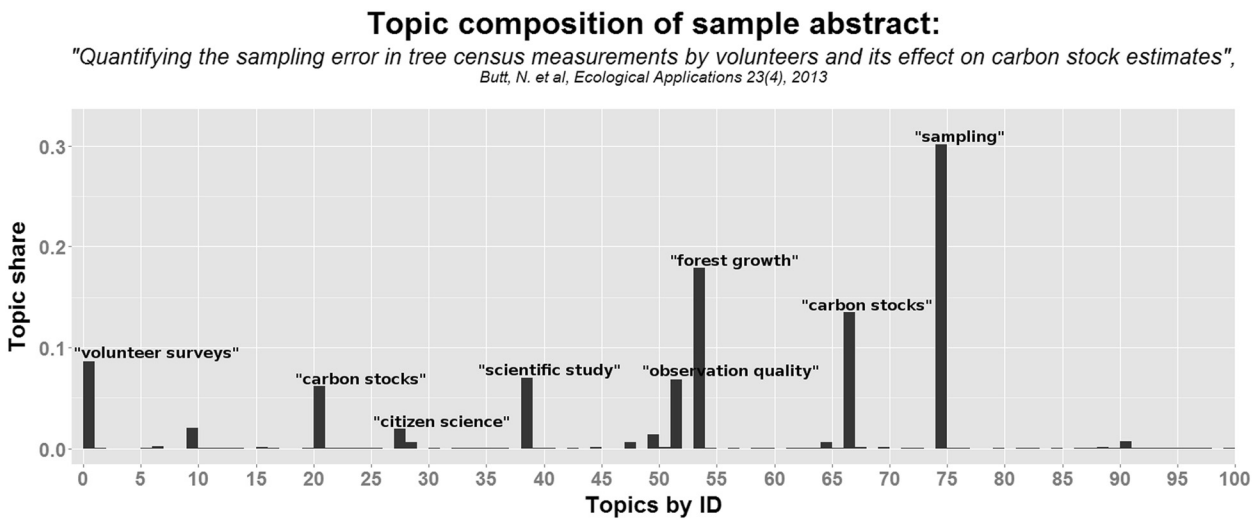

Figure 5 Topic composition of one publication abstract (Butt et al. 2013). The topic proportions reflect the output of a MALLET LDA topic analysis with 100 topics. Topic labels were added for the major topics.

model confirms the results hinted at by only two publications (Roman et al. 2013; Butt et al. 2013) in our document set that matched both the terms "citizen science" and "forest monitoring".

Obviously, the generalizability and conclusiveness of the results is limited by the size of the document collection and the analysed documents. Compared to similar studies – for example (Griffiths and Steyvers 2004), which used 28.154 abstracts, with more than 3 million words and a vocabulary of 20.551 words - our set of 1.015 abstract and a vocabulary of 6.181 terms occurring 100.274 times is significantly smaller. We were nevertheless able to identify many topics with consistent semantics - see for example topic 30 ("education") or topic 95 ("birds") – and the topic composition of sample articles (see Figure 5) seemed to reflect the content well. However, we also found topics like 85, which – while dominated by terms justifying the label "galaxies" - also included terms referring to genomics ("dna", "genetic"), pointing at a lack of granularity that can be attributed to the size of the document set and the vocabulary. A closer look at topic compositions of several sample documents in our set suggests that the quality of the topic assignments for a document correlates with the size of the text - longer abstracts display a more representative topic composition; taking into account other case studies in topic modelling we conclude that larger documents and thus vocabularies would probably deliver more representative topic structures for individual documents and topics with more refined and consistent semantics. A further improvement in the semantic interpretation and consistency of discovered topics might be achieved by exploring variations of topic models that consider word bigrams – the reoccurrence of e.g. the bigram "biodiversity loss" allows a more conclusive

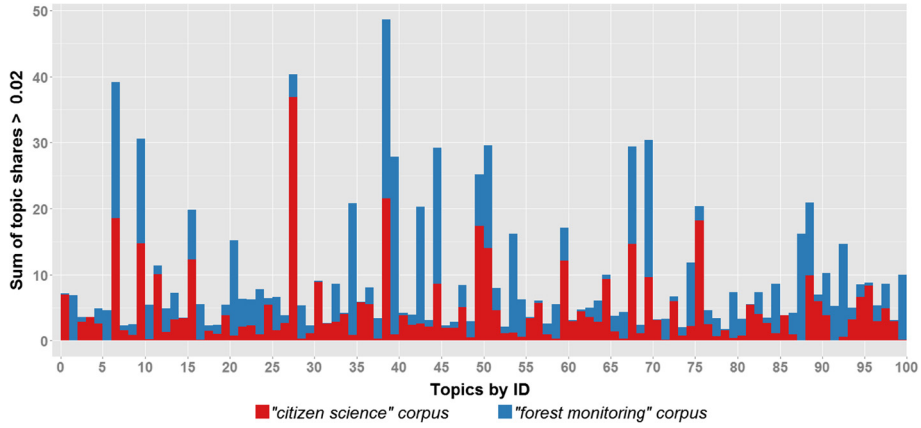

Figure 6 Contributions of topic shares by analysed document corpus. Cumulative shares per topic contributed by the "citizen science" (red) and "forest monitoring" (blue) corpus respectively. Result of a MALLET LDA topic analysis with 100 topics.

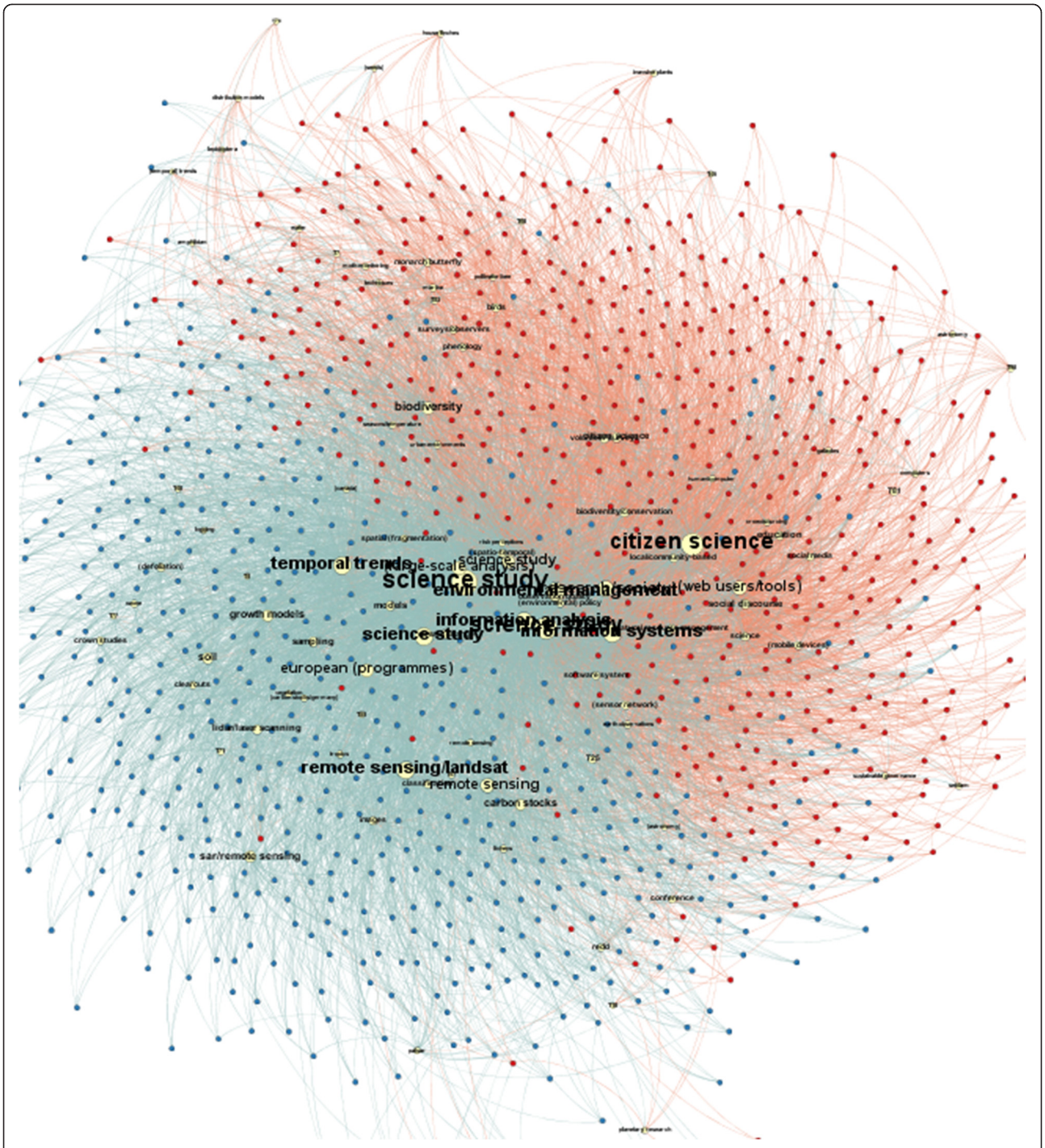

Figure 7 Network representation of the topics and documents in the analysed corpus. Blue nodes represent articles on "forest monitoring", red nodes articles on "citizen science" and yellow nodes the topics they are connected with; only topic shares greater than 0.02 were included as arcs in the network. The graph was generated with the Gephi (https://gephi.org) network visualization tool using a Fruchterman-Reingold network layout.

interpretation of a topic than the individual occurrence of the two words "biodiversity" and "loss".

Despite these potential methodical improvements, we nevertheless gained interesting initial insights in the combined topical landscape of "citizen science" and "forest monitoring" publications. Shared topics can be found and extend beyond the generic topics that would be expected in scientific publications in general. Shared topics such as "urban environments" (96) or "local/community-based" (47) indicate common themes, topics

such as "spatio-temporal" (82) or "software development" (13) common tools and techniques, and topics like "risk perceptions" (76) or "climate change" (32) hint at shared research motivations. It can be argued that these results are "stating the obvious" and similar results could be obtained through traditional manual literature analysis. However, topic modelling is a scalable method that can be applied equally to very large document corpora, full-text analysis of publications and a much larger number of topics, and thus suggests topic modelling as a suitable method not only for a snapshot analysis but also for a continuous analysis of growing document sets. In addition, the consistency of our results with "the obvious" supports our other observations for the document set, particularly that major forest monitoring topics – e.g. "carbon stock" (20) estimates or "forest growth" (53) – are not shared between the document corpora.

In contrast, certain topics with large cumulative shares in the document corpus which are exclusive to or typical for either "citizen science" or "forest monitoring" publications point at interesting opportunities. Topics such as "galaxies" (85) or "astronomy" (43) indicate successful citizen science projects involving the analysis and classification of telescopic images by volunteers. Several articles in our corpus refer for example to the *Galaxy Zoo Supernovae* project (http://supernova.galaxyzoo.org) on the *Zooniverse* citizen science web platform, where volunteer participants were asked to compare changes between images of a specific region of the night sky taken at different times in order to identify supernovae (Smith et al. 2011). Participants were not required to have a background in astronomy, but still delivered classification results of "remarkable quality" (Smith et al. 2011).

In the forest monitoring corpus "remote sensing" (topics 42/87) emerged as a major topic (see Figure 7) and an area that will involve similar tasks and skills as the classification of telescopic images in the supernovae project. While rooted in different domains, both topics focus on image analysis and classification and thus have not only skill sets and techniques in common, but possibly also a citizen science community that could be mobilised for citizen science initiatives in the forest monitoring domain. Indeed, examples of remote sensing projects with volunteer participation can be found, for example in land cover monitoring (http://geo-wiki.org (Fritz et al. 2009), http://forestwatchers.net) or biomass estimates (Fritz et al. 2013), but are still an exception. A possible explanation is that "citizen science" as a research tool is still at an early stage of recognition in the forest monitoring domain, but also that there are concerns over the quality of citizen science data which will determine the applicability of inferred results (See et al. 2013). With reference to the example of remote sensing we propose that an understanding of topical landscapes across domains could contribute to citizen science projects delivering high quality data and results by learning from communities with similar tasks and techniques, finding participants with matching skills and utilising tested frameworks from other domains.

More topic examples largely exclusive to the forest monitoring document corpus which might benefit from intensified monitoring through citizen science are e.g. "carbon stock" estimates (20) or "ozone" effects (91) – citizen science has been explored in these areas (see for example (Sachs 2008)), but not as major research tool in forest monitoring. The topic "education" (30) on the other hand is almost exclusively found in the "citizen science" domain. In light of an increased need to communicate forest policies, threats and values to the general public, this observation points to citizen science as an important communication channel that should find more consideration in the forestry domain.

This exploratory study indicates that the two research areas represented by the document corpora on citizen science and forest monitoring exhibit shared topics, but that promising opportunities to utilise citizen science for key forest monitoring themes still lie dormant. Citizen science projects will be most successful, both in terms of research outcomes and the perceived value for participating volunteers, when projects are designed with a good understanding of the formal models of participation (Bonney et al. 2009; Shirk et al. 2012) and a clear alignment with key research process steps (Newman et al. 2012). We believe that the consideration of the combined topical landscape of citizen science and its (potential) application areas can contribute to the deliberate design of citizen science projects and the success of these projects. The discovery of shared latent topics could be of value when directing researchers and stakeholders in either field to matching resources (articles, studies, methods), connect communities and thus facilitate citizen science projects in the forest domain.

However, these first findings - while intriguing - are still too limited to permit general conclusions. We believe that our initial results confirm topic modelling as a valuable method, but that the conclusiveness of the results could be improved by broadening the thematic scope and the size and number of the analysed documents - for this exploratory analysis we chose to focus on publications explicitly mentioning the terms "citizen science" and "forest monitoring" and hence missed, by design, many citizen science projects in for example forest threat monitoring; furthermore, we analysed abstracts only.

Future research should therefore not only extend the topic analysis to full-text articles but should also pursue a broader thematic focus and include publications from other databases such as NGO project studies as well as publications that apply a different terminology to the subject area for example by using terms like "crowdsourcing",

"public participation in research", "forest inventory", "forest modelling" or "forest planning" instead of "citizen science" and "forest monitoring". When using a larger dataset and running the topic analysis with larger numbers of topics, more-fine-grained topics pointing to specific techniques, skill sets or communities might emerge that would allow to draw conclusions that are more generalizable and point to specific promising citizen science opportunities in the forest monitoring domain.

Conclusions

The application of probabilistic topic modelling for characterizing the shared topical landscape of publications on citizen science and forest monitoring confirmed that the method is useful as a scalable approach for a meta-analysis of large document collections in the chosen domain. While the conclusiveness of the findings is somewhat limited by the number of documents analysed, even this exploratory topic analysis indicates interesting shared motivations and skills, and under-utilised opportunities for citizen science projects in forest monitoring can be inferred from this study.

Citizen science projects in the area of forest monitoring have the potential to contribute to the earlier recognition of forest threats, supplement resources in traditional inventory programs, provide pointers for areas requiring intensified monitoring, indicate public demands on forests and connect forest practitioners and researchers with the general public. In the interest of utilising citizen science for intensified monitoring efforts, communication and public awareness, the presented topic modelling approach should be pursued further and may assist both citizen science and forest monitoring communities in connecting resources and stakeholders, thus possibly aiding in the future deliberate design of more numerous and ambitious citizen science initiatives in the forestry domain.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

The first author (SD) conducted the data collection and topic analysis, compiled the results and wrote the first draft of the article. SD, MA and KG then jointly restructured and improved the presentation during several major revisions. All authors read and approved the final manuscript.

Acknowledgements

This study was inspired by many discussions on topic modelling the first author had with Emma Sundström and Ingo Fetzer of the Stockholm Resilience Centre. This input is gratefully acknowledged. Finally, the authors would like to thank the anonymous reviewers for their valuable feedback which helped to improve this publication.

Author details

[1]Faculty of Forest Sciences and Forest Ecology, Georg-August-University Göttingen, Büsgenweg 5, 37077 Göttingen, Germany. [2]Department of Biodiversity Informatics, Swedish Museum of Natural History, Box 50007, 104 05 Stockholm, Sweden. [3]Northwest German Forest Research Institute, Grätzelstraße 2, 37079 Göttingen, Germany.

Received: 29 May 2014 Accepted: 10 July 2014
Published: 30 July 2014

References

Asuncion A, Welling M, Smyth P, Teh YW (2009) On Smoothing and Inference for Topic Models. UAI '09 Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, AUAI Press, Arlington, Virginia, United States, pp 27–34

Biggs R, Carpenter SR, Brock WA (2009) Turning back from the brink: detecting an impending regime shift in time to avert it. Proc Natl Acad Sci U S A 106:826–831, doi:10.1073/pnas.0811729106

Blei D (2012) Probabilistic topic models. Commun ACM 55:77–84, doi:10.1109/MSP.2010.938079

Blei DM, Lafferty JD (2007) A correlated topic model of science. Ann Appl Stat 1:17–35

Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. J Mach Learn Res 3:993–1022

Blevins C (2010) Topic Modeling Martha Ballard's Diary. In: Pers. Blog. http://historying.org/2010/04/01/topic-modeling-martha-ballards-diary/. Accessed 2 May 2014

Bonney R, Cooper CB, Dickinson J, Kelling S, Phillips T, Rosenberg KV, Shirk J (2009) Citizen science: a developing tool for expanding science knowledge and scientific literacy. Bioscience 59:977–984, doi:10.1525/bio.2009.59.11.9

Butt N, Slade E, Thompson J, Malhi Y, Riutta T (2013) Quantifying the sampling error in tree census measurements by volunteers and its effect on carbon stock estimates. Ecol Appl 23:936–943, doi:10.1890/11-2059.1

European Environment Agency (2011a) Forests, health and climate change - Urban green spaces, forests for cooler cities and healthier people. http://www.eea.europa.eu/publications/forests-health-and-climate-change

European Environment Agency (2011b) Europe's forests at a glance — a breath of fresh air in a changing climate. http://www.eea.europa.eu/publications/europes-forests-at-a-glance

Evans K, Guariguata MR (2008) Participatory Monitoring in tropical forest management - a review of tools, concepts and lessons learned. Center for International Forestry Research (CIFOR), Bogor, Indonesia, p 50

Ferster CJ, Coops NC (2014) Assessing the quality of forest fuel loading data collected using public participation methods and smartphones. Int J Wildl Fire doi:10.1071/WF13173

Fritz S, McCallum I, Schill C, Perger C, Grillmayer R, Achard F, Kraxner F, Obersteiner M (2009) Geo-Wiki.Org: the use of crowdsourcing to improve global land cover. Remote Sens 1:345–354, doi:10.3390/rs1030345

Fritz S, See L, van der Velde M, Nalepa RA, Perger C, Schill C, McCallum I, Schepaschenko D, Kraxner F, Cai X, Zhang X, Ortner S, Hazarika R, Cipriani A, Di Bella C, Rabia AH, Garcia A, Vakolyuk M, Singha K, Beget ME, Erasmi S, Albrecht F, Shaw B, Obersteiner M (2013) Downgrading recent estimates of land available for biofuel production. Environ Sci Technol 47:1688–1694, doi:10.1021/es303141h

Griffiths TL, Steyvers M (2004) Finding scientific topics. Proc Natl Acad Sci U S A 101(Suppl):5228–5235, doi:10.1073/pnas.0307752101

McCallum AK (2002) MALLET: A Machine Learning Language Toolkit. http://mallet.cs.umass.edu/

Millenium Ecosystem Assessment (2005) Ecosystems and Human Well-Being: Synthesis (Millennium Ecosystem Assessment Series). p 160

Newman G, Wiggins A, Crall A, Graham E, Newman S, Crowston K (2012) The future of citizen science: emerging technologies and shifting paradigms. Front Ecol Environ 10:298–304, doi:10.1890/110294

Roman LA, Mcpherson EG, Scharenbroch BC, Bartens J (2013) Identifying common practices and challenges for local urban tree monitoring programs across the United States. Arboric Urban For 39:292–299

Sachs S (2008) Using Students to Monitor the Effects of Ground-level Ozone on Plants. In: Weber, Samantha, and David Harmon (ed) Rethinking Protected Areas in a Changing World: Proceedings of the 2007 GWS Biennial Conference on Parks, Protected Areas, and Cultural Sites. The George Wright Society, Hancock, MI, pp 277–279

See L, Comber A, Salk C, Fritz S, van der Velde M, Perger C, Schill C, McCallum I, Kraxner F, Obersteiner M (2013) Comparing the quality of crowdsourced data contributed by expert and non-experts. PLoS One 8:e69958, doi:10.1371/journal.pone.0069958

Shirk JL, Ballard HL, Wilderman CC, Phillips T, Wiggins A, Jordan R, McCallie E, Minarchek M, Lewenstein BV, Krasny ME, Bonney R (2012) Public participation

in scientific research: a framework for deliberate design. Ecol Soc 17:art29, doi:10.5751/ES-04705-170229
Silvertown J (2009) A new dawn for citizen science. Trends Ecol Evol 24:467–471, doi:10.1016/j.tree.2009.03.017
Smith AM, Lynn S, Sullivan M, Lintott CJ, Nugent PE, Botyanszki J, Kasliwal M, Quimby R, Bamford SP, Fortson LF, Schawinski K, Hook I, Blake S, Podsiadlowski P, Jönsson J, Gal-Yam A, Arcavi I, Howell DA, Bloom JS, Jacobsen J, Kulkarni SR, Law NM, Ofek EO, Walters R (2011) Galaxy zoo supernovae. Mon Not R Astron Soc 412:1309–1319, doi:10.1111/j.1365-2966.2010.17994.x
Steyvers M, Griffiths T (2007) Probabilistic topic models. In: Landauer T, McNamara D, Dennis S, Kintsch W (eds) Latent Semantic Analysis: A Road to Meaning. Lawrence Erlbaum, pp 427–449
Templeton C, Brown T, Bhattacharyya S, Boyd-Graber J (2011) Mining the Dispatch under Supervision: Using Casualty Counts to Guide Topics from the Richmond Daily Dispatch Corpus. Chicago Colloquium on Digital Humanities and Computer Science 2011, Chicago, IL
Wallach H, Mimno D, McCallum A (2009a) Rethinking LDA: Why Priors Matter. In: Advances in Neural Information Processing Systems 22, NIPS 2009 Proceedings. 23rd Annual Conference on Neural Information Processing Systems 2009, Vancouver, British Columbia, Canada, Proceedings of a meeting held 7-10 December 2009
Wallach HM, Murray I, Salakhutdinov R, Mimno D (2009b) Evaluation Methods for Topic Models. Proc. 26th Annu. Int. Conf. Mach. Learn. - ICML '09. ACM Press, New York, New York, USA, pp 1–8
Yang T-I, Torget AJ, Mihalcea R (2011) Topic modeling on historical newspapers. In: LaTeCH '11 Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities. Association for Computational Linguistics, Stroudsburg, PA, pp 96–104

doi:10.1186/s40663-014-0011-6
Cite this article as: Daume *et al.*: **Assessing citizen science opportunities in forest monitoring using probabilistic topic modelling.** *Forest Ecosystems* 2014 **1**:11.

Paper III

Mining Twitter to monitor invasive alien species – An analytical framework and sample information topologies

Stefan Daume[1,2,3,a]

[1]Faculty of Forest Sciences and Forest Ecology, Georg-August-University Göttingen, Büsgenweg 5, 37077 Göttingen, Germany

[2]Stockholm Resilience Centre, Stockholm University, SE-10691 Stockholm, Sweden

[3]Department of Biodiversity Informatics, Swedish Museum of Natural History, Box 50007, 104 05 Stockholm, Sweden

[a]stefan.daume@ecoveillance.org (Corresponding author)

(Manuscript prepared for submission)

Abstract

Social online media increasingly emerge as important informal information sources that can contribute to the detection of trends and early warnings in critical fields such as public health monitoring or emergency management. In the face of global environmental challenges the utilisation of this information in ecological monitoring contexts has been called for, but examples remain sparse. This can be attributed to the significant technical challenges in processing this data and concerns about the quality, reliability and applicability of information mined from social media to the ecological domain.

Here the strength and weaknesses of social media mining for ecological monitoring are assessed using the micro-blogging service Twitter and invasive alien species (IAS) monitoring as an example. The assessment is based on a manual analysis of 2842 Tweets sampled from Twitter data with potential direct or descriptive references to IAS impacting forest ecosystems, which was collected over a period of nearly three years. The results are presented as information topologies for Twitter messages of observational and non-observational character for three IAS with distinctive characteristics (Oak processionary moth, Emerald ash borer, Eastern grey squirrel).

The results show that the social media channel Twitter is a rich source of primary and secondary observational biodiversity information. It also provides useful insights in the topical landscape of public communications on IAS as well as the public perception of IAS and IAS management. The analysis suggests broad application opportunities in IAS monitoring and management, and points at applications for related environmental questions. The results highlight that social media mining for ecological monitoring needs to be approached with the same best practices as ecological monitoring in general, requiring a good understanding of the monitored subjects and specific monitoring questions.

The challenges in utilising this information for operational systems are of technical rather than conceptual nature and include extending the degree of automation, especially with regard to image recognition and the automatic provisioning of location information.

Keywords: *Twitter; invasive alien species; social media mining; ecological monitoring; biodiversity observations; forest ecosystems;*

1 Introduction

Social media such as Twitter or Facebook permeate online communications and have become an ubiquitous information dissemination and conversation channel utilised by 2 billion users worldwide (wearesocial.net, 2014), generating billions of messages on a daily basis. While exhibiting recognized geographical and demographic biases (Smith and Brenner, 2012), specific domain examples demonstrate that the content generated through these channels can provide a good reflection of societal realities and emerging trends (Conover et al., 2011; Salathé et al., 2012).

The sheer volume feeds the expectation that even for very specific domains interesting information for research and practical applications can be obtained. Aside from commercial applications (such as recommender systems (Garcia Esparza et al., 2012)), this is proven by a broad range of studies exploring social media sources with applications in critical fields and direct impacts on improved responses to critical challenges such as epidemic diseases (Signorini et al., 2011), emergency management (Vieweg et al., 2010) or earthquake detection (Crooks et al., 2013).

Applications with ecological focus are emerging, but remain sparse (Barve, 2014; Malcevschi et al., 2012; Stafford et al., 2010). This can be partly attributed to the fact that scholars and practitioners are rightfully sceptical about the usefulness of this information source, since the information stream will be dominated by banal to irrelevant content, the information is unstructured, biased and typically provided outside a specific monitoring context.

This raises practical and theoretical questions if and how this information can be integrated in traditional monitoring data and whether it can be fed as supplementary information into established ecological models commonly utilised to predict change or detect emerging threats (see for example (Graham et al., 2010; van Strien et al., 2013)). Furthermore, both the volume and the structural features of this data present practical challenges in utilising it, as are quality concerns comparable to those in semi-formal citizen science monitoring programs (Bird et al., 2013; Sheppard et al., 2014).

However, given the seriousness of environmental challenges that humanity is facing today (Steffen et al., 2011), an information stream that is easily accessible and reflects the observed reality and ambient trends of a quarter of the global population deserves a

methodical exploration and assessment. At best, social media can act as real-time data source and provide early warnings for pending and potentially irreversible shifts in ecosystems with large implications for human well-being (Biggs et al., 2009). Given the scale of these challenges it is thus paramount to utilise all available information to obtain early warnings or assist with adaptation to ecological changes that have already occurred (Galaz et al., 2010).

This article offers a methodical assessment of social media as sources in ecological monitoring, presents an exploratory framework and applies it to the example of invasive alien species and Twitter.

Twitter is a promising and representative example due to its large volume of users and posted messages (known as "Tweets") (Krikorian, 2013), its predominantly public data accessible through a public API and its focus on textual content. Furthermore, Twitter has been shown to act as a hub to other social media channels (De Longueville et al., 2009) and the constraint on 140 characters of the posted messages suggests a low contribution hurdle with an increased likelihood to share casual observations.

IAS represent a suitable choice for an assessment of social media in ecological monitoring as they are known indicators or drivers of ecosystem change (Crowl et al., 2008), a world-wide phenomenon (IUCN-ISSG, 2015) covering all types of ecosystems with a clear connection to a large variety of specific ecosystem services under threat (Pejchar and Mooney, 2009). Moreover, IAS are recognized as an important issue by policy-makers worldwide (European Commission, 2011; U.S. Government, 2010) from an ecological, economic (Pimentel et al., 2005) and security perspective (Meyerson and Reaser, 2003). Finally, IAS are often notable and easy to identify, and are thus likely to be a subject covered in social online media messages.

With reference to the chosen example of Twitter and IAS this contribution addresses the following major research questions:

1. What amount and type of data with domain-relevance can be found and what are its characteristics?
2. What contributions to ecological monitoring can be expected from this data?
3. Which challenges have to be addressed in order to utilise this information?

2 Methods and data

This contribution is based on samples from Twitter content on IAS that was collected over a period of nearly three years (2012-2014). The data collection and analysis was driven by an initial grouping of characteristic message attributes. Figure 1 summarizes this grouping along two dimensions: information relevance and information completeness (or verifiability). Both dimensions should be perceived as continuous rather than having distinct boundaries.

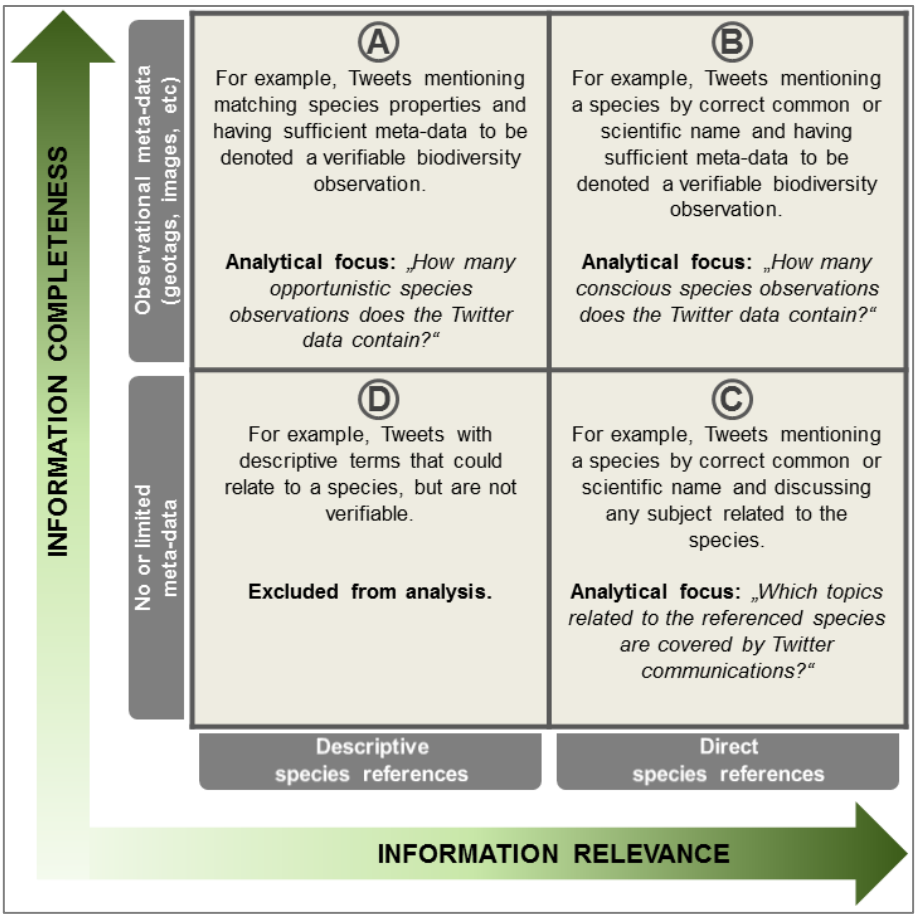

Figure 1. Major information groups in Twitter messages on IAS, organised as a grid around two dimensions with increasing information relevance and (meta-)data completeness implying different functions and significance in the context of IAS monitoring.

When considering Twitter for IAS monitoring, or in fact biodiversity monitoring in general, the two ends of the information relevance dimension can be characterized by ***descriptive references*** to a species (for example “green beetle” or “green insect”) and ***direct references***

(for example “Emerald ash borer” or “Agrilus plannipennis”). Where the latter can be assumed to refer with high certainty to an IAS, the former may be “off-topic” or ambiguous, i.e. a “green beetle” could equally refer to a car and an insect.

The second dimension addresses information completeness or verifiability with regard to the available Tweet meta-data, which extends beyond the textual message content. This can for example include links (URLs) to other resources, media (images, videos), geo-location information (textual, geo-coordinates), references to other users or Twitter network size of the message author.

At one end of the message spectrum purely textual messages are found that can at best indicate ***topical*** foci, and at the other end messages with embedded high-quality images and attached exact geo-coordinates which have ***observational*** quality. In summary, one may thus expect descriptive references to IAS with no attached meta-data that will be impossible to verify as relevant, but also direct references with complete and high-quality meta-data that could approach the observational quality and nature of biodiversity observations obtained in deliberate traditional biodiversity monitoring efforts.

The characterization of the four information classes are presented in Figure 1 together with examples of analytical questions, thus equally highlighting the pursued data collection strategy and the main structure of the presented analysis. Throughout, this contribution will refer to these grid-classes by talking about ***descriptive or direct species references*** as well as ***topical or observational message sets***.

The presented results are largely based on a manual analysis of sample Tweet sets. With a view of potentially establishing social media as a supplementary data source in ecological monitoring and research, and given the complex and unstructured nature of social media data, a detailed manual analysis seems an essential first step in order to provide a solid foundation for an initial assessment as well as possible future research and practical applications.

2.1 Sample Invasive Alien Species

The *Global Invasive Species Database* compiled by the *Invasive Species Specialist Group (ISSG)* at the *IUCN* currently lists 891 invasive alien species (IUCN-ISSG, 2015). This study focuses on IAS with a direct or indirect impact on forest ecosystems, which account for

approximately half of the species in the ISSG database. Forest ecosystems provide a large number of recognized ecosystem services (Nasi et al., 2002) with IAS impacts ranging for example from negative influences on timber production to reduced recreational value, thus offering an interesting coverage of detectable ecological effects of invasions.

Furthermore, IAS were selected that ensured variation with respect to organism types, geographical regions, invasion vectors and invasion progress. Based on these and additional selection criteria (Supplementary Table 1) the list of sample species in Table 1 was compiled in collaboration with IAS experts on the *Aliens-L* mailing list and subsequently included in the data collection from Twitter.

Table 1. Invasive species selected as examples for this study. The selection was compiled in cooperation with IAS experts on the Aliens-L mailing list following selection criteria (**Supplementary Table 1** in the Appendix) ensuring a broad coverage of different regions, organism types, impacts, vectors and invasion history.

Species	Organism	Affected Region	Vector	Impact
European rabbit (Oryctolagus cuniculus)	Mammal	Australia	Deliberate introduction	Prevents regeneration of coastal sheoak forests
Emerald ash borer (Agrilus planipennis)	Insect	North America	Ash wood palettes	Lethal damage to ash trees
Oak processionary (Thaumetopoea processionea)	Insect	Europe	Habitat expansion	(Lethal) damage to oak forests; public health hazard
Asian long-horned beetle (Anoplophora glabripennis)	Insect	Europe, North America	Wood packaging	(Lethal) damage to broad range of deciduous tree species
Coqui frog (Eleutherodactylus coqui)	Amphibian	Hawaii	Accidental introduction	Noise impacts property and recreational value; affects nutrient cycle of tropical forests
Ash dieback (Hymenoscyphus pseudoalbidus)	Fungus	Europe	Unknown	(Lethal) damage to ash tree populations
Sudden oak death (Phytophthora ramorum)	Fungus	Europe, North America	Unknown	Affected oak trees die within 1-2 years
Horse-chestnut leaf miner (Cameraria ohridella)	Insect	Europe	Habitat expansion	Defoliation; reduced ornamental value and resistance to other pests
Pine processionary (Thaumetopoea pityocampa)	Insect	Europe	Habitat expansion	(Lethal) damage to pine forests; public health hazard
Eastern grey squirrel (Sciurus carolinensis)	Mammal	Europe	Deliberate introduction	Reduced biodiversity; tree bark damage
Rhododendron var.	Plant	Europe	Deliberate introduction	Inhibits regeneration of native species

2.2 Data collection using the Ecoveillance platform

Data collection targeted messages ("Tweets") published on the micro-blogging service Twitter, which are accessible through two public APIs: the Streaming API and the Search API, which was utilised here by retrieving Tweets that matched a set of predefined keywords (see Supplementary Table 2). Queries to the Twitter Search API return a rich set of meta-data in structured format, including not only the Tweet content, author and timestamp, but also information on an author's network, utilised source devices and applications, geo-location information, linked media or references to other Twitter users.

Known limitations of these APIs are that they provide only a small sample of all posted Tweets. In the case of the Twitter Streaming API approximately 1% of all Tweets published on Twitter are accessible in real-time, whereas the coverage of the Twitter Search API depends on a combination of a search term's frequency and popularity. Exact numbers are however not available and while the data collected is a - most likely significant - underrepresentation of the total number of matching Tweets, it is not possible to quantify this without resorting to complete, only commercially available, Twitter datasets.

A further limitation of the Twitter Search API is that it does not provide access to Tweets with matching keywords that are older than approximately one week. Data collection thus has to be run continuously. To that end a web-based system (denoted *Ecoveillance* (Daume, 2012)) was implemented which continuously obtains Tweets that match keywords related to the selected sample IAS. For most species, data has been collected continuously since May 2013, for one species (Oak processionary) since May 2012. Table 2 provides an overview of the abundance of Tweets for all monitored IAS by type of reference; the total number of obtained Tweets, the percentage of original Tweets (excluding "RTs" or "retweets"[1]) and ratios of Tweets with embedded media, geotags and a combination of both are shown.

[1] "RTs" or „retweets" are the terms commonly used for Twitter messages that are re-postings of another user's message. Twitter has a built-in mechanism that supports this messaging type.

Table 2. Data collection results for the observation period May 2013 to December 2014, both for direct and descriptive IAS references. Counts for the collected Tweets are provided as well as percentages of original messages ("Non-RTs"), Tweets with embedded media, attached geo-coordinates and a combination of both. In the case of 'Ash dieback' and 'Horse-chestnut leaf miner', references to the host species were used as descriptive references.

Monitored species	Direct references					Descriptive references				
	N	% Non-RTs	% with media	% geo-tagged	% geo & media	N	% Non-RTs	% with media	% geo-tagged	% geo & media
Rhododendron	*73,638*	49.65	20.15	3.28	0.75	-	-	-	-	-
Eastern grey squirrel	*50,048*	69.23	9.83	2.18	0.29	*179,801*	71.13	6.02	4.75	0.33
Emerald ash borer	*30,408*	71.17	5.10	1.26	0.20	*60,188*	76.96	13.14	3.16	0.29
Ash dieback	*15,458*	63.11	2.38	1.52	0.12	*33,518*	72.79	9.75	2.52	0.47
Sudden oak death	*5,306*	58.99	8.85	1.37	0.42	-	-	-	-	-
Oak processionary	*4,447*	74.57	2.77	0.78	0.15	*14,755*	48.53	8.42	3.23	0.35
Asian long-horn beetle	*3,494*	65.51	6.42	1.14	0.22	*8,089*	74.67	7.68	1.95	0.10
European rabbit	*2,612*	71.09	19.06	1.45	0.27	*45,525*	84.03	4.09	3.66	0.20
Coqui frog	*2,030*	63.74	8.35	4.17	0.62	*102,893*	66.97	7.08	3.69	0.36
Pine processionary	*1,615*	91.52	2.77	0.54	0.27	-	-	-	-	-
Horse-chestnut leaf miner	*1,273*	73.68	16.95	2.77	1.28	*101,417*	76.96	8.06	1.65	0.23

2.3 Message samples, analysis approach and information topology profile

This contribution focuses on three of the IAS monitored with the *Ecoveillance* platform, covering different characteristics with regard to introduction vectors, invasion history, organism type and impact:

- The **Oak processionary** (Thaumetopoea processionea) is a moth native to southern Europe that during the last decades has expanded its range into Central Europe (Germany, Belgium, Netherlands, Southern UK and Southern Scandinavia) either induced by climate change (Tubby and Webber, 2010) or recolonization (Groenen and Meurisse, 2012). The caterpillars assemble in distinctive large colonies, which can result in complete defoliation of oak trees (Forestry Commission, 2015) and have poisonous setae which can cause severe allergic reactions in humans (Gottschling and Meyer, 2006).

- The **Emerald ash borer** (Agrilus planipennis) is a small (length less than 1cm) bark-boring beetle indigenous to parts of China, Korea and Russia, was most likely introduced into North America accidentally through wood pallets and has been established in the United States since at least the early 1990s (Herms and McCullough, 2014). All North American species of ash are susceptible to the pest and infested trees typically die within 2-4 years (Herms and McCullough, 2014). It is dramatically expanding its range essentially threatening to wipe out ash trees in North America (Herms and McCullough, 2014), and may pose a similar threat to ash tree populations in Europe (Straw et al., 2013).
- The **Eastern grey squirrel** (Sciurus carolinensis) is indigenous to North America and an actively managed IAS in the UK. Considered a serious forest pest due to bark-stripping of trees, it has significant impact on biodiversity through the displacement of native red squirrels (Bruemmer et al., 2000). First introduced in the UK in 1876 it has been an established IAS in the UK for nearly 100 years (Bertolino, 2008; Bruemmer et al., 2000).

The Tweet analysis was pursued with the objective to provide information topologies for the distinct groups of information shown in Figure 1 in order to explore the "**topical"** and "**observational"** nature of the collected Twitter messages and directly address the posed research questions. Given the large volume of available messages (Table 2) and the need for a largely manual analysis of Tweets, appropriate samples had to be selected for each group of information.

Generally, "RTs" were excluded for all analysed datasets. Furthermore, samples were taken within date ranges that accounted for the lifecycles of the sample IAS, i.e. time periods when observations could be expected (Forestry Commission, 2015; Herms and McCullough, 2014; Thompson, 1977), which is of significance especially for *Oak processionary* and *Emerald ash borer*, to a lesser extent for *Grey squirrel*. Specifically:

- The analysis with a "**topical"** focus utilises Tweets **directly referencing** one of the three sample IAS and posted from May to September (type (C) in Figure 1), thus including the peak emergence or most active periods of all three sample species. The analysis of these datasets focuses on the topics covered by the Tweets (for example IAS impacts, remedies, monitoring) and the share and quality of observational references of the sample species.

- The analysis with an "**observational"** focus encompasses two types of sample datasets utilising Tweets with **direct references** or **descriptive references** of the sample species, limited to the key emergence or activity date ranges, and **containing or linking to images** (type (B) and (A) respectively in Figure 1). The analysis of these samples focuses entirely on the abundance and quality of verifiable conscious or opportunistic species observations.

The selected samples included a total of 2842 Tweets which were subjected to a detailed manual analysis. 2258 of those remained for incorporation to the final results, after removing a small number (62) of duplicates and a larger proportion (522) due to broken links or other stale information that would have been required at some point of the classification process. Figure 3 in the result section shows the frequency of all retrieved Tweets by species and reference type for the complete data collection period and highlights the species' key lifecycle periods that guided the sampling; a detailed listing of samples taken from this Tweet corpus is provided in the Appendix (Supplementary Table 3).

The results of the analysis are presented as an information topology (Figure 2) for each dataset. Figure 2 also illustrates the workflow of the Tweet analysis. Independent of the analysed dataset a Tweet was only subjected to further analysis if it referred to an IAS or a biological observation thus was "on-topic". Furthermore, it was also recorded which information (Tweet text, user information, linked URLs, media etc) contributed to this decision, which is of practical significance with regard to a future automation of such an analysis. The further analysis of all "on-topic" Tweets then followed the elements shown in Figure 2, with the exception that the 'Main message subjects' were only analysed in detail for 'non-observational' Tweets.

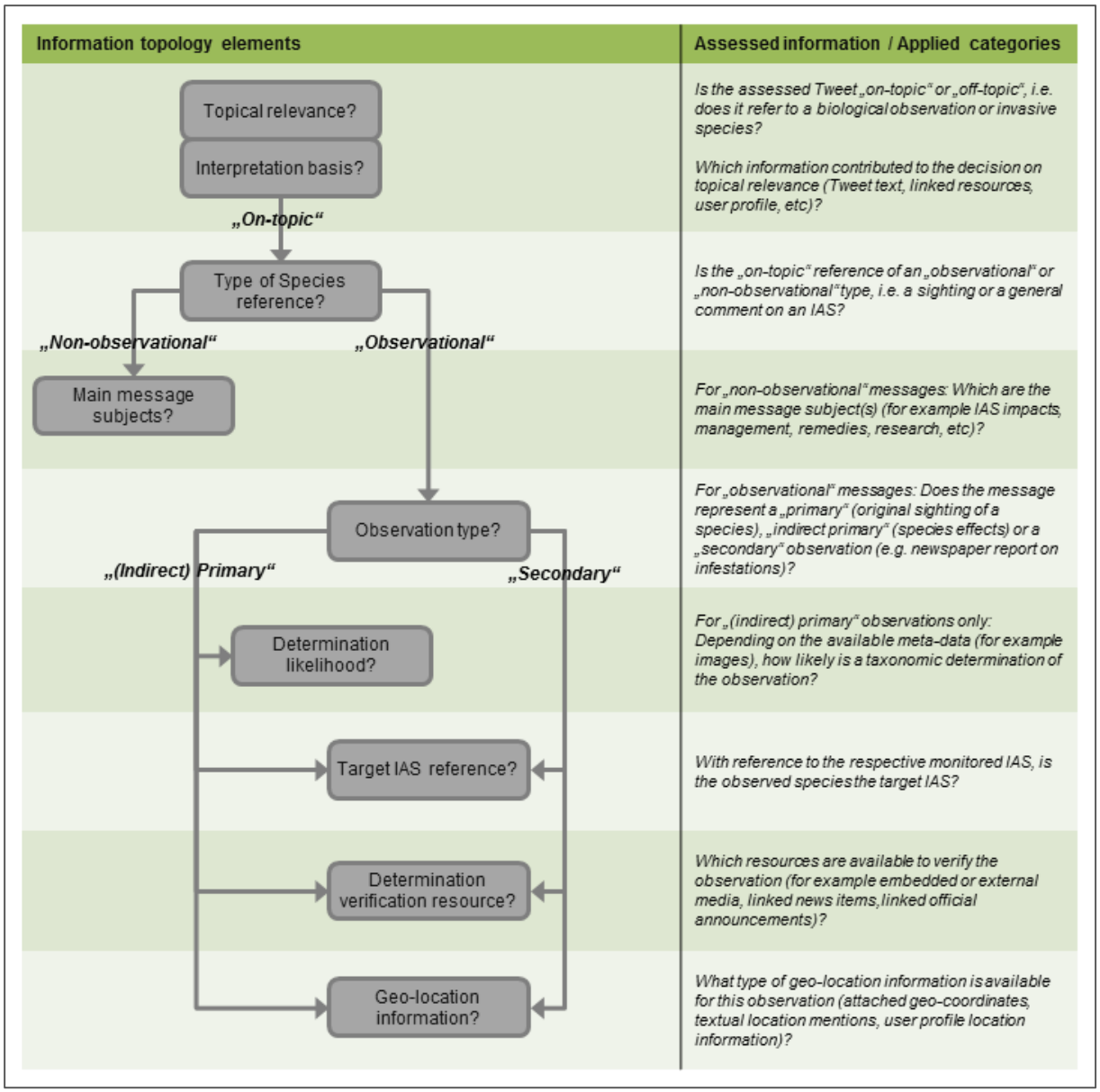

Figure 2. Tweet analysis flow and elements of the information topology obtained for the analysed Tweet samples. A detailed analysis of 'Main message subjects' was only applied to 'Non-observational' Tweets in the "topical" dataset; for 'primary observations' the likelihood for a taxonomic determination based on the available meta-data (specifically) images was assessed for all datasets. A detailed list and descriptions of the applied categories are provided in the Appendix (Supplementary **Table 4**).

The information topologies are complemented by an automatic analysis of additional Tweet meta-data, namely information on source devices and applications, geo-coordinates attached to the Tweets and location information provided in Twitter user profiles.

3 Results

3.1 Messaging frequencies

Figure 3 summarises the weekly Tweet frequencies (without so-called "RTs") for the three sample species, which show significant differences in the messaging patterns and abundance.

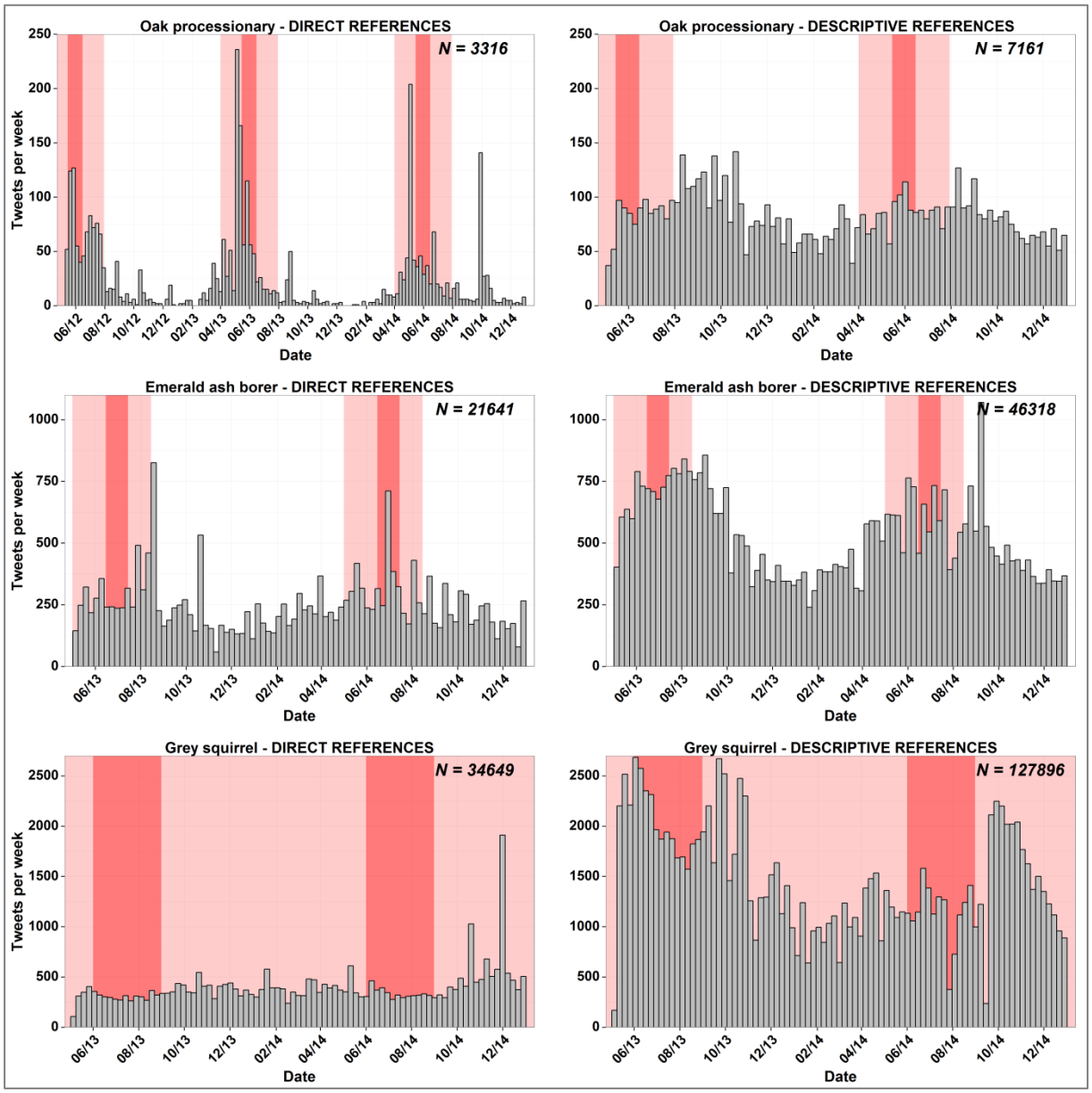

Figure 3. Frequency of original Tweets (excluding "RTs") with direct or potential descriptive references to the three sample species in the data collection period. The light-shaded regions indicate important life cycle and activity periods for each species (larvae development for Oak processionary, adult emergence for Emerald ash borer), the dark-shaded regions key lifecycle phases (Oak processionary caterpillars with poisonous setae building distinctive colonies, and peak emergence of Emerald ash borer (Forestry Commission, 2015; Herms and McCullough, 2014)). Grey squirrel is active and can be observed throughout the year with a peak activity period in the summer months (Thompson, 1977).

In general, messages matching the selected keywords are posted throughout the year and Tweets matching descriptive terms are expectedly more abundant, which can be attributed to their broader coverage, but also larger ambiguity. Tweets matching descriptive terms will thus also contain more "off-topic" Tweets. This observation is confirmed by the analysis in sections 3.2 and 3.3.

The frequencies are in line with known characteristics of the sample species. The modest Tweet numbers for Oak processionary moth (OPM) match the fact that it is a very specific localized seasonal threat that receives attention especially for its health impact. The Emerald ash borer (EAB) is a serious widespread challenge in the U.S., the communications on EAB are much richer and heterogeneous (see also Figure 5), but the species is more difficult to recognize. Eastern grey squirrels (EGS) on the other hand are common, widespread and easy to recognize.

Tweets with direct species references show a good fit with the annual lifecycles. This is very distinct for OPM and indicative for EAB. The very constant messaging pattern for EGS fits the year-round observability of the species. The less apparent fit for descriptive terms indicates the noise in the data. Terms like "hairy caterpillar", "green beetle" and even the more specific "squirrel" are ambiguous and will thus appear not just during the lifecycle period. In all cases it has to be noted that mentions could be found at any time, either because the subject is discussed non-observationally or the observations are happening in geographic regions with different lifecycles.

3.2 Topical Tweets analysis

The results in this section focus on the analysis of Tweets with **direct references** to the sample species **posted in the period May to September 2013**. The purpose of these Tweet sets was to obtain an estimate of the type of all content directly referencing a species and also getting an indication of the abundance of the different message types.

3.2.1 Topical relevance

Following the approach outlined in Figure 2 the sample sets for each of the three sample species were first assessed with regard to their topical relevance (Table 3). Here a Tweet was considered "on-topic" if the matching keywords discussed the species or represented an

observation. It was considered “off-topic” if matching keywords were used out of context or with a different meaning (for example “green beetle” referring to a car).

Table 3. Number of "On-topic", "Off-topic" and “Inconclusive” Tweets with direct references to the three sample species and the required Tweet information items that contributed to the determination of the topical relevance. The interpretation basis percentages in each column add up to more than 100% as multiple information items may have contributed to the decision on topical relevance. A Tweet is considered “on-topic” if it refers to a sample IAS.

		Oak processionary	Emerald ash borer	Eastern grey squirrel
TOPICAL RELEVANCE	N (Tweets)	199	221	190
	% On-topic	*96.0*	*91.9*	*58.9*
	% Off-topic	-	*6.8*	*27.4*
	% Inconclusive	*4.0*	*1.4*	*13.7*
INTERPRETATION BASIS	% Text	*94.2*	*97.0*	*93.8*
	% Links	*16.8*	*17.2*	*8.9*
	% Embedded media	*0.5*	*1.0*	*5.4*
	% User profile	-	-	*0.9*
	% Conversations	*2.6*	*1.0*	*13.4*
	% External media	*0.5*	*0.5*	*5.4*

Table 3 shows that keywords representing direct mentions of OPM and EAB seem to guarantee a high proportion of “on-topic” Tweets, which is not the case for EGS where the combination of the terms “grey” and “squirrel” is apparently more ambiguous.

The results of the interpretation basis assessment for all three samples indicate that the textual content of the Tweets seems to provide sufficient information to decide on the topical relevance of a message. This is quite significant since it has practical implications with regard to future operational uses of this information; if textual content is sufficient to filter relevant Tweets, standard classification algorithms can be applied to support automatic filtering (Liu, 2006; Yin et al., 2012). This also applies to conversations, user profiles and external links which can be treated as textual data, are accessible via the collected Tweets and could thus be utilised in automatic classification routines.

3.2.2 *Message types*

In the next classification step all "on-topic" Tweets were categorised according to their message type. As Figure 4 illustrates this turned out to be primarily a decision on whether Tweets were "observational" or "non-observational", thus if the message indicated some form of species sighting (direct or indirect), or was purely discussing the species in general (for example mentioning species facts, impacts or research related to the species). Tweets could also be coded as belonging to multiple categories; details on all applied categories are provided in Supplementary Table 4.

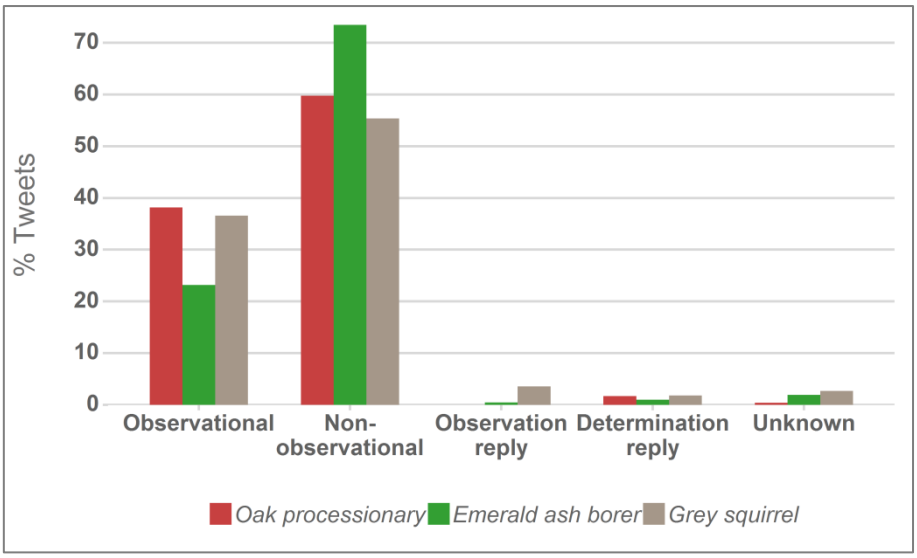

Figure 4. Tweet message type by sample species for all "on-topic" Tweets in the period May to September 2013. The major distinction is between "Observational" and "Non-observational" messages with regard to the respective target species.

The majority of messages are non-observational for all three sample species, with 73% of non-observational messages for EAB indicating a particularly active communication on this subject.

3.2.3 *Non-observational message subjects*

All "non-observational" Tweets were subjected to a detailed analysis of the message subjects, focusing on topics related to the invasive nature of the three sample species (see

Supplementary Table 4 for details). Figure 5 shows that distinctive thematic profiles emerge for the different species.

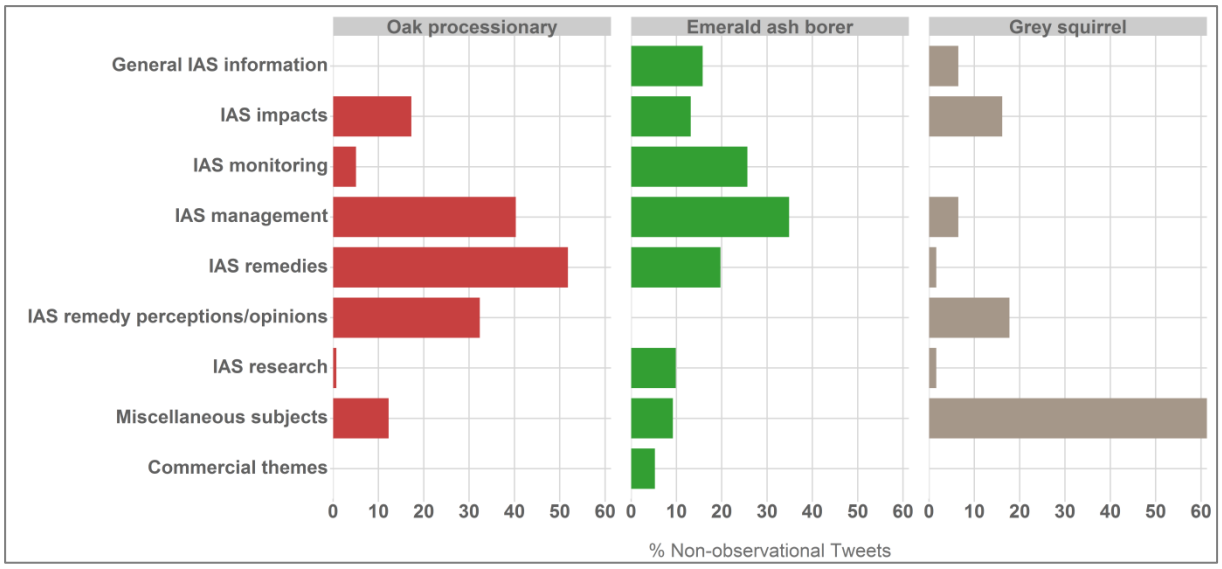

Figure 5. Main message subjects by sample species for all "Non-observational" Tweets in the period May to September 2013. Distinctive subject profiles emerge for the three species. Percentages per species add up to more than 100% as a Tweet may address more than one subject.

IAS impacts (on for example biodiversity, host species, human health) are a similarly frequent subject for all three species, but the topical emphasis shows some notable differences for all other subject areas.

Communications on Emerald ash borer cover for example all aspects of the IAS thematic range: general (educational) information and species facts are shared, the management, remedies and monitoring of EAB play an important role and even research subjects are covered. A significant contrast to Oak processionary and Eastern grey squirrel is the notable lack of perceptions and opinions regarding remedies. There appears to be a consensus on the invasiveness of EAB reflected by the analysed Tweets, whereas for OPM and EGS many opinionated comments - primarily critical statements about the respective IAS management (culling of squirrels and insecticide spraying for OPM) - can be found. For OPM generally "IAS remedies" are the primary subject, which also fits the frequency pattern in Figure 3 where clear messaging peaks can be observed at the point in the lifecycle where remedial action is commonly taken.

Communication about Eastern grey squirrel are however very generic. While impacts are mentioned (especially on biodiversity) and criticism is voiced about the culling of squirrels, the majority of messages do not fall into any of the IAS themes. In fact, messages are dominated by fairly banal content ranging from comments on unusual sightings ("white squirrel"), statements about the "cuteness" of squirrels or humorous statements. This fits both the fact that EGS is a common species in its native range and has been an established species in its invasive range for nearly a century, but also that apparently neither the impact of the species itself nor the remedial actions are perceived as significant.

Notable is also the lack of messages with a commercial background for OPM and EGS. Several messages on EAB were however related to the treatment of ash trees or similar commercial advice and offers. This should be viewed as significant since the degree of commercial importance may be directly related to the spread, impact and invasion progress of a species which is considered dramatic in the case of EAB (Herms and McCullough, 2014).

While this is a snapshot of Twitter communications on the three sample species in a limited time period it is fair to assume that similarly distinctive profiles will emerge when considering the thematic dynamics over much longer time periods. These, might hence provide a good reflection of the progress and perception of an invasive species over time. Whether generic thematic progressions can be found and used for early warnings cannot be answered at this stage and will require long-term sampling and further investigation. However, this snapshot confirms that the information can provide valuable reflections on the public perception of specific IAS which can be of value with regard to IAS management, communication and education. This is of particular interest since mature methods enabling an automated topical analysis (Blei, 2012) of large sets of messages are available.

3.3 Observational Tweets analysis

The results in this section focus on the analysis of Tweets with **direct and descriptive references** to the sample species **posted in the annual key lifecycle periods** (see Figure 3) and **containing or linking to media** (typically images). The objective for these constrained samples was to assess the nature of the potential species observations that can be obtained by mining Twitter. They represent a directly applicable output to IAS management and monitoring, which is likely to apply to other social media sources as well.

Potential observations with direct references to an IAS are of immediate interest, since Tweets that match specific species names are with large likelihood relevant observations. Observations in the form of Tweets matching descriptive references to a targeted species are of particular value as they could potentially expand available observations significantly - few people may know or can name a specific invasive species, but will note and be able to describe a species' distinctive features. This approach can however be expected to introduce a lot of noise in the collected data. The three sample IAS represent a complete range of descriptive terms, including very broad and potentially ambiguous keywords ("green beetle"), more specific descriptive features ("hairy caterpillar") as well as terms falling short of an actual species name ("squirrel").

3.3.1 Topical relevance

Table 4 shows the results for the assessment of topical relevance for the analysed Tweet samples. A Tweet was considered "on-topic" when it referred to any of the IAS or represented a biodiversity observation in general.

Table 4. Topical relevance and interpretation basis for analysed Tweets with direct or descriptive references to the three sample species, posted in the annual key lifecycle periods and containing or linking to media (images). The interpretation basis percentages in each column add up to more than 100% as multiple information items may have contributed to the decision on topical relevance. A Tweet is considered "on-topic" when it represents a biodiversity observation, even if this observation is not of the sample IAS.

		DIRECT REFERENCES			DESCRIPTIVE REFERENCES		
		Oak processionary	Emerald ash borer	Eastern grey squirrel	Oak processionary	Emerald ash borer	Eastern grey squirrel
TOPICAL RELEVANCE	N (Tweets)	85	314	294	331	240	384
	% On-topic	*96.5*	*82.2*	*72.8*	*75.5*	*71.2*	*74.2*
	% Off-topic	*3.5*	*17.2*	*26.2*	*23.6*	*27.9*	*21.9*
	% Inconclusive	-	*0.3*	*0.7*	*0.9*	*0.4*	*3.9*
INTERPRETATION BASIS	% Text	*98.8*	*95.3*	*96.3*	*66.8*	*81.3*	*78.2*
	% Links	*3.7*	*15.9*	*2.3*	-	*1.2*	*0.7*
	% Embedded media	*18.3*	*10.1*	*30.8*	*34.8*	*25.1*	*51.6*
	% User profile	-	*0.4*	-	-	-	-
	% Conversations	-	-	-	*0.4*	-	-
	% External media	*3.7*	*1.6*	*15.4*	*27.6*	*34.5*	*14.4*

Again, direct references (using the species names as a search term) seem to result in a large proportion of "on-topic" Tweets, while samples matching descriptive references have lower shares of "on-topic" Tweets. Furthermore, it is notable that in comparison to the results presented in Table 3 images contribute significantly in determining the topical relevance of Tweets. While this may be expected given that Tweets in this sample were constrained to posts containing or linking to images, this has important practical implications, as it means that in terms of building operational systems utilising these Tweets, automatic filtering based on textual approaches will grab a smaller subset of all potentially relevant posts, automation routines would thus have to extend to images, or manual interventions are required.

3.3.2 Message types

The comparison of Tweet message types (Figure 6) shows that constraining the sample sets on Tweets with embedded or linked media does result in larger shares of "observational" Tweets, in the case of descriptive references this is very significant. Tweets with embedded or linked images and direct references to the species are not exclusively observational, but for all sample species the shares of observational Tweets are higher compared to Figure 4.

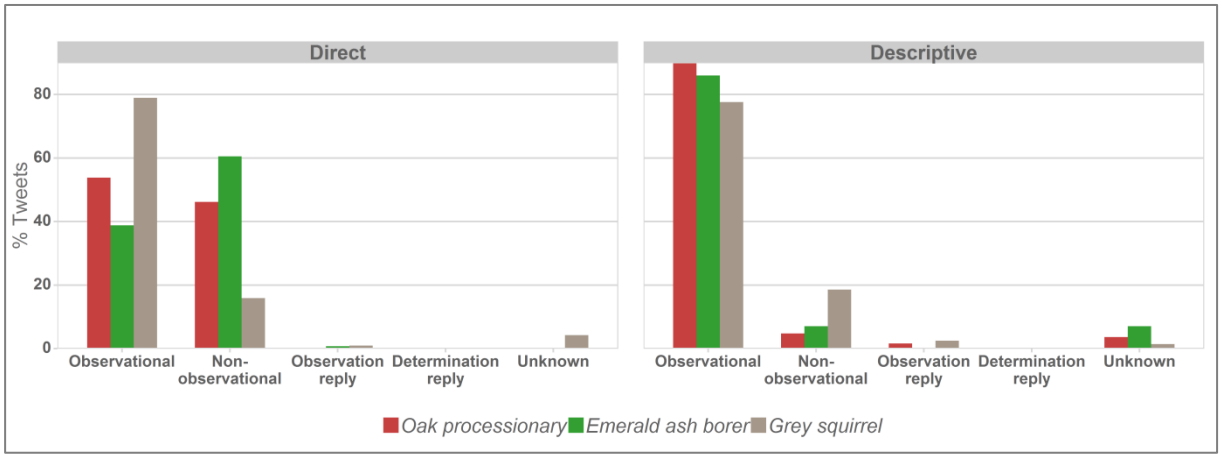

Figure 6. Comparison of Tweet message type by sample species for Tweets with direct and descriptive species references, posted during the sample species' key lifecycle periods and containing or linking to media (images). The major distinction is between "Observational" and "Non-observational" messages with regard to the respective target IAS.

3.3.3 Observational message profile

Figure 7 provides a complete information profile of the observational Tweets obtained using terms that either represented direct or potential descriptive references to the target species. In terms of evaluating, comparing and utilising these biodiversity observations the type of observation, the likelihood of a taxonomic determination of primary species observations given the Tweet information (here specifically the embedded or linked media), the resources available to verify the determination of an observation and finally the type and availability of geo-location information are of interest.

Throughout Figure 7 the share of Tweets in each category that could be confirmed as observations of the specific targeted IAS are indicated with black markers and labels; if those are omitted no Tweet in a given category could be confirmed as an observation of the targeted IAS.

Figure 7. Information profile of "observational" Tweets with embedded or linked media and direct or descriptive references to the target IAS. The information profile specifies separately for direct and descriptive references (A) the type of observation, (B) for primary observations how likely a species determination is considering the quality and availability of the Tweet data, (C) for all observations which resources are available for verification or determination of an observation and (D) what type of geo-location information is available. The black markers and labels per bar indicate which percentage of all Tweets per species could be identified as actually relating to the target species (for example based on the available

information 22.4% of all observational Tweets directly referencing *Oak processionary* were confirmed as primary observations of this species). Percentages can add up to more than 100% as a Tweet can belong to multiple classes.

3.3.4 Observation types

Observational Tweets with descriptive terms are almost exclusively "primary" (Figure 7(A)) that is direct sightings at the source, which are reported by the observer and will typically be novel sightings. Observation types for Tweets with direct references differ significantly: the majority (88.2%) of Grey squirrel observations are classed as primary, whereas only 10.1% of observational Tweets mentioning Emerald ash borer are primary and 44.9% of Oak processionary mentions. For a start this will relate to recognisability. EAB is a small insect which briefly emerges to lay eggs, whereas OPM build distinctive notable colonies and EGS is a common species in urban spaces, which is easy to spot and recognize. This is also reflected by the large share (36.7%) of "indirect primary" observations of EAB, the effects of EAB on the host species (bare branches, peeling bark, feeding patterns on the bark) are easier and more likely to recognize than the species itself, which is not the case for EGS or OPM.

However, the large share of "secondary" observations (mostly news reports and public announcements) of both OPM and EAB underlines that in contrast to a long established invasive species like EGS these threats receive public attention and are perceived and discussed as notable current problems.

A key question for all these observations is how many can be verified as observations of the selected target IAS. With regard to primary observations, the available resources (images in the case of primary observations) can be used to verify the observations and the noted share of actual target species observations (indicated with labels for each observation type in Figure 7(A)) is thus a reflection of both the relevance of the Tweets for IAS monitoring and the quality of the available verification resources.

For Tweets with direct species references EGS does not only have the largest share of primary observations, but also for verified primary sightings: 88.2% of all observational Tweets directly referencing Eastern grey squirrel are primary observations, but only 67.5% of all those Tweets mentioning Eastern grey squirrel could be confirmed as primary observations of this species. The remainder (20.7%) were thus primary observations either

of another species or where the available information did not suffice to determine the species.

This large share for EGS is followed by OPM where 22.4% of all observational Tweets are primary observations of the species. Primary sightings of EAB are rare anyway and only 2.8% of all observational Tweets referring directly to EAB could be confirmed as primary observations of this species. Potential indirect primary observations of EAB are more common, only a small share (5.5%) of all observational EAB Tweets could however be verified as the target species. These numbers correspond again directly with the phenology, commonness and lifecycle characteristics of the three sample species; unsurprisingly, common and easy to identify species or species with distinctive features are more likely to be observed and verified.

Secondary observations in Tweets with direct species references could almost exclusively be verified as target species sightings. The assessment was based on the authoritativeness of the source, thus if the message was based on a confirmed news outlet or official body (for example the US Forest Service or the UK Forestry Commission), it was assumed to be a trustworthy observation of the target species.

Finally, for observational Tweets matching descriptive references of the target species, only descriptive references of EGS using the very specific term "squirrel" could be verified as target species sightings (37.4%). While terms like "green beetle" or "hairy caterpillar" are apparently good keywords to return general primary biodiversity observations none of those were of the target species, which may not be surprising given that insects are known to be the most diverse group of animals. This has practical implications however when considering alternative data collection approaches for operational systems. The options are to either use broader terms ("insect", "caterpillar") and focus on efficient filtering routines, or to use more specific phrases ("green iridescent beetle") or include contextually related terms ("tree", "leaves") and accept that potentially valuable observations are not detected since these terms are less likely to be used in casual, short messages posted on Twitter.

3.3.5 Determination likelihood of primary observations

The determination of primary observations was largely based on the quality of the available embedded or linked media in the analysed Tweets. In cases where no media were available,

could not be used for determination or were of very poor quality the determination likelihood was coded as “Impossible or very unlikely”. Whereas media instances of high-quality that allowed an unambiguous species determination were classed as “Certain” (Figure 7(B)). In addition other information was included in the determination. An example are observations of white or albino squirrels which are typically melanistic variations of EGS, thus even images of poorer quality would suffice for a determination.

Generally for both direct and descriptive references approximately half or more of the available media are of average (“Likely”) or better quality thus supporting species identification. In the case of direct references to EGS, high-quality images accounted even for 53% allowing a reliable determination of the species. A wide quality spread has to be noted though. This has practical implications and hints at limited applicability of automated recognition methods and thus indicates that the utilisation of these observations will to a significant degree involve manual interventions.

3.3.6 Observation verification resources

Figure 7(C) outlines the distribution of available resources for the verification of observations which include embedded and external media, news links and official announcements thus documents from public bodies (like the US Forest Service).

Given the initial filtering of the sample, it is not unexpected that embedded images or externally linked media account for a large share of the verification resources, but they could not be applied for verification of observations in all instances. In the case of direct species references embedded or linked media together represented between 50.5% (EAB) and 88.8% (EGS) of verification resources, whereas news links were noted as the main verification resource for secondary observations of EAB (36.7%) and OPM (30.6%). For descriptive references nearly all images could be used as a verification resource and represented an exclusive verification resource.

In summary, it can be noted that the majority of observational Tweets provide some type of resource to verify an observation. Mostly, these are embedded or external media, thus again implying a certain degree of manual intervention when utilising these observations in operational systems for IAS monitoring.

3.3.7 Geo-location information

The availability of reliable geo-location information is essential for biodiversity observations. Figure 7(D) summarises the available types of geolocation information in the analysed observational Tweets. They include precise geo-coordinates (included in the Tweet meta-data), location mentions in the Tweet text, locations provided by Tweet authors in their Twitter user profiles and indirect location references (such as *"my garden"* or locations that could be inferred from mentions of other Twitter users (*"today at @Caltech"*)).

With the exception of location mentions in the Tweet messages both direct and descriptive references show similar patterns. The large shares of textual location references for EAB (63.3%) and OPM (46.9%) can be directly linked to the large number of secondary observations which are typically of a form like *"Ash Borer found in Jefferson County"*.

The share of users providing resolvable[2] locations in their profile ranges from 46.9% (EGS) to 69.7% (EAB). This information is however showing a coarse granularity (location details range from 8% at country level to 41.2% of users providing a city or more specific location), and generally has to be taken with caution as it cannot be verified, could get of sync and may not correspond to the location from which an observation was reported.

Reliable geo-location information is available in the form of geo-tags representing exact geo-coordinates, typically attached to the Tweet by GPS-enabled devices from which a Tweet was posted. Shares in the analysed samples range from 1.8% to 4.1%; this is in line with averages reported in comparable studies (Croitoru et al., 2014).

There is thus a clear shortage of exact geo-location information associated with this data. However, an exploration of the source devices and location-sharing settings (obtained from the Tweet meta-data) reveals that on average 37.9% of all analysed Tweets were posted from mobile devices or applications such as Instagram (21.3%) geared towards mobile use. Furthermore, 35.2% of users posting Tweets had their profiles geo-enabled, which is a Twitter profile setting that has to be deliberately enabled by Twitter users in order to attach geo-location information to their posts. The mismatch between the large share of geo-enabled Twitter profiles and the small proportion of Tweets with geo-coordinates can most likely be explained by more restrictive settings on the devices used for posting Tweets. Hence, assuming these settings were in sync, the prerequisites for a large pool of

[2] Locations provided in analyzed Twitter user profiles were resolved using Google geo-referencing via the R package *ggmap* (Kahle and Wickham, 2013).

biodiversity observations with precise geo-coordinates are given. In relation to this it is also worth noting that user contact details in form of Twitter accounts are part of the obtained Tweet data and provide the means to follow up with users directly on the reported observation details.

4 Discussion

4.1 Domain-relevance, data abundance and data characteristics

The results presented in this contribution show that a large pool of data on specific invasive species, covering both diverse IAS topics and observations, can be easily obtained via the social media channel Twitter. When extrapolating the observational results for Eastern grey squirrel they even appear to be on par in annual abundance with observational species data stored in repositories such as the Global Biodiversity Information Facility (GBIF, 2015), exceeding in fact the available shares of observations with verifiable media, not matching however the large proportion of geo-referenced records.

Direct and potential descriptive mentions of invasive alien species in Twitter messages are abundant but appear to correlate with the recognisability and commonness of a species. Thus, species that show advanced states in their invasion of an ecosystem will be covered more heavily in Twitter communications, which suggests that this type of social media data primarily mirrors known facts and dynamics on invasion progress and may provide limited contributions to the early detection of IAS.

However, while the presented results, partly due to the selected sample species, cannot provide examples of early detections with regard to initial introductions of an IAS, they do with regard to new invasion seeds and range expansions. Tweets on Oak processionary included for example primary observations of the species in private gardens, which are commonly not routinely monitored in official forest monitoring programs (FVA-BW, 2012; NW-FVA, 2012) and even singular observations can contribute to halting the further spread of a species.

In summary, the results confirm that Twitter is a rich source of general biodiversity observations, and particularly opportunistic observations, gathered via descriptive keywords and contributed without knowledge of the species, hold great potential in

supplementing biodiversity monitoring, even if this will primarily apply to notable or easily recognizable species.

4.2 Contributions to ecological monitoring

From an IAS management perspective the observed Twitter communications are of interest even for established invasions. Opinions, criticisms and the general public perception of IAS remedies are a particularly good example. Public support is critical for effective IAS management (Bremner and Park, 2007) and social media observations offer a direct, efficient and - according to the presented results - representative way to investigate public perceptions. Furthermore, Twitter (as other social media channels) cannot only be utilized as a data source but also as a communication channel, and thus provides the means for IAS managers to engage in a targeted manner in these social media communications for educational purposes, requests for support or general information distribution via identified communication hubs and communities.

The results also show that the collected information will be highly sensitive to the employed search keywords. This is indicated by the results for descriptive search terms, which only for the rather specific “squirrel” returned actual observations of the targeted IAS. Using very specific terms will exclude potentially relevant observations and broader terms will introduce significant noise to the data. Importantly, the results thus also highlight that there is no generic approach to using social media sources in environmental monitoring. Instead, tailored information retrieval efforts, based on a good understanding of the species features, lifecycles and also the interest of casual observers (“What attracts people’s attention?”) is required. Based on the presented results, operational systems aiming to utilize so-called “Big Data” methods (combining for example natural language processing (Miner et al., 2012) with unsupervised machine learning techniques (Liu, 2006)) deserve exploration, but have to be paired with a thorough domain understanding and clear questions for analyzing this data source.

Those questions will vary. Where Twitter data may not be expected to deliver many primary observations for species like Emerald ash borer, secondary observations for these difficult to observe but widespread species can be found, but may need expert analysis and follow-up. Furthermore, EAB is a prime example exhibiting frequent communications on the topic,

and the related channels and networks can be and are utilized in raising awareness to prevent further spread of an invasion or calling for monitoring support.

For threats like Oak processionary similar opportunities are available. In addition, more primary observations can be expected and related threats may “accidentally” also be captured (such as Pine processionary moth). For localized and seasonal threats like OPM, Twitter mining also offers a convenient way to obtain and summarize real-time information for bigger geographical regions, in particular where regular monitoring efforts are fragmented and environmental managers have only access to information of local provenance (Daume et al., 2014).

For common and easily recognized species like the Eastern grey squirrel many primary observations can be expected via social media sources, and this will equally apply to other invasive species which may have established populations but nonetheless need to be monitored as part of general IAS management efforts (examples include Asian carp, Chinese mitten crab or Asian lady beetle).

Finally, that an invasive species (EGS) is essentially not perceived as such in public communications is an interesting result in itself. Monitoring social media communications on any invasive species could thus also be used as a measure to assess the success of communication and education efforts in IAS management.

4.3 Practical challenges in utilizing this information

In general, this study indicates that there are practical rather than conceptual hurdles in using social media data in operational systems that could supplement traditional ecological monitoring.

Operational approaches will have to extend the content coverage and the degree of automation. Content coverage applies to the utilized sources and the representativeness of the obtained samples. As indicated, Twitter was chosen as a representative social media example that shares properties with other social media channels which could hence equally be included in any digital ecosystem surveillance.

Automation routines need to be directed at filtering relevant messages, but also to obtain geo-location information. The scarcity of high-resolution geo-information was noted for the

analyzed samples. While this may limit the number of immediately usable geo-referenced observations, it does not preclude the usefulness of this information. Furthermore, novel approaches utilizing public Twitter user content (Cheng et al., 2010) and network information (Compton et al., 2014), enable the automatic provisioning of Twitter user locations given sufficient resources.

With regard to automated message filtering the results indicate that the textual message data will to a large degree suffice to decide on the topical relevance of Tweets, hence existing toolsets for text classification (Liu, 2009) can be applied. With regard to observational data, manual interventions - not least for the species identification - are however still required, thus suggesting combined manual and automated analysis approaches as demonstrated by related public health monitoring systems (Mykhalovskiy and Weir, 2006). A promising approach would be to engage citizen science communities with this "crowdsourced" data in order to verify and map observations, as has been successfully demonstrated by projects dealing with related biodiversity information (Hill et al., 2012).

5 Conclusions

Twitter proves to be a rich source of information on invasive alien species, ranging from primary species observations to insights in prevalent IAS topics and their public perception. The presented results show that the abundance and the features of this information thus merit dedicated efforts to advance the utilisation and integration with existing ecological information sources.

While operational systems could be automated to a large extent, the presented study also suggests that completely generic approaches will probably have limited applications. Instead, as in any successful ecological monitoring program, a good understanding of the monitored subject and targeted questions must guide the provisioning of the monitoring effort (Lindenmayer and Likens, 2010), here the choice of keywords, social media channels and filtering of the obtained data. The presented information topology can serve as a helpful template to compare the usefulness of social media in ecological monitoring when applied to other species or social media sources.

Finally, particular efforts ought to be directed towards the definition of processing workflows for observational data that includes mappings to formal data structures and integrates with existing biodiversity information sources. Manual interventions in utilising observational data in particular will however be required in many cases, and engaging "citizen science crowds" with this "crowdsourced" data could be a promising route.

6 Acknowledgements

I would like to thank Melanie Josefsson at the *Swedish Environmental Protection Agency* for helpful information on invasive alien species, suggesting examples and connecting me to many other IAS researchers. Furthermore, I am grateful for the help received from all IAS experts on the *Aliens-L mailing list* that allowed me to compile a representative list of IAS for this study.

The idea for this contribution took shape while I was staying as a guest researcher at the Stockholm Resilience Centre (SRC), which has generously supported me with access to an excellent research infrastructure. I am particularly indebted to Victor Galaz who supported me throughout this work.

Furthermore, this work was conducted in parallel to an employment at the Swedish Museum of Natural History and many conversations with my colleagues Fredrik Ronquist, Anders Telenius and Kevin Holston at the Department of Biodiversity Informatics contributed to my understanding of mining and managing biodiversity information in general.

Finally, I would like to thank Klaus von Gadow, Matthias Albert and Victor Galaz for their valuable feedback on the manuscript.

7 References

Barve, V., 2014. Discovering and developing primary biodiversity data from social networking sites: A novel approach. Ecol. Inform. 24, 194–199. doi:10.1016/j.ecoinf.2014.08.008

Bertolino, S., 2008. Introduction of the American grey squirrel (Sciurus carolinensis) in Europe: a case study in biological invasion. Curr. Sci. 95, 903–906.

Biggs, R., Carpenter, S.R., Brock, W.A., 2009. Turning back from the brink: detecting an impending regime shift in time to avert it. Proc. Natl. Acad. Sci. U. S. A. 106, 826–31. doi:10.1073/pnas.0811729106

Bird, T.J., Bates, A.E., Lefcheck, J.S., Hill, N.A., Thomson, R.J., Edgar, G.J., Stuart-Smith, R.D., Wotherspoon, S., Krkosek, M., Stuart-Smith, J.F., Pecl, G.T., Barrett, N., Frusher, S., 2013. Statistical solutions for error and bias in global citizen science datasets. Biol. Conserv.

Blei, D., 2012. Probabilistic Topic Models. Commun. ACM 55, 77–84. doi:10.1109/MSP.2010.938079

Bremner, A., Park, K., 2007. Public attitudes to the management of invasive non-native species in Scotland. Biol. Conserv. 139, 306–314. doi:10.1016/j.biocon.2007.07.005

Bruemmer, C., Lurz, P., Larsen, K., Gurnell, J., 2000. Impacts and Management of the Alien Eastern Gray Squirrel in Great Britain and Italy: Lessons for British Columbia. Management 1, 15–19.

Cheng, Z., Caverlee, J., Lee, K., 2010. You are where you tweet: a content-based approach to geo-locating twitter users, in: Proceedings of the 19th ACM International Conference on Information and Knowledge Management - CIKM '10. ACM Press, New York, New York, USA, pp. 759–768. doi:10.1145/1871437.1871535

Compton, R., Jurgens, D., Allen, D., 2014. Geotagging one hundred million Twitter accounts with total variation minimization, in: 2014 IEEE International Conference on Big Data (Big Data). IEEE, pp. 393–401. doi:10.1109/BigData.2014.7004256

Conover, M.D., Goncalves, B., Ratkiewicz, J., Flammini, A., Menczer, F., 2011. Predicting the Political Alignment of Twitter Users, in: 2011 IEEE Third Int'l Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third Int'l Conference on Social Computing. IEEE, pp. 192–199.

Croitoru, A., Crooks, A., Radzikowski, J., Stefanidis, A., Vatsavai, R.R., Wayant, N., 2014. Geoinformatics and Social Media: A New Big Data Challenge, in: Karimi, H.A. (Ed.), Big Data Techniques and Technologies in Geoinformatics. CRC Press, Boca Raton, FL, pp. 207–232.

Crooks, A., Croitoru, A., Stefanidis, A., Radzikowski, J., 2013. #Earthquake: Twitter as a Distributed Sensor System. Trans. GIS 17, 124–147. doi:10.1111/j.1467-9671.2012.01359.x

Crowl, T.A., Crist, T.O., Parmenter, R.R., Belovsky, G., Lugo, A.E., 2008. The spread of invasive species and infectious disease as drivers of ecosystem change. Front. Ecol. Environ. 6, 238–246. doi:10.1890/070151

Daume, S., 2012. Ecoveillance (Social media analysis web platform).

Daume, S., Albert, M., von Gadow, K., 2014. Forest monitoring and social media – Complementary data sources for ecosystem surveillance? For. Ecol. Manage. 316, 9–20. doi:10.1016/j.foreco.2013.09.004

De Longueville, B., Smith, R.S., Luraschi, G., 2009. "OMG, from here, I can see the flames!": a use case mining Location Based Social Networks to acquire spatio-temporal data on forest fires, in: Proceedings of the 2009 International Workshop on Location Based Social Networks - LBSN '09. ACM Press, New York, New York, USA, pp. 73–80. doi:10.1145/1629890.1629907

European Commission, 2011. European Commission - Environment - Nature & Biodiversity - Invasive Alien Species (http://ec.europa.eu/environment/nature/invasivealien) [WWW Document]. URL http://ec.europa.eu/environment/nature/invasivealien (accessed 4.26.11).

Forestry Commission, G., 2015. Forestry Commission - Pests & Diseases - Oak Processionary Moth [WWW Document]. URL http://www.forestry.gov.uk/oakprocessionarymoth (accessed 5.30.15).

FVA-BW, 2012. Aktueller Hinweis zum Eichenprozessionsspinner. Stand: 17.09.2012, Forstliche Versuchs- und Forschungsanstalt Baden-Württemberg (Germany).

Galaz, V., Crona, B., Daw, T., Bodin, Ö., Nyström, M., Olsson, P., 2010. Can web crawlers revolutionize ecological monitoring? Front. Ecol. Environ. 8, 99–104. doi:10.1890/070204

Garcia Esparza, S., O'Mahony, M.P., Smyth, B., 2012. Mining the real-time web: A novel approach to product recommendation. Knowledge-Based Syst. 29, 3–11. doi:10.1016/j.knosys.2011.07.007

GBIF, 2015. GBIF Portal [WWW Document]. URL http://www.gbif.org/ (accessed 6.22.15).

Gottschling, S., Meyer, S., 2006. An epidemic airborne disease caused by the oak processionary caterpillar. Pediatr. Dermatol. 23, 64–6. doi:10.1111/j.1525-1470.2006.00173.x

Graham, J., Newman, G., Kumar, S., Jarnevich, C., Young, N., Crall, A., Stohlgren, T.J., Evangelista, P., 2010. Bringing Modeling to the Masses: A Web Based System to Predict Potential Species Distributions. Futur. Internet 2, 624–634. doi:10.3390/fi2040624

Groenen, F., Meurisse, N., 2012. Historical distribution of the oak processionary moth Thaumetopoea processionea in Europe suggests recolonization instead of expansion. Agric. For. Entomol. 14, 147–155. doi:10.1111/j.1461-9563.2011.00552.x

Herms, D.A., McCullough, D.G., 2014. Emerald Ash Borer Invasion of North America: History, Biology, Ecology, Impacts, and Management. Annu. Rev. Entomol. 59, 13–30.

Hill, A., Guralnick, R., Smith, A., Sallans, A., Rosemary Gillespie, Denslow, M., Gross, J., Murrell, Z., Tim Conyers, Oboyski, P., Ball, J., Thomer, A., Prys-Jones, R., de Torre, J., Kociolek, P., Fortson, L., 2012. The notes from nature tool for unlocking biodiversity records from museum records through citizen science. Zookeys 219–33. doi:10.3897/zookeys.209.3472

IUCN-ISSG, 2015. Global Invasive Species Database [WWW Document]. URL http://www.issg.org (accessed 5.30.15).

Kahle, D., Wickham, H., 2013. ggmap: Spatial Visualization with ggplot2. R Journal. 5, 144–161.

Krikorian, R., 2013. New Tweets per second record, and how! [WWW Document]. Twitter Off. blog. URL https://blog.twitter.com/2013/new-tweets-per-second-record-and-how

Lindenmayer, D.B., Likens, G.E., 2010. The science and application of ecological monitoring. Biol. Conserv. 143, 1317–1328. doi:10.1016/j.biocon.2010.02.013

Liu, B., 2009. Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data (Data-Centric Systems and Applications). Springer.

Liu, B., 2006. Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data, 1st ed. 20. ed. Springer.

Malcevschi, S., Marchini, A., Savini, D., Facchinetti, T., 2012. Opportunities for Web-Based Indicators in Environmental Sciences. PLoS One 7, e42128. doi:10.1371/journal.pone.0042128

Meyerson, L.A., Reaser, J.K., 2003. Bioinvasions, bioterrorism, and biosecurity. Front. Ecol. Environ. 1, 307–314. doi:10.1890/1540-9295(2003)001[0307:BBAB]2.0.CO;2

Miner, G., Elder, J., Fast, A., Hill, T., Nisbet, R., Delen, D., 2012. Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications, 1st ed. Academic Press.

Mykhalovskiy, E., Weir, L., 2006. The Global Public Health Intelligence Network and early warning outbreak detection: a Canadian contribution to global public health. Can. J. public Heal. Rev. Can. santé publique 97, 42–4.

Nasi, R., Wunder, S., Campos A., J.J., 2002. Forest ecosystem services: can they pay our way out of deforestation?

NW-FVA, 2012. Nordwestdeutsche Forstliche Versuchsanstalt, Abteilung Waldschutz: Hinweise zur Überwachung und Bekämpfung des Eichenprozessionsspinners im Waldschutz (25.10.2012).

Pejchar, L., Mooney, H.A., 2009. Invasive species, ecosystem services and human well-being. Trends Ecol. Evol. (Personal Ed. 24, 497–504. doi:10.1016/j.tree.2009.03.016

Pimentel, D., Zuniga, R., Morrison, D., 2005. Update on the environmental and economic costs associated with alien-invasive species in the United States. Ecol. Econ. 52, 273–288. doi:10.1016/j.ecolecon.2004.10.002

Salathé, M., Bengtsson, L., Bodnar, T.J., Brewer, D.D., Brownstein, J.S., Buckee, C., Campbell, E.M., Cattuto, C., Khandelwal, S., Mabry, P.L., Vespignani, A., 2012. Digital Epidemiology. PLoS Comput. Biol. 8, e1002616. doi:doi:10.1371/journal.pcbi.1002616

Sheppard, S.A., Wiggins, A., Terveen, L., 2014. Capturing quality: retaining provenance for curated volunteer monitoring data, in: Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing - CSCW '14. ACM Press, pp. 1234–1245. doi:10.1145/2531602.2531689

Signorini, A., Segre, A.M., Polgreen, P.M., 2011. The Use of Twitter to Track Levels of Disease Activity and Public Concern in the U.S. during the Influenza A H1N1 Pandemic. PLoS One 6, e19467. doi:10.1371/journal.pone.0019467

Smith, A., Brenner, J., 2012. Twitter Use 2012. Pew Research Center's Internet & American Life Project.

Stafford, R., Hart, A.G., Collins, L., Kirkhope, C.L., Williams, R.L., Rees, S.G., Lloyd, J.R., Goodenough, A.E., 2010. Eu-social science: the role of internet social networks in the collection of bee biodiversity data. PLoS One 5, e14381. doi:10.1371/journal.pone.0014381

Steffen, W., Persson, Å., Deutsch, L., Zalasiewicz, J., Williams, M., Richardson, K., Crumley, C., Crutzen, P., Folke, C., Gordon, L., Molina, M., Ramanathan, V., Rockström, J., Scheffer, M., Schellnhuber, H.J., Svedin, U., 2011. The Anthropocene: From Global Change to Planetary Stewardship. Ambio 40, 739–761. doi:10.1007/s13280-011-0185-x

Straw, N.A., Williams, D.T., Kulinich, O., Gninenko, Y.I., 2013. Distribution, impact and rate of spread of emerald ash borer Agrilus planipennis (Coleoptera: Buprestidae) in the Moscow region of Russia. Forestry 86, 515–522. doi:10.1093/forestry/cpt031

Thompson, D.C., 1977. Diurnal and seasonal activity of the grey squirrel (Sciurus carolinensis). Can. J. Zool. 55, 1185–1189. doi:10.1139/z77-153

Tubby, K. V., Webber, J.F., 2010. Pests and diseases threatening urban trees under a changing climate. Forestry 83, 451–459. doi:10.1093/forestry/cpq027

U.S. Government, 2010. Obama Administration Releases 2011 Asian Carp Control Strategy Framework; Press Release 16 Dec 2010 (http://www.whitehouse.gov) [WWW Document]. URL http://www.whitehouse.gov/administration/eop/ceq/Press_Releases/December_16_2010 (accessed 4.26.11).

Van Strien, A.J., van Swaay, C.A.M., Termaat, T., 2013. Opportunistic citizen science data of animal species produce reliable estimates of distribution trends if analysed with occupancy models. J. Appl. Ecol. 50, 1450–1458. doi:10.1111/1365-2664.12158

Vieweg, S., Hughes, A.L., Starbird, K., Palen, L., 2010. Microblogging during two natural hazards events: what twitter may contribute to situational awareness, in: Proceedings of the 28th International Conference on Human Factors in Computing Systems - CHI '10. ACM Press, New York, New York, USA, p. 1079. doi:10.1145/1753326.1753486

wearesocial.net, 2014. Global Social Media Users Pass 2 Billion [WWW Document]. We Are Soc. URL http://wearesocial.net/blog/2014/08/global-social-media-users-pass-2-billion/ (accessed 6.5.15).

Yin, J., Lampert, A., Cameron, M., Robinson, B., Power, R., 2012. Using Social Media to Enhance Emergency Situation Awareness. IEEE Intell. Syst. 27, 52–59. doi:10.1109/MIS.2012.6

Appendix

Supplementary Table 1. Original selection criteria presented to IAS experts on the Aliens-L mailing list in order to identify representative examples for the study.

Selection criteria	Description
Threat to forest ecosystems	The selected species should present a direct or indirect threat to forest ecosystems.
Notable species or impact	The selected species or their impact/symptoms should be notable and thus easy to identify. Grey squirrels or oak processionary caterpillars are examples for notable species, damages by e.g. horse-chestnut leaf miner or the webs of the white moth (Hyphantria cunea) would be examples for notable symptoms.
Multi-fold impact	The impact of the selected species should ideally include multiple quantifiable impacts: biodiversity threats, reduced yield, decreased recreational value, human health threats, reduced scenic beauty, etc.
Existing threat	The selected species should already be present in European forests, thus have already been sighted and present a real threat.
Threat levels	The selection should cover different threat advancement levels: early stage (first sightings), progressing (multiple sightings at a regional level), established (advanced, potentially large-scale invasions with management programmes in place).
Existing monitoring programmes	The selected species should already be covered by monitoring programmes in order to compare available data, verify trends and explore linkages between the different approaches. Ideally this should include both professional monitoring networks and citizen science projects.
Distribution range	The selected species should present threats to forest ecosystems in North America or Europe, here specifically Germany, France, Benelux, UK, Denmark, Sweden, Norway – the primary reason are available language skills for the analysis of the collected tweets and access to internet technologies and social media.
Introduction pathways	Coverage of different introduction pathways (e.g. exotic plants, timber products, etc) would be beneficial as it would open up opportunities for future research focusing on indirect indicators for invasive species monitoring.
Organism types	Coverage of different organism types would be desirable, but given the small set of species we will initially focus on, this selection can of course not be representative. The most interesting groups are probably invasive insects and plants.
Indicator function	Certain species may allow further insights in ecological changes, such as habitat shifts that may be interpreted as signs of changing climate conditions. If the selected species offer this or similar indicator functions this would provide a promising additional perspective.

Supplementary Table 2. List of keywords by species and reference type applied to obtain Tweets via the Twitter Search API potentially referring to the sample IAS. For all species two sets of keywords were compiled with terms that represent potentially direct or descriptive references to the species.

Species	Direct species references	Descriptive species references
European rabbit	*european rabbit; oryctolagus cuniculus*	*saw a rabbit; saw rabbits*
Emerald ash borer	*ash borer; agrilus planipennis; ash borers*	*green insect; green bug; green beetle*
Oak processionary	*eichenprozessionsspinner; thaumetopoea processionea; eichen-prozessionsspinner; oak processionary; prozessionsspinner*	*hairy caterpillar; long haired caterpillar; long hairs caterpillar; haarige raupe; lange haare raupe; langhaarige raupe; hairy caterpillars; long haired caterpillars; long hairs caterpillars; haarige raupen; lange haare raupen; langhaarige raupen*
Asian long-horned beetle	*asian long-horn beetle; anoplophora glabripennis; asian longhorned beetle; asian longhorn beetle; asiatiska långhorningar; asiatiska langhorningar; anoplophora chinensis; asiatischer laubholzbockkäfer; asiatischer laubholzbock; laubholzbock*	*black white beetle; black white bug; black white insect*
Coqui frog	*coqui frog; eleutherdactylus coqui*	*saw this frog; saw a frog; saw these frogs; saw frogs; seen this frog; seen a frog; seen these frogs; seen frogs; saw frog; seen frog*
Ash dieback	*hymenoscyphus pseudoalbidus ; chalara fraxinea; chalara disease; ash dieback; ash disease; eschensterben; eschenkrankheit; eschen krankheit*	*ash tree; ash trees*
Sudden oak death	*sudden oak death; phytophthora ramorum; oak disease; ramorum dieback; ramorum blight*	-
Horse-chestnut leaf miner	*kastanienminiermotte; kastanienkrankheit; rosskastanienminiermotte; miniermotte; leaf miner; cameraria ohridella*	*kastanie; kastanien; horse chestnut; horse chestnuts; kastanienbaum; chestnut tree; rosskastanie; rosskastanien; aesculus hippocastanum*
Pine processionary	*pine processionary; thaumetopoea pityocampa; pinien-prozessionsspinner; prozessionsspinner*	*hairy caterpillar; long haired caterpillar; long hairs caterpillar; haarige raupe; lange haare raupe; langhaarige raupe; hairy caterpillars; long haired caterpillars; long hairs caterpillars; haarige raupen; lange haare raupen; langhaarige raupen*
Grey squirrel	*sciurus carolinensis; grey squirrel; grey squirrels; gray squirrel; gray squirrels*	*saw squirrel; saw squirrels; seen squirrel; seen squirrels*
Rhododendron var.	*rhododendron*	-

Supplementary Table 3. Size and coverage of the selected sample Tweet sets for "topical" and "observational" analysis of the three selected IAS. For each group of datasets the tables show for each species the date range from which Tweets were sampled, the approximate proportion of all Tweets in that date range covered by the sample, the actual number of Tweets obtained as a sample, and the number of Tweets in that sample that eventually had to be excluded at some point of the classification process, due to broken links or unregistered user profiles that were required for the assignment of categories in the information topology.

"Topical" analysis				
Sampling of all Tweets with DIRECT species references				
Species	**Date range**	**~sampled %**	**N (sampled)**	**N (excluded)**
Oak processionary	01.05.-30.09.2013	25.0	225	24
Emerald ash borer	01.05.-30.09.2013	4.2	279	55
Eastern grey squirrel	01.05.-30.09.2013	3.6	258	59
"Observational" analysis				
Tweets with DIRECT species references and embedded or linked MEDIA				
Species	**Date range**	**~sampled %**	**N (sampled)**	**N (excluded)**
Oak processionary	01.04.-31.07.2012	100.0	26	7
Oak processionary	01.04.-31.07.2013	100.0	32	4
Oak processionary	01.04.-31.07.2014	100.0	43	2
Emerald ash borer	15.06.-15.07.2013	100.0	121	5
Emerald ash borer	15.06.-15.07.2014	100.0	214	7
Eastern grey squirrel	01.06.-30.08.2013	50.0	201	54
Eastern grey squirrel	01.06.-30.08.2014	25.0	182	23
Tweets with DESCRIPTIVE species references and embedded or linked MEDIA				
Species	**Date range**	**~sampled %**	**N (sampled)**	**N (excluded)**
Oak processionary	01.04.-31.07.2013	100.0	133	32
Oak processionary	01.04.-31.07.2014	100.0	275	40
Emerald ash borer	15.06.-15.07.2013	25.0	147	49
Emerald ash borer	15.06.-15.07.2014	25.0	193	41
Eastern grey squirrel	01.06.-30.08.2013	12.5	262	77
Eastern grey squirrel	01.06.-30.08.2014	12.5	251	43

Supplementary Table 4. Applied categories in the Tweet analysis for each element of the information topology in Figure 2.

Category group	Categories	Description
Topical relevance	On-topic	References to a species as an IAS or a general biodiversity observation.
	Off-topic	Matching keywords but different meaning (e.g. "green beetle" referring to a car).
	Inconclusive	Undecidable topical relevance due to lack of information or ambiguity.
Interpretation basis	Textual content	Tweet text sufficient for or contributes to decision on topical relevance.
	External links	Externally linked information (URLs) contributes to decision on topical relevance.
	Media (embedded)	Media embedded in the Tweet contributes to decision on topical relevance.
	Media (external)	Externally linked media (e.g. Instagram) contributes to decision on topical relevance.
	User profile	Information in a Tweet author's profile contributes to decision on topical relevance.
	Tweet conversation	Conversation ensuing from a Tweet contributes to decision on topical relevance.
Message type	Observational	Messages representing some form of species sighting.
	Non-observational	Messages referring to an IAS or a biodiversity observation in general.
	General reply to observation	Tweets send in reply to observational messages.
	Determination reply	Tweets send in reply to observational messages and providing taxonomic determinations.
	Unknown	None of the previous message types.
Observation type	Primary observation	A direct species sighting reported by the Tweet author ("I saw this caterpillar").
	Secondary observation	A report of a sighting provided not provide by the original observer (e.g. news items, official announcements).
	Indirect primary observation	A direct sighting of a species' impact or effects reported by the Tweet author (bare branches, feeding patterns, etc).
	Inconclusive	None of the previous observation types.
Species determination likelihood	Impossible/Very unlikely	Verification information (typically media) is lacking or of very poor quality.
	Unlikely	Verification media are available, but of such poor quality that a determination is unlikely even for experts.
	Likely	A combination of medium verification media quality and/or distinctive species features may allow a determination for trained observers.
	Very likely	A combination of good verification media quality and/or distinctive species features will typically allow a species determination (especially for trained observers).
	Certain	A combination of very good verification media quality and/or distinctive species features will guarantee a species determination (often even for casual observers).
Referenced species	Target species	The referenced species can be verified as the target IAS for which a Tweet was collected.
	Non-target species	The referenced species can be verified as the NOT being the target IAS for which a Tweet was collected.
	Inconclusive	Target/Non-target species cannot be verified.
Message subject	General IAS information	General facts about an IAS species, including information on introduction pathways.
	IAS impacts	IAS impacts on host species, human health, biodiversity, etc.
	IAS monitoring	Messages relating to all aspects of IAS monitoring including official monitoring, citizen science, call for support, species identification information, etc.
	IAS management	Message relating to the active management of an IAS including general IAS management information, preventative advice, warnings and alerts, events, stakeholders, etc.
	IAS remedies	Message relating to remedial actions directed at IAS including general

		information, activity reports, new remedies, etc.
	IAS remedies perceptions/opinions	Messages representing public perception of IAS remedies including general criticism, health and environmental concerns, etc.
	IAS research	Messages reporting research on an IAS or IAS remedies.
	Commercial themes	Messages advertising commercial services in relation to IAS such as e.g. treatments, tree removals, consultations, etc.
	Miscellaneous	General message referencing the targeted species but not covering any of the above IAS topics.
Observation verification resources	Embedded media	Media embedded in the Tweet can be applied for determination and/or verification of an observation.
	External media	External media linked from the Tweet can be applied for determination and/or verification of an observation.
	News links/references	News items linked from the Tweet can be applied for verification of an observation.
	Official announcements	Official announcements linked from the Tweet can be applied for verification of an observation.
	Other resources	Other resources (for example other social media) linked from the Tweet can be applied for verification of an observation.
	None	No resources are available to verify and observation or support species determination.
Location references	Tweet text location mention	Place names (e.g. UK, Scotland, Edinburgh, etc) resolvable to geo-coordinates are mentioned in the Tweet text.
	User profile location	Place names (e.g. UK, Scotland, Edinburgh, etc) resolvable to geo-coordinates are provided in the location field of the Tweet author's Twitter user profile.
	Indirect location reference	Relative locations ("my garden", "today at @Caltech") that could be determined on follow-up are mentioned in the Tweet text.
	None	No location information is associated with an analysed Tweet.

Paper IV

“Anyone know what species this is?” – Twitter conversations as embryonic citizen science communities

Stefan Daume [1,2,3 a], Victor Galaz [2 b]

[1] Faculty of Forest Sciences and Forest Ecology, Georg-August-University Göttingen, Büsgenweg 5, 37077 Göttingen, Germany

[2] Stockholm Resilience Centre, Stockholm University, SE-10691 Stockholm, Sweden

[3] Department of Biodiversity Informatics, Swedish Museum of Natural History, Box 50007, 104 05 Stockholm, Sweden

[a] stefan.daume@ecoveillance.org (Corresponding author)

[b] victor.galaz@stockholmresilience.su.se

(Under review at PLOS One)

Abstract

Social media like blogs, micro-blogs or social networks are increasingly being investigated and employed to detect and predict trends for not only social and physical phenomena, but also to capture environmental information. Here we argue that opportunistic biodiversity observations published through Twitter represent one promising and until now unexplored example of such data mining. As we elaborate, it can contribute to real-time information to traditional ecological monitoring programmes including those sourced via citizen science activities. Using Twitter data collected for a generic assessment of social media data in ecological monitoring we investigated a sample of what we denote biodiversity observations with species determination requests (N=191). These entail images posted as messages on the micro-blog service Twitter. As we show, these frequently trigger conversations leading to taxonomic determinations of those observations. All analysed Tweets were posted with species determination requests, which generated replies for 64% of Tweets, 86% of those contained at least one suggested determination, of which 76% were assessed as correct. All posted observations included or linked to images with the overall image quality categorised as satisfactory or better for 81% of the sample and leading to taxonomic determinations at the species level in 71% of provided determinations. We claim that the original message authors and conversation participants can be viewed as implicit or embryonic citizen science communities which have to offer valuable contributions both as an opportunistic data source in ecological monitoring as well as potential active contributors to citizen science programmes.

Keywords: *citizen science, social media, biodiversity, ecological monitoring, taxonomy, Twitter*

1 Introduction

Social online media have emerged as important sources in the monitoring, prediction and modelling of trends and patterns in a broad range of domains. Commercial motivations drive many applications [1,2], as do political and sociological perspectives [3,4]. Applications include support in emergency situations [5,6] or better prediction of natural phenomena [7,8]. Some of the best researched and operational systems can be found in the domain of public health monitoring [9–11]. Increasingly, the potential of social media sources is also recognized in the environmental and ecological domain [12–15]. The volume, real-time nature and simple accessibility of this type of information source as well as advances in Big Data processing methodologies and tools, are major factors in support of this growing body of research on applications of social media mining.

An analysis of social media contributions with ecological significance could start by focusing on general mentions of environmental subjects interpreted as thematic trends [16]), but the pervasiveness of mobile devices with cameras combined with a broad set of social media channels provides great potential for real-time observations of ecologically relevant information [17] that may be contributed casually without knowledge of their ecological significance.

The value of such casual non-expert observations is underlined by an increasing number of so-called citizen science projects where members of the general public contribute to scientific research for example by providing or verifying biological observations [18]. This type of volunteer-driven monitoring contributes both to a wider coverage of monitoring efforts in general, and with the emergence of new technologies [19] may also have to offer more timely monitoring data compared to formal monitoring networks. Opportunistic biological observations in general therefore have the potential to contribute to early warnings of ecological changes [13], not least for potentially irreversible shifts in ecosystems [20]. At the same time, citizen science can serve as a tool to raise the public awareness for ecological changes and challenges [21], thus exhibiting characteristics not unlike social media which equally combine the profiles of data source and communication channel.

There is thus a large potential for ecological applications of this diverse set of social media information types. Compared to the prevailing themes in social media channels (such as music, entertainment, or news) specific ecological subjects may be marginal. However, the

breadth of social media applications and the volume of major social media channels such as Facebook or Twitter hints at a significant amount of valuable information given the right tools [22]. Despite their acknowledged potential, very few tangible applications of these methodologies have been presented in the ecological domain to date.

1.1 Challenges: data quality and compatibility

One explanation for this scarcity of applications is that scholars and practitioners alike are rightfully sceptical about this type of data source. In contrast to professional ecological monitoring programmes social media data is unstructured, contributed outside a monitoring context, and exhibits known demographic and geographic biases [23]. This thus raises concerns about usability, representativeness, reliability and quality – the same concerns frequently voiced when the general value and impact of data generated by citizen science projects is discussed.

Examples from a broad range of domains can be cited to show that public participation in scientific research, can produce high quality data that serves as a valid basis for scientific results [24–27], although specific analytical tools [28] or adjustments through domain-specific contextual models [29] may be required. The growing interest in and importance of citizen science data has however led to a more thorough exploration of data quality, and rather than demanding standard formats and quality scales, approaches to formally capture quality and provenance as meta-data of a data set have been advocated [30,31]. Moreover, this should extend to "traditional" scientific data sources. A review of data managed as part of professional research workflows and large-scale data hubs [32] reveals for example that while of overall good quality it is not exclusively observational data sources originating from citizen science endeavours that struggle with incompleteness or exhibit errors and biases.

It can thus be argued that the concerns raised towards informal information sources such as social media are general, as these apply to varying degrees to all ecological data sources. While requiring specific tools and formal shared standards to ensure usability, it does not preclude the usefulness of this data and its integration in the canon of ecological information sources. Given the volume of this data and the scope of current environmental challenges it seems thus both promising and critical to formally explore social media as data sources in ecological monitoring.

1.2 Social media, citizen science and ecological monitoring

Biodiversity observations are of particular interest from an ecological perspective. As a means to assess the usefulness and feasibility of social media as informal ecological monitoring sources, we collected a broad set of social media posts using the example of invasive alien species (IAS) in forest ecosystems. More precisely, we collected IAS mentions in messages posted on the micro-blogging service Twitter. These sorts of observations provide a good case to explore, as IAS are often not only highly visible for non-experts, but at the same time have well-known ecological impacts [33].

These Twitter messages include cases of users seeking input from their social media networks in order to get clarifications or species determinations on original observations. Thus Twitter users posting requests for species identification and users in their networks answering these requests and providing species determinations show an interest and knowledge in environmental and biodiversity observations.

Biodiversity observations posted on Twitter and the ensuing conversations thus appear to align closely with ecological citizen science data, have to address the same quality concerns and exhibit activity patterns that fit common citizen science typologies [19,34]. Hence, ecological observations shared via social media may at least partially match the models and activities for public participation in scientific research, specifically the data-centric activities typical for "contributory", "collaborative" or "co-created" projects which are frequently indicated as the most common models in citizen science typologies [34,35].

We therefore propose to explore and assess this instance of social media information in relation to citizen science data and activities, and moreover inquire whether these evolving and virtual small ad-hoc communities can be viewed as embryonic citizen science communities that could lead to active contributions to biodiversity monitoring. More precisely, we address two questions in this contribution:

1. What is the type and quality of the attainable social media data, specifically in relation to comparable citizen science projects?
2. What potential do these ad-hoc social media communities hold in engaging actively with citizen science projects?

2 Methods and Data

The data for this contribution was drawn from data originally collected for the *Ecoveillance* project [14], a research initiative that assesses the potential of online social media as informal sources in ecological monitoring in general and as providers for early warnings in particular. The project concentrates on Twitter as one social media source and focuses amongst other, on the example of invasive alien species. The *Ecoveillance* platform [36], developed as a web-based tool utilising the Twitter Search API, has been employed for the last three years to continuously obtain Tweets matching certain keywords that could indicate references to relevant ecological observations.

Keywords range from direct references to selected species ("oak processionary moth", "emerald ash borer"), descriptive references ("hairy caterpillar", "green bug") or general observational statements ("I saw a moth"). From this large pool of Tweets (approaching one million messages) we concentrated for this contribution on a small subset (N=356) containing embedded media or linking to external media (e.g. Flickr or Instagram), and with message texts specifically indicating a request for a determination of a species, thus Tweets containing phrases such as "anyone know what species", "what kind of .." or "what type of ..." (see Supporting Information 1 for a complete list). Figure 1 shows an example of such a Tweet with the triggered conversation.

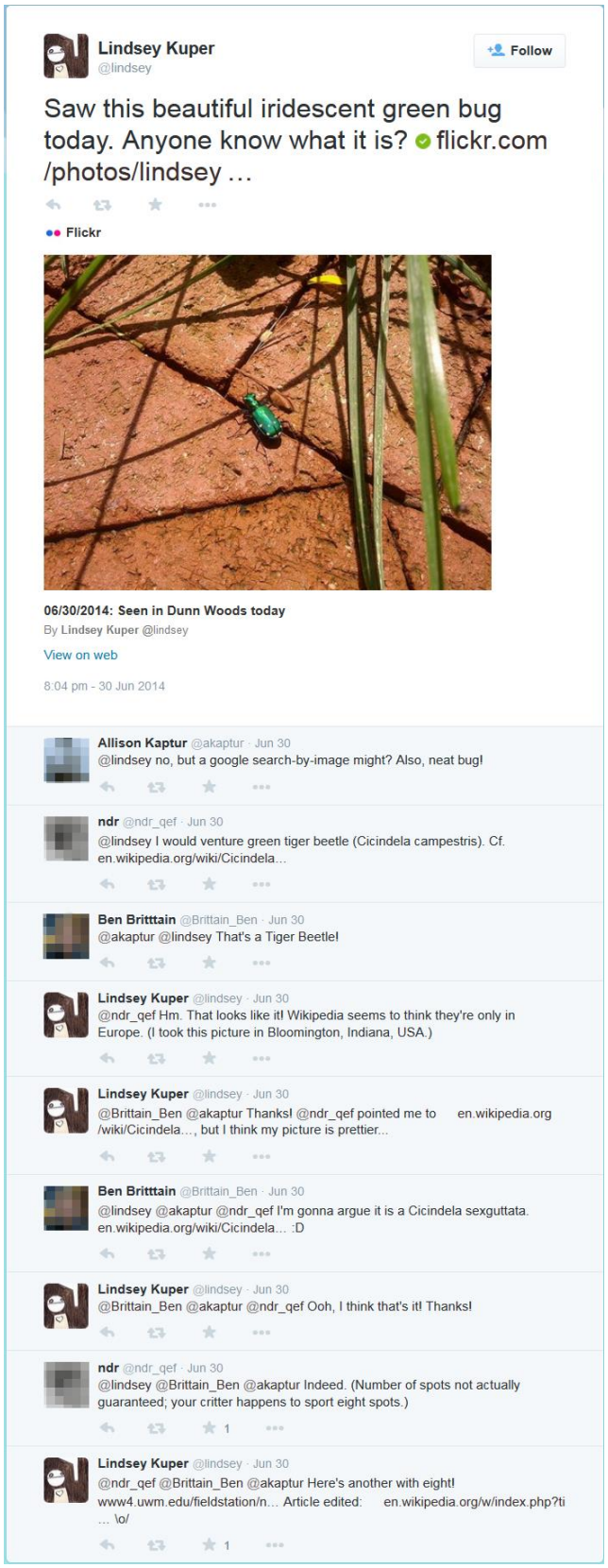

Figure 1. Sample Tweet [37] **requesting a determination with embedded photo and ensuing Twitter conversation (included with permission of the Tweet author).** Two alternative determinations are suggested including scientific names and URL links to a taxonomic reference for verification. The media source is external (Flickr) and includes text with a location reference ("Dunn Woods").

It is important to note that both the subset of Tweets used for this study and the nearly one million Tweets matched by the Ecoveillance platform will represent only a small proportion of all Twitter messages that could be classed as relevant biodiversity observations. Firstly, we concentrated only on English language keywords as search terms, thus limiting the geographic and demographic coverage of the obtained messages. Secondly, both the public Twitter Search API (Application Programming Interface) utilised by the Ecoveillance platform and the alternative Twitter Streaming API provide access to a small share of all potential Tweets; informal estimates for the coverage of these APIs vary significantly with some sources stating that for example the Twitter Streaming API provides a 1% sample of all Tweets in real-time whereas the coverage via the Twitter Search API depends on a combination of a search term's frequency and popularity since this API is geared towards popularity rather than completeness. Operational systems should pursue alternative, and certainly computationally more resource-intensive, approaches to obtain matching data and estimates of the abundance of this information, should cover other languages and apply search terms that specifically target requests for a species determination.

We further filtered our dataset, by removing duplicates and excluding Tweets that themselves were no longer accessible (or essential resources they were linking to, i.e. media links, user profiles), which left us with 215 unique Tweets for analysis; the complete list of Tweets is included as supplementary material (Supporting Information 2).

In an initial classification we concentrated on deciding whether these Tweets with the above matching phrases were indeed "on-topic", thus whether they represented examples for biological observations with a request to a Twitter user's network for a taxonomic determination of the observed species. The results are summarised in Table 1, which also provides information on the interpretation basis of the "on/off-topic" categorisation. While it is primarily the textual content of a Tweet that allows a decision on topical relevance, this is not exclusively the case. The decision basis for topical relevance is of importance when considering a future automatic approach to obtaining, classifying and analysing such Tweets and conversations – if textual content (Tweet messages, user profiles) suffices for this classification, automatic processing can be deemed more feasible.

Table 1. Number of "On-topic", "Off-topic" and "Undecidable" Tweets and the required information items that contributed to the determination of the topical relevance. The interpretation basis percentages in each row add up to more than 100% as multiple information items may have contributed to the decision on topical relevance. Linked URLs, user profiles and ensuing conversations were also considered as a potential interpretation basis, but were not required for this dataset.

	N of Tweets	Interpretation basis (%)		
		Tweet text	Embedded media	External media
On-topic	191	99.5	11.5	8.4
Off-topic	22	95.5	9.1	-
Undecidable	2	-	-	-

The 191 "on-topic" Tweets with embedded or linked media and (where applicable) conversations were subjected to further analysis. Specifically, we

- assessed the quality of the embedded or linked media with regard to a likely determination of the observed and imaged species,
- extracted textual references to geo-locations in the Tweets, geo-coordinates attached to the Tweets and location information provided in Twitter user profiles,
- noted whether the posted Tweet triggered a conversation, how long it was and where it took place (Twitter or external media such as Instagram or Facebook),
- whether the conversation included one or more answers to the requested species determination, what level of taxonomic detail it covered, who provided it and if it was (as far as determinable) correct or not,
- and finally what type of environmental background the requesting and answering Tweet authors had.

Furthermore, we utilised the rich metadata for each Tweet - accessible through the Twitter API - to obtain additional information of relevance such as geo-location information associated with Tweets and user profiles, size of a user's network ("followers" and "friends") or the number of Twitter user's mentioned in a Tweet. The applied categories and utilised metadata will be explained in more detail in the relevant results sections.

3 Results

3.1 Conversations

In an initial assessment each of the 191 "on-topic" Tweets was reviewed for the occurrence of conversations on either Twitter or social media sources linked from the Tweet (specifically Instagram, Flickr, Facebook). Table 1 summarises the results of this analysis.

Table 2. Conversations identified for the analysed data set, showing total number and shares by conversation medium and reply type (i.e. with or without species determinations). A 'determination reply' is a conversation that contains at least one reply that suggests a taxonomic determination of the posted biodiversity observation. Shares add up to more than 100% as some Tweets received parallel replies on multiple media.

Conversation	None	Twitter	Instagram	Facebook	Other	Σ
No reply	**69** (36.1%)					*69*
General reply		**12** (6.3%)	**5** (2.6%)	**1** (0.5%)	**1** (0.5%)	*19*
Determination reply		**77** (40.3%)	**35** (18.3%)	**2** (1.0%)	**1** (0.5%)	*115*
Σ	*69 (36.1%)*	*89 (46.6%)*	*40 (20.9%)*	*3 (1.5%)*	*2 (1.0%)*	

Overall, 64% of all Tweets analysed are answered, and 86% of those conversations do contain at least one reply providing a species determination in response to the original Tweet author's request. Twitter and Instagram, an image sharing tool with options to reply to posted images, are the primary conversation media, with no significant difference in the share of replies with determinations. In 12 instances parallel conversations on Twitter and Instagram could be observed, but the majority of conversations happen exclusively on one medium, mostly on Twitter.

3.2 Observational characteristics

Each Tweet analysed in this contribution represents a unique biological observation. In the following sections we provide an overview of the media type and quality, temporal patterns, geo-information and source meta-data associated with these observations. This will help assess the type and quality of the attainable data.

3.2.1 Media type and quality

All analysed Tweets contained images as the sole shared media type rather than videos or sound files, which are also frequently shared on Twitter or other social media sites. Figure 2 summarises the found image types, highlighting that the majority (65%) are embedded in the Tweet, thus visible directly to a user viewing the post. This may in fact apply to other media as well such as Flickr and Twitpic, which have only negligible shares though. Images shared through Instagram and linked from the Tweet account for approximately 27% of all posted images.

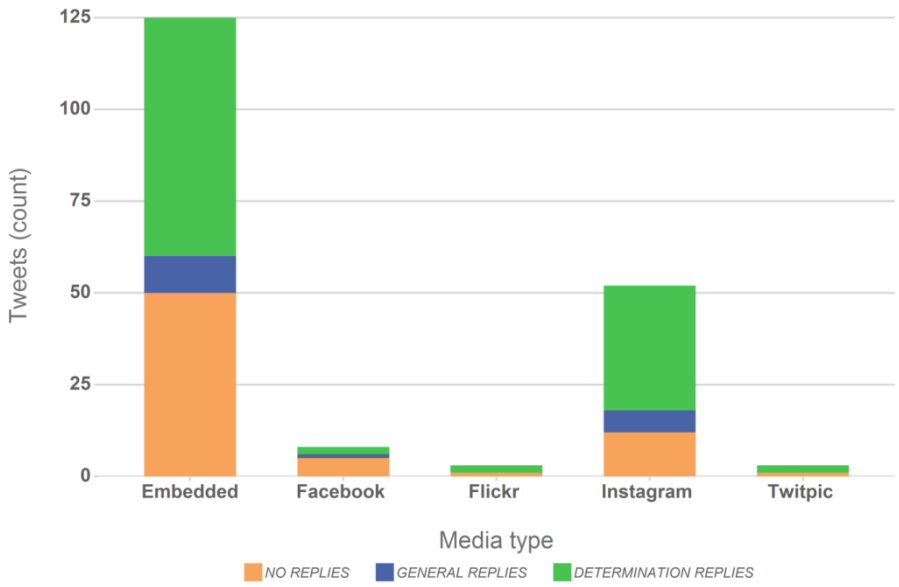

Figure 2. Type of embedded or external media associated with the analysed Tweets and type of reply. 'Embedded' media are images that are shown embedded within the Tweet text. External media appeared as URL links to Facebook, Flickr or Instagram in the Tweet text. The share of unanswered Tweets (orange), Tweets receiving replies with (green) and without (blue) suggested taxa is highlighted for each media type.

Aside from relevant expertise in a Tweet author's network a key factor in receiving determinations will be the quality of the posted images. Images of higher quality will be more likely to support a conclusive identification of an imaged species. Factors contributing to a higher quality and in turn likelihood for species determination are the general quality features of an image (resolution, lighting, sharpness, contrast, colour space), the relative size of the photographed species, distinctiveness of the species itself as well as the availability of direct or indirect scales, helpful peripheral information or textual content

that provides context to the image and the captured species, such as for example mentions of colours or geo-locations.

Focusing on these features and with a view on the likely determination of the imaged species the quality of all images posted with the 191 "on-topic" Tweets was assessed manually and assigned to five quality classes ranging from *"very poor"* to *"very good"* with the former assuming that a determination may be near impossible while for the latter a species determination was assumed near certain if an expert would be given access to the picture. Figure 3 summarises the results of this quality assessment, again distinguishing between images that received no reply, general replies and determination replies. Overall, the majority (81%) of shared images were of satisfactory or better quality thus lending itself to verify or identify an observed species.

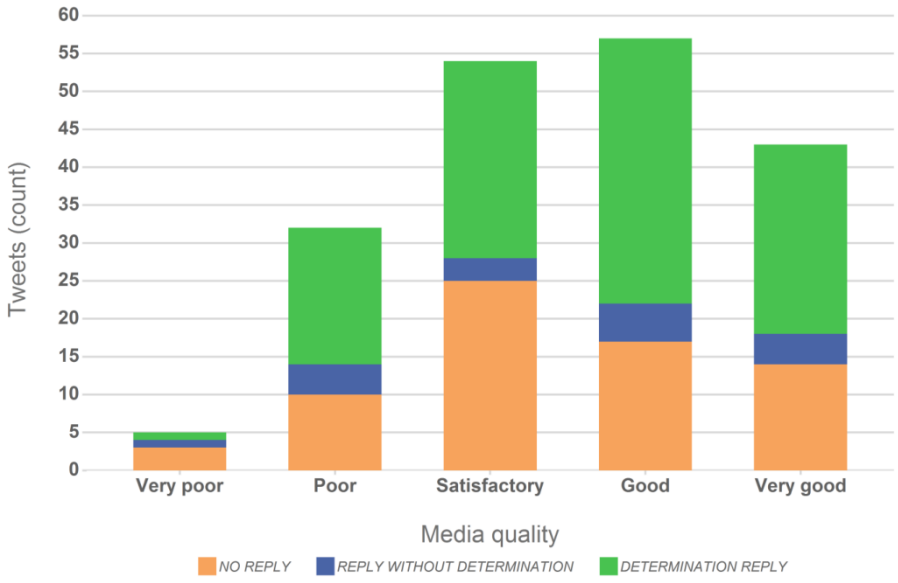

Figure 3. Quality of the posted media with regard to a likely determination of the captured species distinguished by type of reply. The share of unanswered Tweets (orange), Tweets receiving replies with (green) and without (blue) suggested taxa is highlighted for each quality class.

3.2.2 Temporal characteristics

The analysed "on-topic" Tweets were posted during the period from May 2013 to December 2014. Figure 4 (A) shows the weekly frequencies of these Tweets in the data collection timeframe. For a start, the date distribution actually underlines the bias in our data. Since

we used Tweets for this analysis that were originally collected with a focus on invasive alien species in forest ecosystems, the distribution is seemingly a reflection of the lifecycle of the originally targeted species, rather than necessarily the observational activity of the authors posting the photos and requesting determinations. While our original sampling focus and sample size may not allow a generalization, it is however fair to assume that the type of casual observation we analysed here are more likely made during core lifecycle phases of the observed organisms, not least because daylight and weather conditions will probably coincide with general and recreational outdoor activities of the potential observers.

Figure 4(B - D) illustrates that the posted observations exhibit some interesting additional temporal features. Based on the content and wording of the analysed Tweets, we can assume that these observations are casual rather than deliberate monitoring events. The weekday distribution of the posted Tweets confirms this (Figure 4(B)), with a clear spike on Sunday (22% of all Tweets), thus a day where people can generally be expected to be off work and engage in recreational activities. While a similar pattern may be expected for Saturday, it is very pronounced for Sunday.

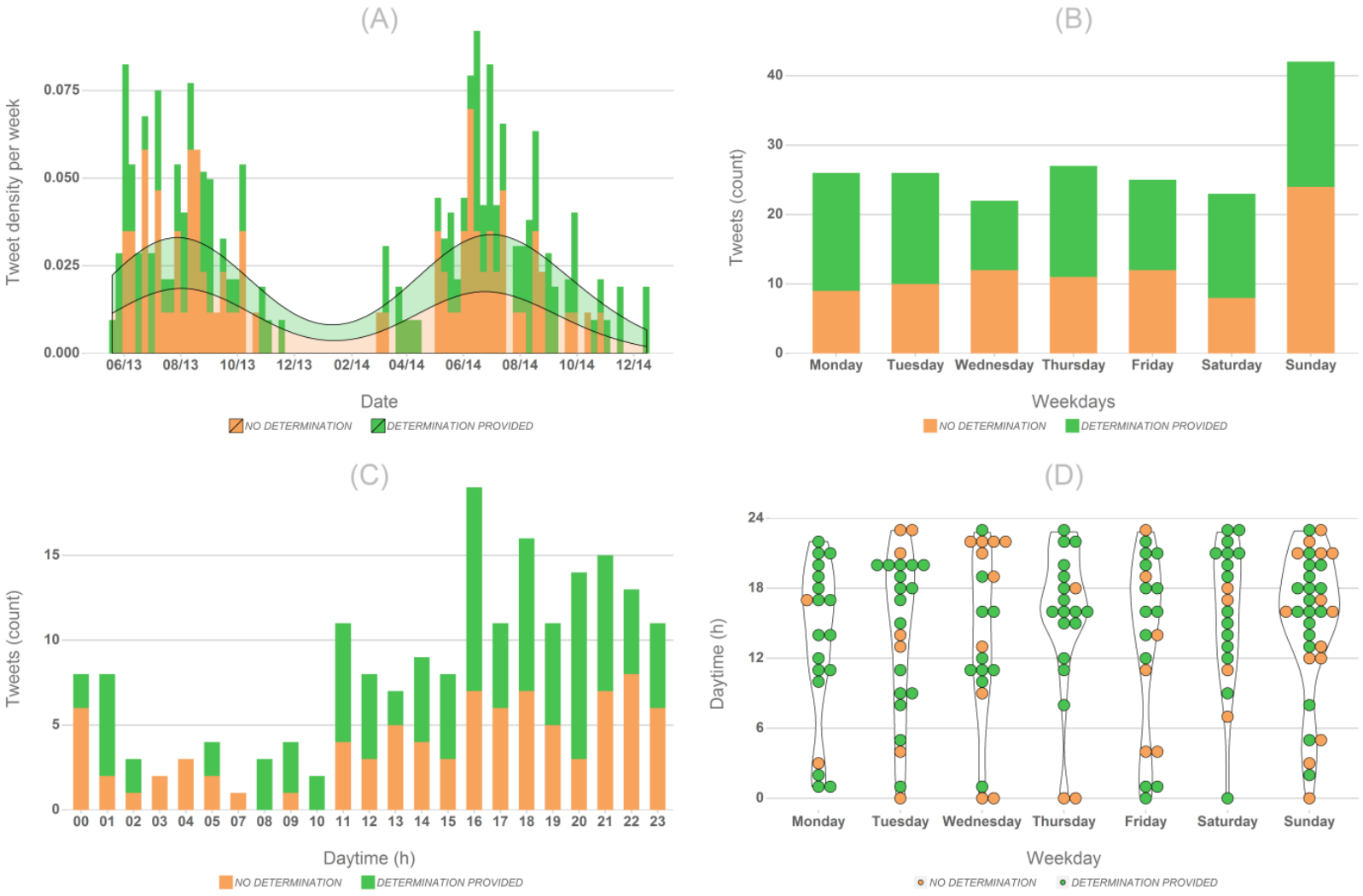

Figure 4. Temporal characteristics of the analysed Tweets. Tweet frequencies (A) by week in the data collection time period, (B) by weekday, (C) by hour of day and (D) as a combined distribution over weekday and daytime.

Looking at the daytime distribution of the Tweets (Figure 4(C)), it is not surprising that there is little activity at night-time and early morning. A slight peak can be observed around lunchtime, again suggesting the opportunistic and casual nature of observations probably made during lunch breaks (also confirmed by the combined weekday/daytime distribution in Figure 4(D)). However, it is notable that many observations are posted late in the day and evening, even more so given that the review of the pictures suggested that they were almost exclusively taken in daylight. There is thus a notable reporting latency and the Sunday reporting peak could also be attributed to observations made on Saturdays and reported with a latency of a whole day rather than just few hours. However, the pronounced reporting peak at the weekend seems to support a reporting latency of up to 24 hours rather than several days or weeks, and significantly larger latencies (weeks or months) can be excluded given the majority of observed species (Figure 10) and the good fit of the observation frequencies with the lifecycles of those species (Figure 4(A)).

Hence, while pictures were apparently taken casually in daylight, the latency in posting the request for determination suggests a more than casual interest in the subject. We will revisit this observation when discussing the potential of the Tweet authors' contributions in the context of citizen science in general.

3.2.3 Geo-location information

With few exceptions, biodiversity observations will be of value only in connection with geo-location information, both for ecological monitoring in general as well as input to ecological models. Twitter applications typically offer a mechanism to attach geotags (detailed geo-coordinates) to a posted Tweet, which are then available as part of the Tweet metadata via the Twitter API. These geotags will be of particular quality and accuracy when Tweets are sent from GPS-enabled mobile devices.

However, of the 191 analysed Tweets only two were available with geo-coordinates. This is in line with results in other studies. [38] found that the share of Tweets with geo-coordinates is typically in a range of 0.5% to 3%, but depending on the studied subject and messages' geographic origin the proportion of precisely geo-tagged Tweets can be significantly higher; [8] cite several studies where geo-tagged posts accounted for 5% to 16% of collected Tweets.

Geo-information, albeit less accurate and reliable, is however available in other forms as well: as volunteered location information in a Twitter user's profile or as textual references. User profile location information is obtainable through the Twitter API. Figure 5 summarises the granularity of the available geo-location information in user profiles of the authors of the analysed Tweets. Nearly one third (31%) of users do not provide usable location information in their profiles, but 43% of user profiles hold locations at the granularity level of "City", which may still cover a large region (New York, London), but narrows the geo-placement of the observed species. This again matches results in other studies: both [38] and [8] report that descriptive toponyms associated with analysed Tweets vary within an equally broad range (21-70%). However, there is no guarantee that Tweet authors took the posted images at their profile location and profiles may get out of sync with a Twitter user's actual location.

Figure 5. Granularity of location information in user profiles for the analysed Tweets and share of geo-enabled user profiles in each location category.

Textual location references were found in 20% of analysed Tweets. While more reliable they exhibit a similar granularity (ranging from e.g. "USA" to "Rainham Marshes" or "Ashford train station") and require a similar validation and mapping step if extracted automatically from the message text.

Figure 5 highlights another characteristic with regard to Twitter geo-information: the proportion of Tweet authors with "geo-enabled" profiles, a Twitter platform setting that triggers the automatic geotagging of Tweets. Interestingly, 37% of all analysed Tweet authors and 49% of those providing a "city" in their user information had geo-enabled profiles, and yet only two Tweets in our dataset carried geo-coordinates. The most likely explanation is that while user's geo-enabled their Twitter profiles (which is not the default setting), they did not geo-enable their devices or blocked geo-tagging on those devices for certain applications. We can only speculate if this is a deliberate choice for the specific Tweets we analysed or if this setting merely was forgotten resulting in the lack of geo-coordinates. We nevertheless can observe that the authors in our dataset must have made a deliberate choice to geo-enable their profiles. Thus, while geo-information is lacking or does not propagate through, if settings on the used devices were in sync with this choice we could expect a large amount of observations with high-resolution geo-information.

3.2.4 Tweet source devices and applications

A look at the prevalent source devices and applications of the analysed Tweets (Figure 6) underlines the point made in the previous section; information on the source device or application is part of the metadata obtainable for a Tweet via the Twitter API. We summarised instances clearly identifiable as originating from *Mobile devices* (e.g. Blackberry, iPhone, Android) or *miscellaneous web* applications (e.g. Twitter website, TweetDeck) in two classes, the remainder originated from Twitter integrations of Instagram or Facebook.

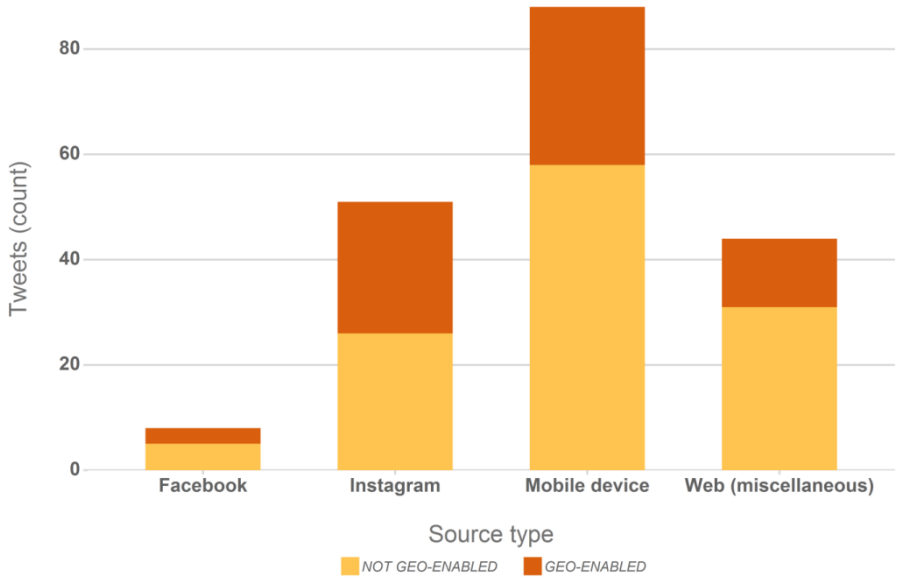

Figure 6. Source devices and applications from which the analysed "on-topic" Tweets originated with an indication of the share of geo-enabled profiles associated with Tweets in each category.

Mobile devices were clearly identifiable as sources of the postings for 46% of the posted Tweets. The actual number is however almost certainly significantly higher, since Facebook or even miscellaneous web applications could have been used from mobile devices and Instagram (27%) is geared towards mobile usage. We can thus assume that the potential rate of geo-tagged Tweets could be much higher if a user would choose to share geo-coordinates from a mobile device together with the published Tweet.

3.3 Species determinations

In Table 2 we distinguished between conversations with general replies and those containing at least one suggested determination for the species a Tweet author photographed – in 86% of all conversations at least one of the participants provided a determination. Figure 7 explores the number and nature of these determinations in greater detail. The majority of conversations (56%) contain only one suggested determination. Of the 37 conversations with more than one determination 32% contain alternative determinations. We consider those "conflicts" as resolved, if the contributing conversations authors settle on one determination or at least one of the determinations is correct, which applied in all but 4 instances.

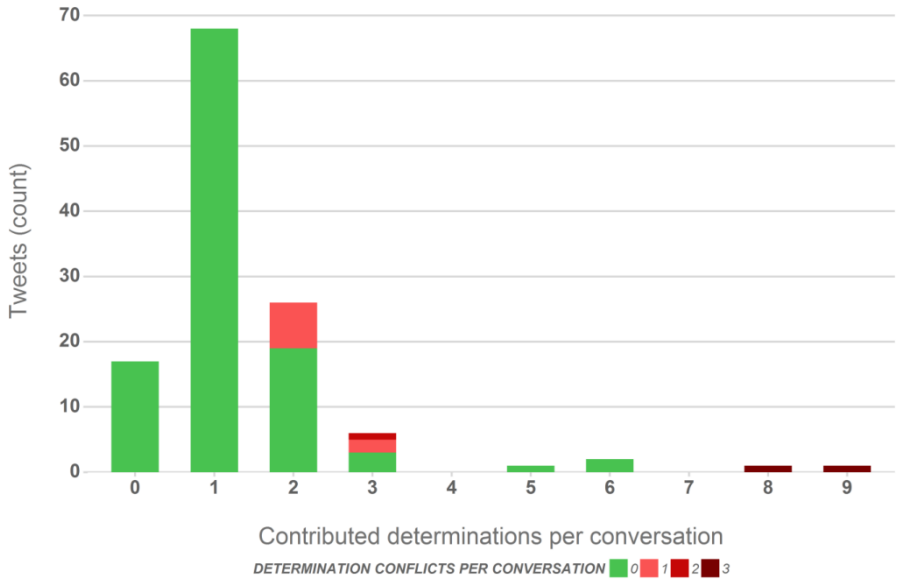

Figure 7. Number of suggested taxa (determinations) per Tweet for Tweets receiving replies and number of conflicting determinations. The share of Tweets with conflicts and the number of conflicts is indicated with a red colour scale.

On closer inspection, these "determination conflicts" or longer determination conversations represent valuable information by itself since they capture a vetting process that can be interpreted as explicit meta-data on the reliability of the information. Sometimes these conversations take the form of singular determination statements, sometimes additional information is requested and provided, leading to improved determinations. Moreover, these type of conversations offer contextual information that will not be available in standard biodiversity observation databases, for example when contributors express surprise about a sighting at a particular location or outside an expected time window, mention the rarity or commonness of a species, or comment on the reliability of a determination in the context of geo-information, lifecycles or other environmental contexts. All these variations were represented in our dataset, but given the size of the available Tweet sample are illustrative and do not yet permit provision of a detailed profile of this interesting contextual meta-data.

Figure 8 illustrates the relation between the number of replies and determinations per conversation, which expectedly suggests that longer conversations contain more determination replies. This trend is however not very pronounced. As Figure 7 illustrated as well, the majority of conversations are short and only have one or two determination

replies. This observation could be explained by either assuming that the authors requesting a determination only have access to a small pool of experts in their network or that a provided determination reduces the motivation for others to contribute an additional answer.

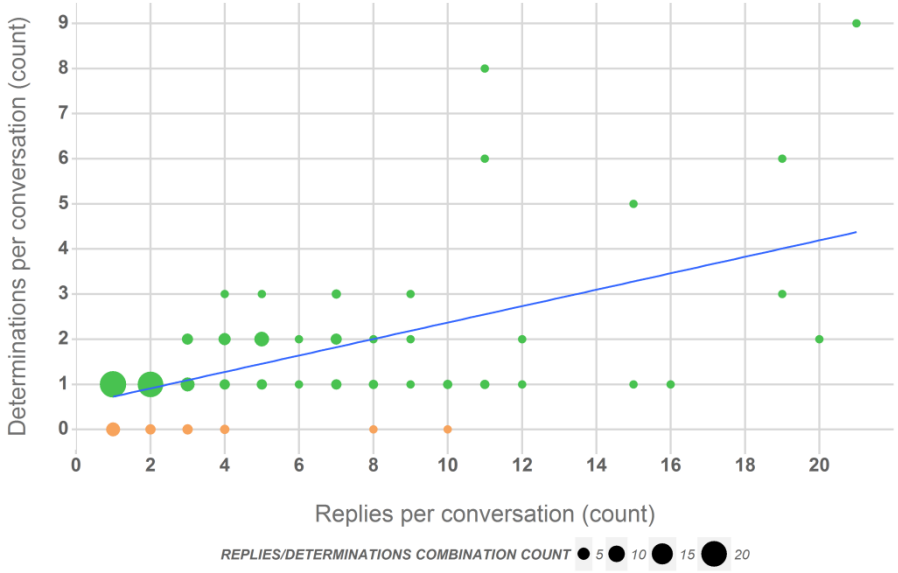

Figure 8. Relation of number of replies containing a species determination to the total number of replies per analysed conversation. For readability orange circles were used to highlight conversations without determination replies as opposed to conversations with determination replies (green). The size of the circles indicates the frequency with which a specific combination of reply and determination counts was found.

In a further analysis of the provided determinations we noted the level of taxonomic detail, the used terminology and the actual provided determinations. Figure 9 summarises the highest taxonomic level provided for each conversation with determinations and whether common or scientific names were used. In 71% of all conversations the request is answered with a determination at the species level. Only in 12% of cases however the determination providers contribute determinations using scientific names. In 16% of conversations the determination providers back up their claim with a link to taxonomic references such as for example *ukmoths.org* or *Wikipedia*. Figure 10 is included for illustration, quoting all determinations provided in the analysed conversations.

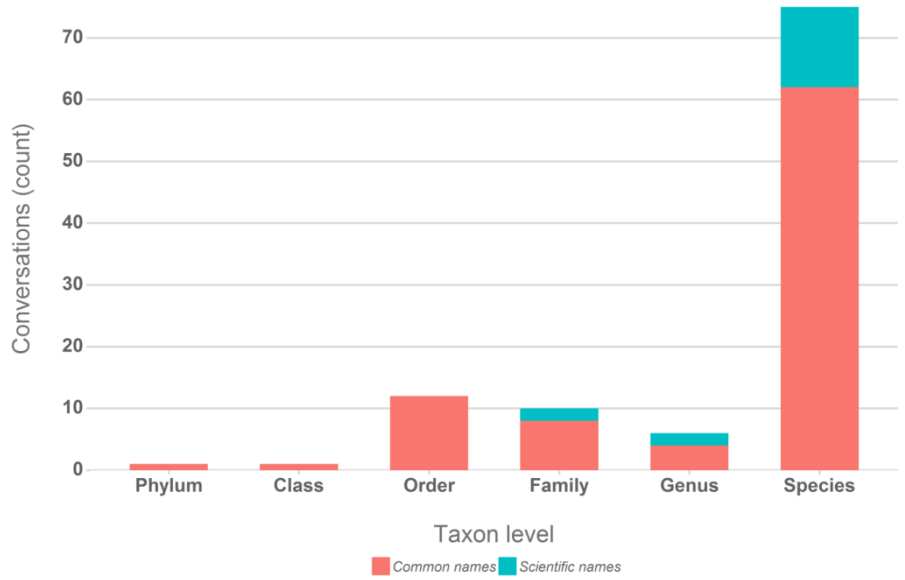

Figure 9. Highest taxonomic detail and choice of terminology (scientific or common names) for provided determinations per conversation.

Figure 10. Word cloud showing all provided determinations in the analysed conversations.

In a final evaluation of the provided determinations we followed up the claims and tried to assess if a determination conversation resulted in a correct determination. Some of these assessments had to be marked as uncertain, due to the quality of the posted images, limited visibility or lack of distinctive features of the assessed organisms as well as the authors' taxonomic expertise.

Table 3. Assessment of correctness of contributed taxonomic determinations in the analysed conversations. The numbers in parentheses indicate assessments where correctness could not be decided with absolute certainty.

Determination assessment	Correct	Partially correct	Incorrect	Undecidable
Conversation count	80 (29)	2 (1)	9 (7)	14
Conversation %	76.2% (36.3%)	1.9% (50.0%)	8.6% (77.8%)	13.3%

Overall the quality and reliability of the provided determinations can be assessed as high. With caveat of the noted uncertainty margins, only 9% of the determination conversations produced incorrect results while 78% were correct or partially correct.

3.4 Contributor classification

In order to reflect on contributors to the analysed Tweet observations and determination conversations in the context of citizen science, we carried out a categorisation of the original Tweet authors and users providing determinations. Our classification scheme was motivated by the question whether the two groups of observation and determination contributors are dominated by contributors with a documented environmental interest, education or profession.

We included all users (N=191) contributing observations (including unanswered ones), and all users (N=114) providing determinations in Twitter conversations. Users contributing determinations in Instagram conversations were not included because the user information accessible on these sites did not provide a sufficient basis to assess the background or interest of the users. For Twitter users their environmental interest or formal domain

education was assessed manually based on their Twitter user profiles, external sites linked from those profiles and the content of their other Tweets. Table 4 provides a complete list and explanation of the applied author classes.

Table 4. Classification scheme applied to all Twitter authors requesting and providing species determinations in the analysed Tweet set.

Class	Classification criteria
Domain professionals	Individuals with a formal education and/or profession within the environmental or biological domain including for example researchers, foresters, farmers or professional gardeners, etc.
Amateur biologists	Individuals with a specialised biological subject interest (entomology, ornithology) pursued as a recreational activity but following professional standards and methods. This includes individuals with a documented participation in citizen science projects.
General nature enthusiasts	Individuals with a strong documented personal but not professional interest in nature and outdoor activities (e.g. gardening, photography), including environmental activists.
Environmental organisations	Organisations with a documented association to environmental or biological subjects, including research organisations, conservation groups or gardening associations, entomological or ornithological societies, etc.
Social media aggregators	Special Twitter channel dedicated to "retweeting" Tweets by other users reporting biological observations.
Miscellaneous organisations	Miscellaneous public or private organisations, including companies, with no discernible environmental background or domain function.
"Incidental" biologists	Contributors of the analysed Tweets or conversations with no discernible domain background or documented environmental activities.

Figure 11 compares the shares of the different contributor types in the two groups of Twitter authors posting observations and requesting species determinations, and those providing determinations. Both groups are dominated by individuals, organisational contributors account for a marginal share only. Furthermore, in both groups contributors with no discernible environmental background (denoted *"Incidental biologists"*) represent the largest share, 64% of determination requesters and 46% of determination providers. The second largest contributor type – with 21% and 22% respectively – are those termed *"General nature enthusiast"*. Very few individuals (8%) with a professional or quasi-professional domain background (*"Domain professional"*, *"Amateur biologist"*) request determinations but this group accounts for 25% of all provided determinations.

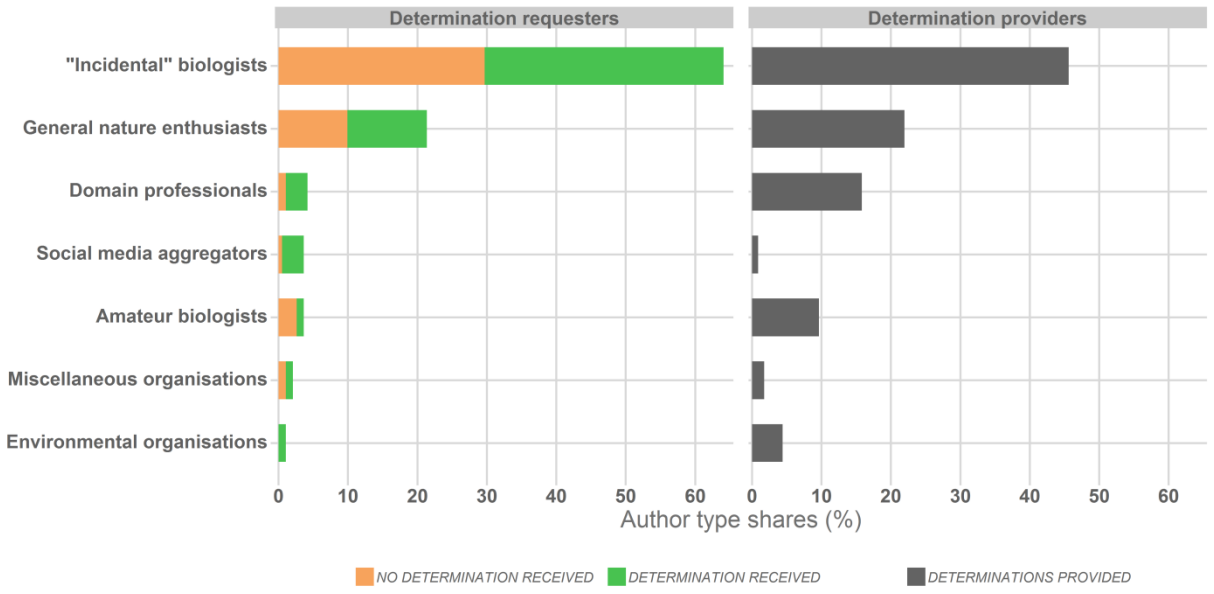

Figure 11. Type of users requesting determinations and providing determinations, with categories indicating contributors with a documented environmental interest, education or profession if any. The classification of users is based on available Twitter profiles, linked personal pages and the content of Tweets authored by them; *"Incidental" biologists* denote users with no discernible biological/environmental background or activities.

Figure 12 adds an additional dimension to the contribution of different author types requesting and providing determinations. The request for and provision of determinations is represented as a network that captures the frequency of author types and the frequency of certain combinations of author types.

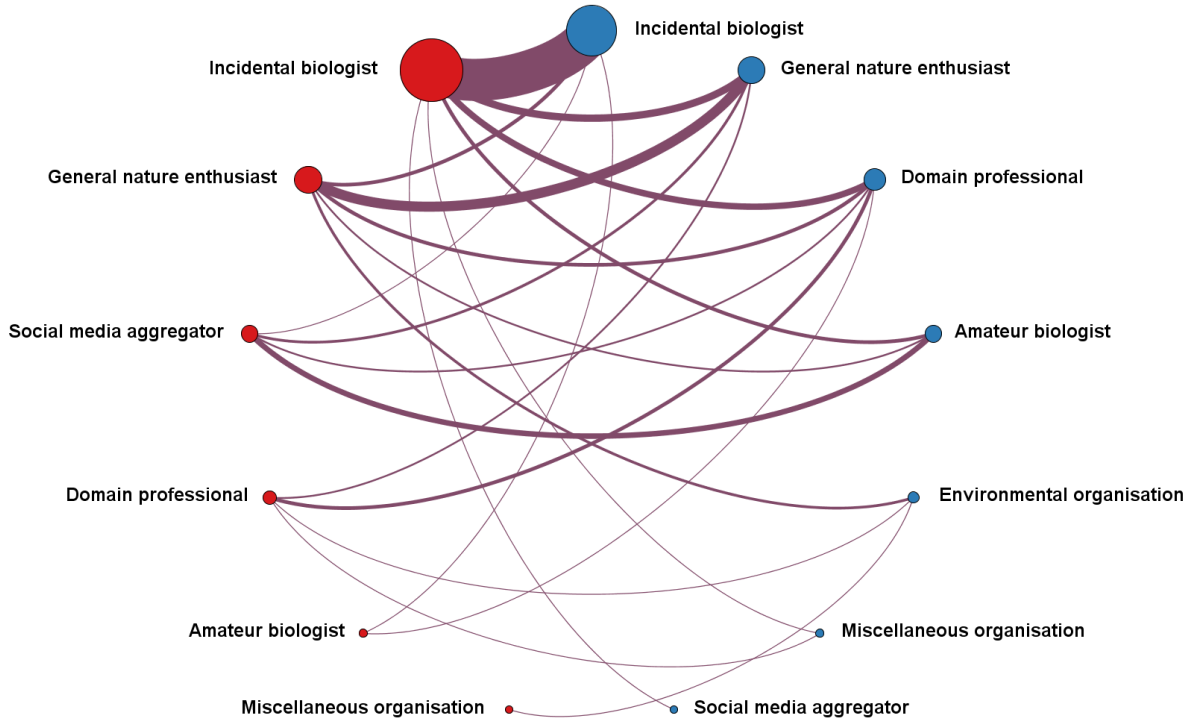

Figure 12. Connections between different types of determination requesters and providers. Red circles represent user types requesting and receiving determinations, blue circles user types providing determinations. The size of the circles indicates the frequency of an author type, the size of the edges the frequency with which a particular pairing can be found. The graph was generated with the Gephi (https://gephi.org) network visualization tool using a Circular network layout.

As already noted in Figure 11 the largest group are Twitter users with no discernible environmental background (*"Incidental biologists"*). They also represent the most abundant connection, thus requests by users with no documented environmental background are answered by the same type of users. This may not be surprising considering the dominance of this user type in our dataset, but notable when considering the large proportion of correct determinations.

The second most abundant connection is between the contributor classes termed *"General nature enthusiast".* We can observe that determination requests by this contributor type are primarily answered by the same type of contributors. At the same time *"General nature enthusiasts"* are the second most frequent determination providers to users with no environmental background (*"Incidental biologists"*).

Interestingly, we can also observe that if users with a formal or professional background ask for determinations (*"Domain professionals"*) they are predominantly answered by users

with a similar background (*"Domain professional", "Environmental organisation")*. It can be argued that this should not be surprising since connections in Twitter networks will be driven by shared interests, thus researchers can be expected to have other researchers in their network. However, Twitter networks will not be completely homogenous and one can ask if the willingness by non-experts to provide suggestions to experts is influenced by the self-assessment of a user's own domain expertise. However, given the abundance of non-experts providing determinations and the overwhelming correctness of those, we can speculate if there is not even more contributory potential that can be mobilised within the ranks of those considered or considering themselves non-experts.

3.5 Results summary

We analysed 191 Tweets with biodiversity observation posted with a species determination request, 64% received replies, 86% of those contained at least one suggested determination, of which 76% were assessed as correct. All posted observations included or linked to images with the overall image quality categorised as satisfactory or better for 81% of the sample and leading to taxonomic determinations at the species level in 71% of provided determinations.

While acknowledging that we used a dataset originally collected for another purpose and thus working with a comparatively small sample, the above summary of some of the main results suggests that we are dealing with a valuable resource both with regard to the published biodiversity observations as well as the contributions of the participating community. Importantly, this data can be considered as lost since it is published outside an ecological monitoring context and channel, thus not collected, assessed and utilised, which highlights the potential contribution of this data source in ecological monitoring efforts.

4 Discussion

One of the key features of observational data obtained via social media channels such as Twitter, Facebook or Instagram is its real-time nature. In light of a recent critique of shortcomings of traditional ecological monitoring programmes [39] the value of real-time monitoring data in particular can be stressed, and an exploration of social online media as

additional data sources in ecological monitoring seems merited as it may help to address not only issues such as timeliness, but also contribute to question-driven monitoring [14,39].

This type of data can be considered even more valuable if it extends beyond plain and undetermined observations and is instead vetted and reviewed, thus possibly approaching the level of detail and quality contributed in common non-expert, volunteer-driven citizen science monitoring efforts such as Artportalen [40] (https://www.artportalen.se), OPAL [41] (http://www.opalexplorenature.org), eBird [42] (http://ebird.org) and many others [18]. In that context we explored a set of Twitter observations and ensuing conversations. Our analysis was motivated by the potential these observed ad-hoc virtual communities hold with regard to active contributions to citizen science initiatives. We discuss the analysed social media data and its "embryonic citizen science nature" with reference to the two research questions we posed in the introduction.

4.1 What is the type and quality of the attainable social media data, specifically in relation to comparable citizen science projects?

While identifying certain differences and gaps in the data profile, we claim that overall the analysed biodiversity observations in the form of Twitter messages and conversations do approach the type and quality of comparable citizen science data, and under consideration of the highlighted shortcomings deserve an intensified scientific and practical exploration.

A key difference between the analysed social media data and data sourced through citizen science projects is that the latter imposes a structure that is largely lacking for the analysed Twitter posts. "Rapporteurs" to the Swedish Species Observation System (Artportalen) are for example required to provide the full species name (verified against the taxonomic backbone *Dyntaxa*), geo-referenced location, the time of observation and the name of the observer [40]. However, this information is available in our analysed Twitter observations and conversations in a semi-structured format. Specifically:

- The key data item in our analysed Tweet observations are the embedded or linked images, generally accessible without restrictions, predominantly of good quality (Figure 3), thus providing sufficient detail to enable a taxonomic expert validation and determination.

- The temporal features suggest that the provided data is real-time, reported with a low latency and a good reflection of the lifecycle of the majority of the observed species, hence in line with typical biodiversity observation programmes.
- Observer and determiner information is equally available through the used Twitter accounts. The associated user profiles do not only provide background information on the contributors, but also a direct communication channel to follow up on observations or determinations.
- Precise geo-coordinates are scarce, but geo-information, albeit of lower granularity, is also available in the form of user profile locations and textual location references.
- Finally, the available information enabled 71% of determinations at the taxonomic level of "Species", 76% of determinations were assessed as correct, although only for 16% of the determination conversations the use of scientific taxonomic names could be observed.

While acknowledging the lower quality level, we argue that there is thus only a technical rather than a conceptual challenge to utilise this data, possibly by feeding it into existing citizen science portals like Artportalen. The most notable challenges are the current lack of high-quality and reliable geo-location information as well as the level of taxonomic detail.

With regard to the first challenge, we find however that with little effort on the part of the Tweet authors the majority of observations could come with exact geo-coordinates: more than 2/3 of postings are apparently submitted from mobile devices which can be assumed to have GPS functionality, hence allow the provisioning of geo-tags; furthermore, more than half of the Tweet authors in our dataset already had their Twitter profiles geo-enabled. Thus, if users could be encouraged to actively contribute observations, the utilised devices, applications and social media settings would suffice to guarantee a high degree of detailed and reliable geo-information which would not require any regular manual intervention by the user, but could possibly be of even higher quality than manually contributed data on certain citizen science platforms. While this observation is encouraging from a technical perspective, we have to take it with the caveat that we can only speculate about the reason for the surprising mismatch between the large share of geo-enabled user profiles and the lack of geo-coordinates.

The second challenge concerns the quality of taxonomic determinations. Artportalen requires observations to be reported at species level and with full scientific names.

Especially the latter is not matched by our social media sample. We still argue that the quality of the recorded determinations can be judged as fairly good considering the casual conversational context and primary background of the users. Moreover, we could possibly expect contributions of higher quality, greater detail and using scientific terminology if contributors knew that they were submitting determinations to a biodiversity monitoring project. Finally, in the case of multiple (conflicting) determinations, these conversations capture a determination process in addition to determination result, which represents interesting meta-data in itself and deserves a broader and more detailed exploration with larger samples.

4.2 What potential do these ad-hoc social media communities hold in engaging actively with citizen science projects?

We claim that the posted biodiversity observations and ensuing determination conversations clearly match typical data collection and interpretation activities in citizen science projects [35], the data is comparable to that collected in citizen science projects and the contributor profiles hint at a large pool of contributors previously not engaged in citizen science, thus showing significant potential should the participants in our study be encouraged to graduate from a passive to an active citizen science status.

While we were not able to address those Twitter users directly and thus had to employ an indirect approach to elucidate the likely motivations, we can infer some triggers and motivations based on specific Tweet samples. In some cases the motivations were of practical nature, such as questions about the impact of a species on gardening plants and possible remedies, mostly however the basic desire for knowledge, an interest in learning what species an observation (often with a distinctive appearance) belonged to and in some cases the authors of the Tweets seemed to be motivated by a sense of discovery as indicated by for example enquiries about the potential rarity of a species. Similarly, determination providers appear to enjoy sharing their knowledge with others, and in some cases their comments and questions and the sharing of supplementary information suggested that they may also be motivated by an educational element of their participation.

Our results indicate that posted biodiversity observations and requests for determinations receive significant interest and active participation from within a Tweet author's network

(Table 2), which suggests that there is a notable implicit community detectable around these types of casual biodiversity observations. At the same time we have to note however, that the observable communities per Tweet are comparatively small; the majority of conversations receive one or two determination replies (Figure 8) and few determination conversations have more than two determinations including discussions around alternative determinations (Figure 7). While our results suggest only a small proportion of true experts in these networks, this does not necessarily imply that there is also small share of people able or willing to reply a determination request. This can equally be attributed to conversational etiquette (i.e. it is unlikely that a user contributes a concurring opinion if the question has already been answered) rather than the number of knowledgeable potential contributors in a Twitter user's network.

This is further supported by our categorisation of the author types: it is notably users who are not active citizen scientists, amateur biologists or domain professionals with a formal biological education that contribute observations and provide determinations (Figure 11), and non-experts or general nature enthusiasts communicating with each other (Figure 12) account for the majority of conversation replies producing determinations with a high correctness (Table 3).

In combination with the observed latency in "tweeting" the captured images, which indicates an interest in the shared observations that extends beyond the moment when the Tweet authors casually take a photo, we argue that this suggests the presence of a large pool of contributors that are currently not actively participating in formal monitoring activities, but could possibly be mobilised to regularly and actively contribute to biodiversity monitoring when such an activity involves interaction patterns comparable to the informal activities analysed here, which is the case for many citizen science biodiversity monitoring programmes.

Exact quantifications of the potential size of these embryonic citizen science communities, the mobilisation potential and the potential number of additional biodiversity observations sourced through these communities will require not only larger samples, but also an engagement with the analysed communities through direct surveys. Precise estimates are further complicated by the lack of exact numbers on the actual sample coverage of Tweets obtained through the public Twitter APIs in general and require computationally more resource-intensive directions to improve the thematic, geographic and temporal coverage

and access to this data. Finally, in estimating the potential number of observations and contributors we have to take into account other social media channels as well, such as Facebook, Flickr or Instagram, and would have to include other languages and regions rather than the exclusively English language search terms used for this study. This sketches not only the technical challenges that need to be addressed for operational applications, but also highlights the potential of the presented approach given the abundance of social media channels, users and data.

5 Conclusions

Biodiversity observations posted on Twitter and conversations with taxonomic determinations triggered by those posts appear to provide a rich, real-time data source of good quality and containing core characteristics of comparable data provided in related citizen science projects.

We can state that observational data characteristics of the "tweeted" observations and the triggered determination conversations show all elements that would be found in comparable citizen science project data. The reporting latency is low, images provide a reliable determination basis leading to conversations that produce determinations of good quality and have to offer interesting additional meta-data. The lack of detailed and reliable geo-location information stands out as a significant weakness though. We elaborated however that there is reason to believe that this could easily be alleviated. In addition, a unique feature of Twitter or similar social media tools as a data source for ecological observations is that they come with a communication channel built in, thus if the observations and determinations were to be used as monitoring data, the associated social media accounts offer a convenient way to immediately and directly follow up with the users providing the original observations.

Generally, we can conclude that a large pool of individuals with access to GPS-enabled mobile devices, no current documented but apparently more than casual interest in biodiversity observations are actively carrying these biodiversity observations into their respective social media networks, and could thus make an important active contribution to general or targeted citizen science biodiversity monitoring initiatives, both in providing and validating observations. Hence, in terms of the activity type, the contributed data and the

type of participants the analysed Twitter conversations may well be termed “embryonic citizen science communities”, which merit a further exploration and have to offer practical applications for ecological monitoring and citizen science activities.

6 Acknowledgements

The idea for this contribution was developed while the first author was staying as a guest researcher at the Stockholm Resilience Centre (SRC), which has generously supported the author with access to an excellent research infrastructure. Many conversations with colleagues at the SRC have contributed to this research. This support is gratefully acknowledged. Furthermore, the authors would like to thank Matthias Albert, Klaus von Gadow, Kevin Holston and Juan Rocha for their valuable feedback on the manuscript.

7 Author contributions

The authors jointly developed the idea for the presented approach. The first author implemented the data collection and analysis tools, collected and analysed the data and wrote the first draft of the manuscript. Both authors then jointly restructured and improved the manuscript during several major revisions. Both authors read and approved the final manuscript.

8 References

1. Xu K, Li J, Song Y. Identifying valuable customers on social networking sites for profit maximization. Expert Syst Appl. 2012;39: 13009–13018. doi:http://dx.doi.org/10.1016/j.eswa.2012.05.098

2. Bollen J, Mao H, Zeng X. Twitter mood predicts the stock market. J Comput Sci. 2011;2: 1–8. doi:doi:10.1016/j.jocs.2010.12.007

3. Sobkowicz P, Kaschesky M, Bouchard G. Opinion mining in social media: Modeling, simulating, and forecasting political opinions in the web. Gov Inf Q. 2012;29: 470–479. doi:doi:10.1016/j.giq.2012.06.005

4. Crooks A, Pfoser D, Jenkins A, Croitoru A, Stefanidis A, Smith D, et al. Crowdsourcing urban form and function. Int J Geogr Inf Sci. Taylor & Francis; 2015; 1–22. doi:10.1080/13658816.2014.977905

5. Vieweg S, Hughes AL, Starbird K, Palen L. Microblogging during two natural hazards events: what twitter may contribute to situational awareness. Proceedings of the 28th international conference on Human factors in computing systems - CHI '10. New York, New York, USA: ACM Press; 2010. p. 1079. doi:10.1145/1753326.1753486

6. Qu Y, Huang C, Zhang P, Zhang J. Microblogging after a major disaster in China: a case study of the 2010 Yushu earthquake. Proceedings of the ACM 2011 conference on Computer supported cooperative work - CSCW '11. New York, New York, USA: ACM Press; 2011. pp. 25–34. doi:10.1145/1958824.1958830

7. Sakaki T, Okazaki M, Matsuo Y. Earthquake Shakes Twitter Users : Real-time Event Detection by Social Sensors. WWW '10 Proceedings of the 19th international conference on World wide web. ACM; 2010. pp. 851–860. doi:10.1145/1772690.1772777

8. Crooks A, Croitoru A, Stefanidis A, Radzikowski J. #Earthquake: Twitter as a Distributed Sensor System. Trans GIS. 2013;17: 124–147. doi:10.1111/j.1467-9671.2012.01359.x

9. Mykhalovskiy E, Weir L. The Global Public Health Intelligence Network and early warning outbreak detection: a Canadian contribution to global public health. Can J public Heal Rev Can santé publique. 2006;97: 42–4. Available: http://www.jstor.org/stable/41994676

10. Salathé M, Bengtsson L, Bodnar TJ, Brewer DD, Brownstein JS, Buckee C, et al. Digital Epidemiology. Bourne PE, editor. PLoS Comput Biol. 2012;8: e1002616. doi:doi:10.1371/journal.pcbi.1002616

11. Gomide J, Veloso A, Meira W, Almeida V, Benevenuto F, Ferraz F, et al. Dengue surveillance based on a computational model of spatio-temporal locality of Twitter. Proceedings of the ACM WebSci'11, June 14-17 2011, Koblenz, Germany. 2011. pp. 1–8. doi:doi:10.1145/2527031.2527049

12. De Longueville B, Smith RS, Luraschi G. "OMG, from here, I can see the flames!": a use case mining Location Based Social Networks to acquire spatio-temporal data on forest fires. Proceedings of the 2009 International Workshop on Location Based Social Networks - LBSN '09. New York, New York, USA: ACM Press; 2009. pp. 73–80. doi:10.1145/1629890.1629907

13. Galaz V, Crona B, Daw T, Bodin Ö, Nyström M, Olsson P. Can web crawlers revolutionize ecological monitoring? Front Ecol Environ. Ecological Society of America; 2010;8: 99–104. doi:10.1890/070204

14. Daume S, Albert M, von Gadow K. Forest monitoring and social media – Complementary data sources for ecosystem surveillance? For Ecol Manage. 2014;316: 9–20. doi:10.1016/j.foreco.2013.09.004

15. Cha Y, Stow CA. Mining web-based data to assess public response to environmental events. Environ Pollut. 2015;198: 97–9. doi:10.1016/j.envpol.2014.12.027

16. Malcevschi S, Marchini A, Savini D, Facchinetti T. Opportunities for Web-Based Indicators in Environmental Sciences. Perc M, editor. PLoS One. 2012;7: e42128. doi:10.1371/journal.pone.0042128

17. Stafford R, Hart AG, Collins L, Kirkhope CL, Williams RL, Rees SG, et al. Eu-social science: the role of internet social networks in the collection of bee biodiversity data. PLoS One. 2010;5: e14381. doi:10.1371/journal.pone.0014381

18. Silvertown J. A new dawn for citizen science. Trends Ecol Evol. 2009;24: 467–471. doi:10.1016/j.tree.2009.03.017

19. Newman G, Wiggins A, Crall A, Graham E, Newman S, Crowston K. The future of citizen science: emerging technologies and shifting paradigms. Front Ecol Environ. Ecological Society of America; 2012;10: 298–304. doi:10.1890/110294

20. Biggs R, Carpenter SR, Brock WA. Turning back from the brink: detecting an impending regime shift in time to avert it. Proc Natl Acad Sci U S A. 2009;106: 826–31. doi:10.1073/pnas.0811729106

21. Cooper CB, Dickinson J, Phillips T, Bonney R. Citizen science as a tool for conservation in residential ecosystems. Ecol Soc. 2007;12. Available: http://www.ecologyandsociety.org/vol12/iss2/art11/

22. Kaplan AM, Haenlein M. Users of the world, unite! The challenges and opportunities of Social Media. Bus Horiz. 2010;53: 59–68. doi:doi:10.1016/j.bushor.2009.09.003

23. Smith A, Brenner J. Twitter Use 2012 [Internet]. 2012. Available: http://pewinternet.org/Reports/2012/Twitter-Use-2012.aspx

24. See L, Comber A, Salk C, Fritz S, van der Velde M, Perger C, et al. Comparing the quality of crowdsourced data contributed by expert and non-experts. Preis T, editor. PLoS One. Public Library of Science; 2013;8: e69958. doi:10.1371/journal.pone.0069958

25. Crall AW, Newman GJ, Stohlgren TJ, Holfelder KA, Graham J, Waller DM. Assessing citizen science data quality: an invasive species case study. Conserv Lett. 2011;4: 433–442. doi:10.1111/j.1755-263X.2011.00196.x

26. Butt N, Slade E, Thompson J, Malhi Y, Riutta T. Quantifying the sampling error in tree census measurements by volunteers and its effect on carbon stock estimates. Ecol Appl. Ecological Society of America; 2013;23: 936–943. doi:10.1890/11-2059.1

27. Smith AM, Lynn S, Sullivan M, Lintott CJ, Nugent PE, Botyanszki J, et al. Galaxy Zoo Supernovae. Mon Not R Astron Soc. 2011;412: 1309–1319. doi:10.1111/j.1365-2966.2010.17994.x

28. Hochachka WM, Fink D, Hutchinson RA, Sheldon D, Wong W-K, Kelling S. Data-intensive science applied to broad-scale citizen science. Trends Ecol Evol. 2012;27: 130–7. doi:10.1016/j.tree.2011.11.006

29. Van Strien AJ, van Swaay CAM, Termaat T. Opportunistic citizen science data of animal species produce reliable estimates of distribution trends if analysed with occupancy models. Devictor V, editor. J Appl Ecol. 2013;50: 1450–1458. doi:10.1111/1365-2664.12158

30. Reichman OJ, Jones MB, Schildhauer MP. Challenges and Opportunities of Open Data in Ecology. Science (80-). 2011;331. doi:10.1126/science.1197962

31. Sheppard SA, Wiggins A, Terveen L. Capturing quality: retaining provenance for curated volunteer monitoring data. Proceedings of the 17th ACM conference on Computer supported

cooperative work & social computing - CSCW '14. ACM Press; 2014. pp. 1234–1245. doi:10.1145/2531602.2531689

32. Otegui J, Ariño AH, Encinas MA, Pando F. Assessing the Primary Data Hosted by the Spanish Node of the Global Biodiversity Information Facility (GBIF). Raghava GPS, editor. PLoS One. 2013;8: e55144. doi:10.1371/journal.pone.0055144

33. Crowl TA, Crist TO, Parmenter RR, Belovsky G, Lugo AE. The spread of invasive species and infectious disease as drivers of ecosystem change. Front Ecol Environ. 2008;6: 238–246. doi:10.1890/070151

34. Shirk JL, Ballard HL, Wilderman CC, Phillips T, Wiggins A, Jordan R, et al. Public Participation in Scientific Research: a Framework for Deliberate Design. Ecol Soc. 2012;17: art29. doi:10.5751/ES-04705-170229

35. Wiggins A, Crowston K. Goals and Tasks: Two Typologies of Citizen Science Projects. 2012 45th Hawaii International Conference on System Sciences. IEEE; 2012. pp. 3426–3435. doi:10.1109/HICSS.2012.295

36. Daume S. Ecoveillance (Social media analysis web platform) [Internet]. 2012. Available: http://www.ecoveillance.org

37. Lindsey Kuper (@lindsey). "Saw this beautiful iridescent green bug today. Anyone know what it is? https://www.flickr.com/photos/lindseykuper/14360920608/ ...", 30 June 2014, 8:04 pm. Tweet. [Internet]. 2014. Available: https://twitter.com/lindsey/status/483808338015047680

38. Croitoru A, Crooks A, Radzikowski J, Stefanidis A, Vatsavai RR, Wayant N. Geoinformatics and Social Media: A New Big Data Challenge. In: Karimi HA, editor. Big Data Techniques and Technologies in Geoinformatics. Boca Raton, FL: CRC Press; 2014. pp. 207–232.

39. Lindenmayer DB, Likens GE. The science and application of ecological monitoring. Biol Conserv. Elsevier Ltd; 2010;143: 1317–1328. doi:10.1016/j.biocon.2010.02.013

40. Gärdenfors U, Jönsson M, Obst M, Wremp AM, Kindvall O, Nilsson J. Swedish LifeWatch — a biodiversity infrastructure integrating and reusing data from citizen science, monitoring and research. Hum Comput. 2014;1. doi:10.15346/hc.v1i2.6

41. Davies L, Bell JNB, Bone J, Head M, Hill L, Howard C, et al. Open Air Laboratories (OPAL): a community-driven research programme. Environ Pollut. 2011;159: 2203–10. doi:10.1016/j.envpol.2011.02.053

42. Wood C, Sullivan B, Iliff M, Fink D, Kelling S. eBird: engaging birders in science and conservation. PLoS Biol. 2011;9: e1001220. doi:10.1371/journal.pbio.1001220

9 Supporting information

Supporting Information 1. List of phrases used to identify Tweets that qualify as determination requests.

Determination request phrases
“anyone know what”
“anybody know what”
“anyone know which”
“anybody know which”
“what this is”
“what species”
“what is this”
“what kind it is”
“know what kind of”

Supporting Information 2. List of the 215 analysed Tweets. In accordance with Twitter terms and policies on data sharing the Tweet information is limited to the Tweet identifiers.

310264486902767616, 320648466181279744, 322155392723464192, 331960333969264640, 336284534687555584, 338498329430339584, 338937212651585536, 339750583898607616, 340580859243536384, 341708939928408064, 341949563638272000, 342630306228088832, 342836020653457408, 343808856654233600, 346023045506404352, 347897418840416256, 348155379756183552, 348618973526491136, 350089262949076992, 350317679443324928, 351379998931832832, 352435125943930880, 354297229592375296, 354305613397893120, 354385207564070912, 354670829759500288, 356409976677343232, 359839316261879808, 360702577253445632, 360765951819534336, 361080267965874176, 362737424952070144, 363692965840957440, 364769961958309888, 364785243967877120, 366318729514078208, 366514263424118784, 366628421364748288, 366685546245935104, 367154694223052800, 367734690725126144, 367776722189836288, 368102508066922496, 368872789974847488, 370645068988035072, 370944941612367872, 371769450389061632, 372019767148052480, 373927964817502208, 374064056686350336, 375773109678514176, 379690185879220224, 379976434695294976, 381853537837129728, 382599489950740480, 384760397217013760, 386964769078390784, 386969673691975680, 387248532388057088, 390947697630670848, 391167906668904448, 396078282246221824, 399457932880117760, 409740532026732544, 441078393728008192, 443807260666769408, 448541622259974144, 450023214727442432, 451608526381936640, 454326603963973632, 461852673319239680, 461875140771725312, 466427775390609408, 469916284067807232, 470727475295956992, 472687828707442688, 472828017400492032, 474083128198582272, 475025560129638400, 475496407516721152, 476493977953509376, 477904276132737024, 478206573928787968, 478226066000322560, 478251419116793856, 480306095374618624, 481736117667241984, 482956933465653248, 483210762911875072, 483357228036935680, 484688723272302592, 484814021783089152,

484814250553008128, 485036371003080704, 488064685514948608, 488268501136982016, 488345669854576640, 492032153094717440, 492303905590874112, 492413354678681600, 494175663583002624, 494991476276084736, 497495646774247424, 498510599639142400, 499921841910476800, 501055044314476544, 501177768064090112, 501789491913564160, 503700423380041728, 505161373215895552, 509454998699540480, 511408200093286400, 513370406804529152, 516084816417402880, 516292800393277440, 526689205066694656, 527015311372214272, 529330944982134784, 529404148312465408, 537587538139967488, 543507413257945088, 544412820478361600, 340932759914684416, 341594233377009664, 341958301828915200, 342271807459581952, 342382752672391168, 343594139482484736, 343808856654233600, 346315492534927360, 346721344345501696, 347897418840416256, 348829838779502592, 352388331205697536, 352899772241281024, 353181438096572416, 353788884242550784, 355428832200765440, 358625705401126912, 359028120558383104, 360886178628636672, 365014168685912064, 366628421364748288, 367776722189836288, 368647674095874048, 370245329972506624, 372021354532306944, 374272141719371776, 374817341759975424, 374898220968984576, 379991807553466368, 384697262393532416, 386442289058299904, 386514746351947776, 393798278766419968, 394760931042791424, 401699719761514496, 443807260666769408, 443974185082056704, 447442360595542016, 450322221576503296, 463003533655412736, 463319800216035328, 463742418387357696, 464752657803247616, 464753071613296640, 467087814921048064, 467089058934841344, 468115421137485824, 468634103336931328, 474183978220781568, 474655229699301376, 474971408662294528, 476027328385253376, 476493977953509376, 477090006943956992, 477572016694767616, 477997265198804992, 478474617808777216, 479218872722391040, 481259193459306496, 481859395156844544, 483233705607974912, 483239530431270912, 483240172969279488, 483808338015047680, 484471096985808896, 484670714197606400, 484769497405194240, 484817616289013760, 485528769130954752, 487705291958792192, 487721549605007360, 488312122435465216, 488345669854576640, 494516565610471424, 495329385059454976, 496483736180256768, 497827970460160000, 499267353109336064, 500068441290973184, 500402109108744192, 501344632186609664, 502851804502425600, 505795069191131136, 506737683637747712, 509454998699540480, 514456675026493440, 515118920571695104, 515269220335370240, 521284183545888768, 522697860333469696, 532217620410671104, 534372435874033664, 534410848669016064