NERC Open Research Archive sharing our research

Centre for Ecology & Hydrology
NATURAL ENVIRONMENT RESEARCH COUNCIL

Article (refereed) - postprint

August, Tom; Harvey, Martin; Lightfoot, Paula; Kilbey, David; Papadopoulos, Timos; Jepson, Paul. 2015. **Emerging technologies for biological recording [in special issue: Fifty years of the Biological Records Centre]** *Biological Journal of the Linnean Society*, 115 (3). 731-749. 10.1111/bij.12534

Copyright © 2015 The Linnean Society of London

This version available http://nora.nerc.ac.uk/511643/

NERC has developed NORA to enable users to access research outputs wholly or partially funded by NERC. Copyright and other rights for material on this site are retained by the rights owners. Users should read the terms and conditions of use of this material at http://nora.nerc.ac.uk/policies.html#access

This document is the author's final manuscript version of the journal article, incorporating any revisions agreed during the peer review process. Some differences between this and the publisher's version remain. You are advised to consult the publisher's version if you wish to cite from this article.

The definitive version is available at http://onlinelibrary.wiley.com

Contact CEH NORA team at
noraceh@ceh.ac.uk

The NERC and CEH trademarks and logos ('the Trademarks') are registered trademarks of NERC in the UK and other countries, and may not be used without the prior written consent of the Trademark owner.

Title: Emerging technologies for biological recording

Running Title: Emerging technologies for biological recording

Authors:

Tom August

Centre for Ecology and Hydrology, Wallingford

Martin Harvey

Faculty of Science, The Open University

Paula Lightfoot

National Forum for Biological Recording

David Kilbey

IT services, University of Bristol

Timos Papadopoulos

Biodiversity Institute, University of Oxford

Paul Jepson

School of Geography and the Environment, University of Oxford

Abstract (100-200 words):

Technology has played an important role in biological recording for hundreds of years, from the invention of the microscope to the microprocessor. We review current and emerging technologies that are changing the way we study and record the natural world.

From websites to smartphones, data capture is becoming easier, faster and more accessible. Increases in data volume and wider participation raises concerns over data quality which are being addressed with accurate sensors, automated validation tools, and verification platforms that utilise expert taxonomists and collective intelligence to ensure the highest level of quality possible.

Data curation and interoperability have been transformed in the information age. The need to collate data at continental and global scales and across institutions continues to drive the formation of standardised data formats and taxonomies. Once collated these data can be analysed using high performance computing, and used to provide valuable feedback in the form of interactive visualisations, computer generated text or even game-like rewards.

We also address issues arising from these technological developments. For example: how will the role of the expert naturalist change? Is biological recording undergoing a revolutionary or evolutionary process? and how is technology leading to the empowerment of the public?

1. Introduction

Widespread access to computational and communication technology is one of the defining characteristics of our generation, prompting talk of humanity embarking on an information revolution (e.g. Sliwa & Benoist, 2011; Saylor, 2012). The implications of social media, apps, low cost sensors, search engines, and predictive analytics associated with 'big data' have profound implications for many areas of society including biological recording. We can expect a step-change in biological recording as developments in sensor technologies (LiDAR, UAVs, acoustic arrays, eDNA etc) and biodiversity informatics infrastructures generate 'big ecological data' (Snaddon *et al.*, 2013). Big data is changing the way scientists view, study, and analyse the world (e.g. National Ecological Observation Network; www.neoninc.org) and it will likely change the nature of biological recording in terms of techniques, practice and the relationship between expert and non-expert.

The processes driving the uptake of technologies in biological recording are two-fold. Firstly, the passion and curiosity for the natural world that has motivated naturalists for centuries is inspiring present day naturalists to develop and use technologies to enhance their recording activities. Secondly, for many of the pressing questions in conservation, such as identifying species at risk of extinction, structured survey data are available for only a small group of charismatic taxa (e.g. UK Butterfly Monitoring Scheme; www.ukbms.org). The next best alternative are unstructured data, which are based on ad-hoc observations. Methods have recently been developed to compensate for sampling biases associated with unstructured data which identify trends and predict species responses to environmental change (van Strien, van Swaay, & Termaat, 2013; Isaac *et al.*, 2014). When collated nationally in a standard format, biological records data can be used for assessments of extinction risk (Maclean & Wilson, 2011), impacts of climate change (Hickling *et al.*, 2006), informing emergency responses to environmental threats (Department for Environment Food and Rural Affairs, 2012), and reporting on the Convention on Biological Diversity (CBD) targets (Department for Environment Food and Rural Affairs, 2013). This has led to an increased interest from data users in the ways in which technology might be able to enhance biological recording. For example, the ecosystem approach of the revised CBD (www.cbd.int/ecosystem) requires data on multi-taxa assemblages of species as indicators of habitat condition or proxies for ecosystem services, such as pollination or soil biodiversity. Examples of projects that take on this challenge in the UK include PondNet (www.freshwaterhabitats.org.uk/projects/pondnet), the National Plant Monitoring Scheme (www.brc.ac.uk/npms/content/welcome), and SPLASH (Survey of Plants and Lichens associated with

Ash; www.brc.ac.uk/splash/home). Emerging technology plays a key role in such projects, both in engaging participants and improving standards of data quality and data flow.

The scope of this paper is on current, cutting edge, and future applications of technologies in volunteer biological recording. Technological innovations have influenced the practice of biological recording since the outset, for example field guides and distribution maps were an outcome of the print press and new cartographic techniques (Demeritt, 2001) respectively. The technological innovations we examine here are computational - hardware and software - and we include 'emerging' to flag our focus on technologies that are emerging from the interplay of cloud and mobile computing, social media and networks, sensors, informatics, and big data within institutions of biological recording. While we will include international examples, our focus will be on examples from the UK which exemplify these developments.

The explosion of technologies of the late 20th century and early C21st century have provided both opportunities and challenges for biological recording. A brief history of these developments is presented in Figure 1. We will explore these as we review the role of technology at each stage of the recording process. Specifically we will consider the quantity, quality and diversity of data afforded to us by new data capture technologies; the importance of record verification and dataset interoperability; and new ways in which technology is being used to share, analyse and engage people with data.

2. Data capture

The mechanisms for recording data have changed dramatically over the past 50 years, from recording based solely on paper to schemes that are now solely based on electronic systems. This change has been possible due to developments in hardware, such as mobile computing and the internet, and software such as databasing and HTML.

2.1 Websites

Websites allow data to be entered and shared quickly and reliably and have become an integral part of modern wildlife recording (Table 1). The ubiquity of computers in contemporary life has led paper forms to be replaced by digital forms presented via a PC or mobile web-app. These technologies afford

standardised forms with restricted data entry, drop down lists, and mapping interfaces that help assure data quality. In addition they allow the capture of any data that can be stored digitally, notably photographs. Because the need for distributing and collecting physical forms has been removed, websites have radically extended the reach, accessibility and interactivity of biological recording schemes which can now be promoted via a simple link embedded in an email, news report, or tweet. Websites may be considered by many as old technology as the first website was created over 20 years ago, however, website design is constantly evolving. The design of websites, and any other human-computer interaction, is key for the success of biological recording systems. Good design engages users, maintains their motivation and makes the process of recording and data exploration efficient and enjoyable.

Xeno-canto (www.xeno-canto.org) is a website that successfully leverages the engagement of the general public from across the globe to record and archive bird vocalisations with minimal needs in funding and centrally invested human effort. Launched less than 10 years ago and maintained by only 4 administrators, with the assistance of the xeno-canto community, this website now contains recordings from more than 9000 avian species collected by more than 2000 registered users (covering roughly the 80% of all vocalising avian species). Species labelling as well as audio quality characterisation of each individual recording is done based on a peer assessment scheme. Collecting digital data on a global scale, as in this example, presents exciting possibilities, and would not have been possible without modern technologies.

Most websites invite users to register and provide basic information about themselves. This generates secondary data relating to the recorders use of the website. The Zooniverse project (www.zooniverse.org), which deals with data classification rather than capture, is using such data to generate algorithms that profile users and automatically present tasks suited to their skill and commitment levels, in much the same way Google targets adverts. Such approaches could be applied to large-scale biological recording schemes in the future.

Websites, such as Zooniverse, can also themselves become a place for data generation. An ever increasing number of projects take advantage of 'cognitive surplus' (Shirky, 2010), the idea that for much of our spare time we do not make use of our cognitive capacity. These projects use this surplus to undertake useful tasks. One task that humans are particularly good at, which computers are not, is pattern recognition. A number of the most successful citizen science projects have developed websites to harness this skill. Zooniverse projects (e.g. Snapshot Serengeti and Seafloor explorer) have attracted over 1.2 million users, generating data whose analysis has led to numerous scientific publications (e.g. Lintott *et al.*, 2008). Other websites seek to encourage volunteers to digitise physical records, for

example Herbaria@Home (www.herbariaunited.org/atHome) has led to the digitisation of over 142,000 herbarium specimens by volunteers, while the Atlas of Living Australia's DigiVol project (www.volunteer.ala.org.au) has expanded this idea to other taxonomic groups and successfully engaged over 700 volunteers.

Typically seen as a portal through which data is submitted or viewed, biological recording websites can also provide tools to aid recorders. Multi-access keys are a notable example with great future potential. In contrast to traditional dichotomous keys, users of multi-access keys can consider the morphological characters of a specimen in any order. One of the benefits of this approach is that users can avoid characters that are absent on their specimen or which they are not confident identifying, however, multi-access keys are very hard to produce in paper form. Computer-based multi-access keys take advantage of the flexibility of a graphical user interface and can incorporate elements absent from paper versions. For example, probabilistic species matches can be provided at any point in the key, making use of information such as the location or date of the identification (Burkmar, 2014). Given the benefits of multi-access keys the uptake of this approach has been slow (Morrison, 2011), in part because the technology required (i.e. well designed user-interfaces and portable hardware that can be taken into the field) has been absent, but also because of loyalty to existing paper-based guides and single-access keys (Burkmar, 2014). These technological challenges have now been overcome and the field is advancing rapidly (Nimis & Lebbe, 2010), with standards in place - the Structured Descriptive Data (SDD; http://wiki.tdwg.org/SDD) - to ensure interoperability between systems. The potential to take tablet computers into the field, loaded with tens or hundreds of guides, user friendly multi-access keys, and with access to online libraries of photos, sounds and videos, is exciting and will result in increased quality of data gathered and widened participation in biological recording.

2.2 Smartphones

In the fifth wave of computing (Figure 1) users became liberated from their work stations, being able to carry around computers in their pockets (Saylor, 2012). The first mobile computers were the PalmPilots (PDAs) launched by Hewlett Packard, which were applied to biological recording in a limited way, notably Cybertracker (www.cybertracker.org/background/our-story). However, the introduction of smartphones represented a step change in mobile computing. Since their introduction 20 years ago smartphones, and more recently tablets, have undergone a meteoric rise in popularity and have become almost ubiquitous in the UK: an estimated 72% of UK consumers aged 16-64 owned a smartphone device in 2013 a rise of 14% since the previous year (Deloitte, 2013). The latest generation

of smartphones integrate enhanced sensors and computational power with the capacities of cloud computing, big-data, social networking, crowdsourcing and the human user. 'Apps', small task oriented programmes, create this integration and have proliferated exponentially since the launch of Apple's iTunes App store (2008) and Google's Play Store (2009).

The wide adoption of smartphones in the developed world has presented some unique opportunities for biological recording by facilitating data entry in the field and providing built in sensors. The sensors used most in biological recording are: GPS chipsets for deriving accurate location, camera, and a microphone. Some newer phones also feature sensors to measure temperature, humidity, air pressure and movement (gyroscopes). These provide opportunities to add valuable metadata to recordings, however it is important to capture uncertainty from these sensors since they may be imprecise and inaccurate.

Scientists are harnessing the ubiquity and functionality of smartphones by designing apps to facilitate biological recording by experienced biological recorders but also by members of the public with only a limited or casual interest in recording. Examples include PlantTracker (http://planttracker.naturelocator.org/) for crowd sourcing location data on invasive plant species. Users photograph and geo-locate the specimen and then submit the record to a database to be verified by experts. The New Forest Cicada Hunt app (http://www.newforestcicada.info/) deploys the capacity of smartphone microphones to detect the near ultrasonic call of the UK's only native cicada, and submits an audio recording and sonogram should something likely be detected, placing itself at the cutting edge of biological recording apps.

The way in which apps are produced is changing too. With the continuous evolution of web standards there has been an increase in hybrid and pure web apps, which facilitate a "write once, deploy anywhere" approach. This is in contrast to natively coded apps, which are specific to a particular phone platform. This is making apps both more economical to produce and easier to maintain although currently with some compromise in functionality and overall user experience. Using this approach a number of apps have been developed for biological recording which work across multiple platforms such as the Nature Locator family of apps (www.naturelocator.org).

Pure web apps have become more feasible options owing to the increased capabilities of the HTML5 standard. HTML5 is the latest iteration of the language used to create websites and offers new features such as support for touch based interaction, as is common on touch screen devices. These apps are run from the host's server and have both advantages and disadvantages compared with native and hybrid apps. Among the pros include the ability to update the app website and for these

updates to be instantly visible to all users. Additionally, since pure web apps run from a server, they can use the increased computing power this affords to undertake complex tasks such as sophisticated sound and image recognition. However, these types of app have no App Store presence nor do they yet have access to some smartphone features such as touch gestures or access to certain hardware. These apps also suffer from a limitation that affects many smartphone apps, the need for an internet connection.

Many apps navigate the issue of unreliable internet connectivity by storing content on the device rather than loading it from a server, however the storage space on smartphones is limited. Additionally many apps allow users attempting to submit records without an internet connection to store them on their phones for submission later. While these two features address the main issues of reduced connectivity in rural areas, improved connectivity would allow for a greater diversity of features in these apps.

2.3 Repurposed technology

The data capture technologies we have considered so far are designed with a top-down approach. They are built by organisations or individuals who desire data, with the intent that volunteers will use the technology to collect data for them. Communities are now emerging across the globe that seek to develop technologies that address local concerns, a bottom-up development process.

Through the use of repurposed technologies - technology that has been adapted for a purpose other than that for which it was originally designed - it is possible for citizen science projects that are technology dependent to become independent of institutions who would previously have provided the technology needed. This provides the opportunity for technology-dependent citizen science independent of industry or academia, however, co-development of projects is often advantageous to both parties. This movement is aided by the relatively low cost of electronics and computers, and online communities that are dedicated to supporting and sharing information on repurposing technologies (e.g. instructables; www.instructables.com and The Public Laboratory for Open Technology and Science; www.publiclab.org).

Common technologies that are repurposed include digital cameras, UAVs (unmanned aerial vehicles) and computers. Projects in this area have included documenting the deep horizon oil spill using modified digital cameras attached to kites and balloons (The Public Laboratory for Open Technology and Science; www.publiclab.org/wiki/gulf-coast), monitoring pollution using cheap open source

sensors and electronics platforms (FLOAT; www.f-l-o-a-t.com) and monitoring water quality (The Public Laboratory for Open Technology and Science; www.publiclab.org/wiki/open-water), amongst others.

There are few repurposed technologies that have been constructed for biological recording, however a number of opportunities exist. There are projects which have developed microscopes from webcams which could help identification of small organisms, cheap camera traps can be constructed from digital cameras and infrared sensors, and advances in UAV technology provide opportunities for low cost, high resolution, imagery for mapping species occurrence (Martin *et al.*, 2012).

2.4 Community creation and support

The social aspect of biological recording is a key motivation for many volunteers and has resulted in vibrant recording communities in many parts of the world. However, communities of recorders can often be sparsely distributed and often form around geographic regions such as towns/cities or counties/states. This can result in high spatial variability in recording activities.

Many recording communities now have websites which are a centre point for organising events, publishing news and promoting recording. These can help to unify a community in a formal way by promoting common projects such as creating regional atlases, and activities such as trips or conferences.

While websites can help to unify a community they are often used as a notice board rather than a place for social interactions. Instead, social networking sites have become a useful tool for enabling easy communication between like minded members of the recording community. In the 1990s biological recorders were quick to adopt the Yahoo! system of email discussion groups, with a plethora of groups being set up for different taxonomic groups and interests, sharing information and allowing sightings of interesting species to be rapidly shared. Many of these groups are still active today. More recently naturalists have been quick to adopt new social media including Facebook and Twitter. Social media offers users the ability to form their own communities and communicate with individuals or groups easily and instantaneously (Box 1). It also offers established organisations the ability to communicate easily with their members in a more interactive and less formal way. For example Butterfly Conservation has an estimated 24,100 followers on Twitter, and organisations such as local bat groups use Facebook groups to organise events and have discussions.

Box 1 - Garden Bioblitz is an example of the impact of social media on biological recording. In a blog post in 2011 a UK-based naturalist described how she had carried out a 'bioblitz' survey of her garden, recording as many species as possible. Her experience was shared via Twitter, which led a number of people with similar interests to set up a trial project in 2012 using Twitter as the focus for a larger group of people to take part (Comont, 2013). The success of this trial led to the launch of a nationwide Garden Bioblitz in 2013, which has continued to use Twitter and Facebook to engage more people to take part, with online resources including iRecord (www.brc.ac.uk/irecord) and iSpot (www.ispotnature.org) used to store biological records and provide help with identifications. The swift development of such a project, bringing together novice and experienced wildlife recorders from many widely dispersed locations, was only possible by embracing the possibilities offered by social media.

Some biological recording websites also include social elements to gain the best of both worlds. For example both iSpot and iNaturalist (www.inaturalist.org), allow users to submit photos of sightings, comment on each others' photos as well as give their own taxonomic identifications to other people's observations.

3. Data management and quality assurance

The biological recording community has always been quick to adopt new technologies to help manage data: spreadsheets and databases have long formed the backbone of data management for recording schemes. Recent and emerging technologies provide further possibilities for increasing efficiency in data management through shared online systems and automated approaches. However, while some of these new approaches are making it easier to capture greater amounts of data more quickly and from a wider range of recorders, the need remains for such data to be checked to ensure that its quality is known and documented. Not all data uses require high quality data, especially if data volume is large (see section on analysis). However, for many questions data of a known quality is required and as such metadata is important. Data quality can be quantified, and in some cases improved through quality assurance steps.

Quality assurance for biological records has largely been carried out by people working or volunteering for local records centres and for national and local recording schemes. The terms "validation" and

"verification" are often used to refer to, respectively, carrying out standardised checks on the completeness and legitimacy of a record's contents, and ensuring the accuracy of the recorded location and taxon name (James, 2011).

Some aspects of quality assurance can be automated, while others require significant input from experts. The extent to which emerging technologies will be able to augment and assist in quality assurance is not yet clear, but there is scope for them to play a greater role in the future. However, it should also be recognised that expert verifiers also use the verification process as a means of communicating with, educating, and enthusing the recorders who provide data. Automated procedures can provide help to human verifiers, but are unlikely to replace them.

3.1 Automation

3.1.1 Automated location data

In field location data has previously relied on the observer's map reading skills and can be a source of error. Many websites for biological records offer the recorder an interactive map so that they can provide the location by simply clicking on a map, thus avoiding the need for a detailed understanding of coordinate systems. In addition fully automated location recording is available from many mobile devices containing GPS. This is a welcomed advance, however it is important that the GPS accuracy is also captured since this can vary greatly depending on habitat structure and device.

3.1.2 Automated validation and verification tools

Using our understanding of organisms' traits (e.g. distribution, phenology, etc) tools are now available to automatically detect unusual records, helping to direct the effort of experts whose time is often in short supply. For example, NBN Record Cleaner (www.nbn.org.uk/Tools-Resources/Recording-Resources/NBN-Record-Cleaner.aspx) is an automated validation and verification decision-support tool for recorders and biodiversity data managers. It is designed to improve the efficiency of data flow and ensure the quality of biodiversity datasets by enabling automated checking of large datasets in a variety of formats against validation and verification 'rule sets' developed by national taxonomic experts. Verification rule sets flag up records that fall outside the known temporal or spatial distribution of that species, as well as highlighting records of species that are inherently difficult to identify. The human verifier can then choose whether or not to accept these 'outlier' records, or to go

back to the original recorder to seek further evidence for the record. A similar system has been developed by eBird (www.ebird.org) which uses similar rules to highlight unusual records, and many others are documented by the Global Biodiversity Information Facility (GBIF; www.gbif.org.

Published rulesets are defined at a point in time, but in the future it would be possible to facilitate live updating of some of the rules as new data arrives. For instance, fixed rules based on the known range of the species can go out of date quickly as ranges change, but in principle geographical rules could be based on the current set of verified records, updating them in real-time as further records are verified.

3.1.3 Automated observation and identification

Automated collection of images and sound recordings for biological recording is becoming increasingly familiar. Work using camera traps has traditionally focussed on mammals (Trolliet *et al.*, 2014), but is also being extended to other taxa, including insects, for example via the Rana system (http://www.tumblingdice.co.uk/rana/detecting-pollinators). Different types of microphones have been used to record bats (Walters *et al.*, 2012), marine mammals (Mellinger *et al.*, 2007), birds (e.g. Zwart *et al.*, 2014) and insects (e.g. Chesmore & Ohya, 2004). These approaches can produce images and sounds alongside metadata such as date, time and often location. Human input is usually required to identify the species that have been recorded, but this aspect is also becoming amenable to automation.

Arguably, one of the most pertinent examples of automated detection and identification of biological sounds is the case of birdsong recognition. Recordings of birdsong have become increasingly available due to the widespread interest of the general public, the existing networks of bird enthusiasts, and the fact that bird vocalisations reside largely in the same frequency region as human speech (and hence can be recorded with cheap voice recorders and mobile devices). This has resulted in increased research interest in the application of machine learning methods to the problem of automatic bird sound recognition. Early research works on the subject in the 1990s made use of simple pattern matching techniques from the early days of speech processing (e.g. Anderson, Dave, & Margoliash, 1996). The similarities of birdsong to human speech (e.g. Doupe & Kuhl, 1999) led the research area to grow based on methods and findings from the speech processing literature through the 2000s. This included the use of audio features originally devised for human speech recognition, the use of Hidden Markov Models and in most cases the need of a manual preprocessing syllable segmentation stage for the training of the proposed learning algorithms (for a review see Stowell & Plumbley, 2011). More recently, the severe practical limitations associated with manual syllable segmentation as well as the

need for classification algorithms that can discriminate between hundreds of different bird species labels for which no dictionaries of phonemes/syllables exist has resulted in the gradual departure from speech recognition methods in favour of a wider toolset of probabilistic machine learning methods (e.g. Briggs *et al.*, 2012; Stowell & Plumbley, 2014). A significant advantage of structured probabilistic models is the ability to provide human users with an ordered list of most likely results thus allowing the crowd-sourcing of human corrective input to the automatic identification scheme.

An increasing number of online sound recognition competitions are organised as part of international conferences and other initiatives. This is as a result of increasing interest in designing biological sound recognition algorithms, using machine learning, to analyse audio obtained in realistic conditions. Such competitions attract worldwide entries from several tens of competing design teams at a time. They facilitate agile forms of knowledge and expertise sharing and provide a constant update on the state of the art in the field. The most prominent examples of recent competitions include the UCR Insect Classification Contest in 2012 (www.cs.ucr.edu/~eamonn/CE/contest.htm), the Multi-label Bird Species Classification - NIPS 2013 (www.kaggle.com/c/multilabel-bird-species-classification-nips2013), the Machine Learning for Signal Processing 2013 Bird Classification Challenge (www.kaggle.com/c/mlsp-2013-birds) and the two bird sound and one whale sound competitions organised as part of the International Conference on Machine Learning 2014 (sabiod.univ-tln.fr/ulearnbio/challenges.html)

As a result of this research activity, various computer programmes and smartphone apps have emerged that attempt to identify species from images or sounds, thus opening the way for complete automation of the biological recording process – species, location and date/time. The Leafsnap app (www.leafsnap.com) utilises algorithms based principally around leaf shape to aid identification by refining the number of possible matches a user needs to consider (Kumar *et al.*, 2012), while the Echometer touch app (www.wildlifeacoustics.com/products/echo-meter-touch) and the iBatsID web application (www.sites.google.com/site/ibatsresources/iBatsID) aim at identifying bat species from their calls. Wildlife sound identification computer software and mobile apps that are based on proprietary algorithms are offered by private companies such as SoundID (www.soundid.net) and Isoperla (www.isoperla.co.uk). Other examples of recent and ongoing projects include the crowdfunded project Warblr (www.warblr.net) and the BioSound project (http://oxlel.zoo.ox.ac.uk/research/projects/biosound/) both on bird song identification, as well as the Google Impact Challenge funded collaboration between the Royal Botanic Gardens Kew and the University of Oxford (https://impactchallenge.withgoogle.com/uk2014/charity/kew) which aims at

crowd-sourcing data to help prevent mosquito-borne diseases. The true impact of this proliferation of automated identification applications is yet to be evaluated.

3.2 Crowd-sourcing

Biological recording employed 'crowd-sourcing' long before the term had been coined. Online technologies have enabled this collaborative approach to evolve further, allowing people from all over the world to communicate more quickly (Box 2).

> Box 2 - In October 2009, a 6-year-old girl found an unusual-looking moth in her house in Berkshire, UK. To seek help with identification her father posted a photograph on to the iSpot website, developed by The Open University (www.ispotnature.org/node/7407). Within an hour a provisional identification had been suggested: *Pryeria sinica* Moore (Lepidoptera: Zygaenidae), a species not previously recorded in Europe. Within 24 hours this had been confirmed by experts from the Natural History Museum in London, and their colleagues in Taiwan, within the moth's native range (Ansine, 2013). It has been suggested that the larvae of the moth may have been imported on the moth's food-plant *(Euonymus spp*.). This chain of events demonstrates the potential of crowd-sourcing biological recording.

There is no doubt that online interactions enable people to learn from 'the wisdom of the crowd' (Galton, 1907) when seeking help with species identifications. However, there is also plenty of evidence that not all such wisdom is accurate. For the less familiar taxon groups expert knowledge may be confined to few people within the crowd, and it is not easy to assess whether identifications suggested online are correct. For this, input from experienced naturalists is invaluable. The iSpot project addresses this issue by using social media to put novices in touch with experts, enabling them to learn about species identification, gain experience themselves, and go on to help others within the crowd. Additionally, iSpot uses a 'reputation score' to indicate how much experience each participant has in identifying species within broad taxonomic groups (Silvertown *et al.*, 2015). Participants can suggest identifications, and if others agree with their identification they receive an addition to their reputation score – the amount added to the score depends on the amount of reputation that has already been assigned to the person agreeing with the identification. The reputation scores are indicated by icons on the website, and thus it is possible to make a judgement about how much weight of experience may be behind any particular suggested identification.

The combination of web technology and mobile apps allows a novice naturalist observing wildlife in the field to have almost immediate access to expertise from a community of users who may in principle be based anywhere in the world (Scanlon, Woods, & Clow, 2014), providing new opportunities for learning and sharing information. A challenge associated with crowd-sourcing verification is false-positives; instances where a species identification is verified when it is in fact incorrect. Were this type of error to be prevalent we would expect a weak relationship between the number of identifications per species and their difficulty to identify. When results from iSpot were tested there was a strong relationship between the number of observations of a species and its difficulty of identification (Figure 2), suggesting that these false-positives are not common. However, it also raises a concern that increasing use of crowd-sourced data based on photographs of wildlife will bias species recording datasets towards those species that are easy to identify, along with other potential biases towards more conspicuous and colourful species and those that are easier to photograph.

Crowd-sourced identification such as that used by iSpot and Project Noah (www.projectnoah.org) encourage people to learn how to identify species, and as a result biological records are accumulated. More direct online systems to collate records are provided by sites such as iRecord and eBird whose focus is directed primarily at data collection rather than education. These systems allow records for any taxon at any location to be entered online and on smartphone apps, often with accompanying photos. Verification is typically provided by expert verifiers with local expertise in species' ecology and identification.

3.3 Data centres

Increased participation in biological recording as a result of technological innovation is generating large volumes of digital data. While this does not currently pose a problem in terms of physical memory space, it is more important than ever to curate digital data to ensure it is as accessible as possible to those who might want to use it, including those who helped to collect the data. Interoperability with other datasets and future-proofing against technological changes and increasing volumes of data is also key.

New data types such as digital photographs and sound recordings (e.g. bat calls and bird song), have required modifications to databases and in the near future we should expect genetic sequences to become a common addition to records. eDNA analysis is likely to increase the volume of biological

records data and have an impact on biological recording more generally, this is discussed in another paper in this special issue (Lawson-Handley, in press).

Despite the increasing number of biological records being submitted digitally, and the increase in data that is submitted with these, such as images and sound files, data storage is not currently thought to be a major concern. For example iSpot currently holds ~300,000 observations and ~470,000 associated images which uses ~500GB of storage space. With storage costing under £50 per terabyte, this is not currently problematic, however, mobile internet connectivity and data limits currently restrict the ability to readily upload and download large files such as images and videos from mobile devices.

Standardised, quality assured, permanently archived data centres are needed to ensure longevity and security of data. Examples of permanent data archives include the Natural Environment Research Council's network of data centres (www.nerc.ac.uk/research/sites/data), GBIF, the BioFresh platform (www.freshwaterbiodiversity.eu) and GenBank (www.ncbi.nlm.nih.gov/genbank), a global repository of genetic information. While some data centres are static, others are dynamic; interfaced with a growing suite of tools for editing and managing data and metadata.

With the number of data centres increasing it is a challenge to ensure that data can be aggregated across centres if needed. One solution is to create a centralised system, all data flowing into one data warehouse, which is possible using the Indica toolkit (www.indicia.org.uk). Indicia can be used to create a centralised data warehouse which can receive data from a number of entry points. These include web forms that can be implemented on the websites of many different recording schemes and other organisations, as well as a variety of apps. For instance, in the UK the National Garden BioBlitz (www.gardenbioblitz.org), the PlantTracker app (planttracker.naturelocator.org), and the iRecord Butterflies app (www.brc.ac.uk/article/irecord-butterflies-mobile-app) all feed data into the same data warehouse. These projects all look very different, and most people who participate in them will be unaware that they are linked in any way. However, behind the scenes, rather than having separate data storage solutions for each of these projects, a single central data warehouse, maintained by the Biological Records Centre (www.brc.ac.uk), takes in the data from each project (Figure 3). Experts acting as verifiers can see all the data that is relevant to their taxonomic group or geographical area in one place, rather than having to deal with multiple datasets from each project. The data becomes available to recording schemes and records centres for analysis purposes, with the level of verification documented as part of the record.

Globally unique identifiers (GUIDs) enable data objects, such as taxon names, field observations or voucher specimens, to be identified and accessed via the Internet. They are an effective way of ensuring biodiversity data are easily citable and that credit is given to those responsible for the data

collection. Data centres can assign these identifiers to data they hold and some, such as GBIF, already do so. Identifiers can also be used to capture the state of a dataset at a set point in time, ensuring that results of analyses are reproducible in the future. GUIDs must be unique, persistent and provide a route, such as a web address, for accessing the data object. Examples that are suitable for use with biodiversity data include Life Science IDs (LSIDs) and Digital Object Identifiers (DOIs) (Taxonomic Databases Working Group, 2011).

Data centres should not be thought of as isolated entities but considered as nodes in a network of data centres. By connecting data across data centres we can bring together complementary data and address questions that cannot be answered by any one data centre in isolation.

3.4 Interoperability

Interoperability allows data from across regions, nations, or institutions to be combined without loss in quality. Combining data is this way is necessary to monitor change in biodiversity at the scale at which policy decisions are being made such as the European Union (EU). For example, the EU water framework directive intercalibration exercise (Water Information System for Europe, 2008) sort to combine data on the ecological status of water bodies across 27 member states by ensuring comparability and consistency in classification results between members. In this case, combining data at the EU level is important for reporting against targets and informing future policy. In the EU the INSPIRE directive aims to make spatial environmental data better connected, saleable and accessible. The INSPIRE directive was transposed into UK law in 2009 as the INSPIRE regulations. Not only does interoperability afford the analysis of data over a large spatial extent, it also allows the combination of different data types, for example, distribution and climate data for species distribution modelling.

To allow data to be shared and combined efficiently standardised data exchange formats are used. These formats specify the structure and content of data and metadata, enabling data stored in a variety of formats and software to be analysed in combination. Traditionally, these exchange formats have been applied to datasets post-collection, but standardisation can now be built into the collection phase via online recording websites and apps. The most well developed format for biological recording purposes are the Darwin Core standards (Wieczorek *et al.*, 2012), which are "intended to facilitate the sharing of information about biological diversity by providing reference definitions, examples, and commentaries" and are used to facilitate the sharing of biological data worldwide by GBIF. Darwin Core Format (DwC; http://rs.tdwg.org/dwc/index.htm) uses a suite of text files, each with their own

attributes, allowing efficient representation of the various aspects of a record without duplication of information while retaining the flexibility to the add new fields. DwC has been adopted by GBIF, who have started to develop tools based on the standard, notably the Integrated Publishing Toolkit (IPT; http://ipt.gbif.org/) which provides an efficient mechanism to publish and share existing biodiversity databases in Darwin Core Archive (DwC-A) format.

The development and use of standards for biological recording is not new (e.g. Burnett, Copp, & Harding, 1995), but the rapid advances in technology and changes in the focus of policy means that existing standards need to be periodically reviewed to ensure they remain fit for purpose, and continue to facilitate the combination of data for analyses. New technologies are likely to lead to further developments in standards to cover things such as the addition of digital photographs or genetic data (Lawson-Handley, in this special issue) as part of the biological record, the explicit recording of verification decisions, and sharing of unique record identifiers that allow records to be tracked across multiple platforms. While standards clearly have advantages it is also important to note that many biological recording schemes are small enterprises with limited resources and technological expertise. For these groups implementing international standards is not a priority, and the effort required to fully implement such standards may be greater than the effort needed to reformat the data when passed on to global systems

Protocols for sharing information such as HTTP and XML, which are widely used in other disciplines have been used to collate and deliver biodiversity data over the internet. The Distributed Generic Information Retrieval (DiGIR) and BioCASe protocols were developed to retrieve structured data over the internet from independent, heterogeneous databases using Darwin Core and ABCD (Access to Biological Collections Data) schemas respectively. These protocols allow website and application developers to access biological data programmatically, opening up new possibilities for data re-use.

One of the challenges of combining biological records data is matching up taxonomies between datasets. This process has been aided by standardised taxon dictionaries. The task of coordinating and updating taxon dictionaries is extremely complex and would not be possible without modern information technology. The Catalogue of Life (CoL; www.catalogueoflife.org) holds data on over 1.5 million species, an estimated 70% of the world's biodiversity, and provides the taxonomic backbone for global partnership projects including GBIF, the IUCN Red List and the Encyclopaedia of Life. This database and others like it are vital for making datasets comparable. For example, the UK Species Inventory (UKSI), based on over 230 separate taxonomic checklists in the UK, allows data to be passed between numerous UK systems with relative ease. Additionally dictionaries allow searches based on

taxonomy, such as species within a family, as well as incorporation of synonyms (including old names and common misspellings).

Taxon dictionaries are constantly being updated by taxonomic experts and this is likely to increase as DNA sequencing becomes more pervasive in systematics and existing taxa are re-assessed. However, updates to taxon dictionaries are currently not passed on automatically to systems that use them. For example, custodians and managers of taxonomic websites could be alerted automatically when a taxonomic split or dispute occurs to help them to respond.

Both data standards and taxon dictionaries are well developed tools that can significantly increase the value of biodiversity data. However, there is a lack of awareness of standards, dictionaries, and the freely available tools to aid the collection, collation and dissemination of biodiversity which leads to wasted effort and loss of data quality. Several biological recording websites and apps have been developed which are not underpinned by a standard taxonomic dictionary, do not include georeferencing aids such as an interactive map, do not integrate automated checks for validation and verification support, and do not export data in a standard exchange format. Valid biological records can still be produced and disseminated by such websites and apps, but greater effort is required to achieve this.

Combining datasets can increase their value above the sum of their parts. This can be achieved by aggregating data from different sources and making it available via a single portal, such as the Atlas of Living Australia (www.ala.org.au). The Atlas of Living Australia combines data on regional boundaries, marine regions, species occurrences data, habitat data, and can additionally call in layers from web mapping services (WMS) or use user uploaded data. Making data available via web services, web feature services or Open Geospatial Consortium WMS, enables developers to create their own combinations of data from different sources for analysis or dissemination (e.g. Plantwise; www.plantwise.org).

Increased interoperability and data exchanges between data centres pose a challenge for tracking changes to data and avoiding duplication. Currently it is possible for a single record to be submitted to a number of different websites, apps or recording schemes, for each platform to perform bespoke data validation and verification and pass the record on to a centralised data centre such as GBIF. These actions result in duplication of records which is particularly problematic when the record has been amended in one version but not another. Creating an audit trail for biological records would address these issues and could be achieved by creating an interoperable standard for the allocation of GUIDs to observations at the point of first submission.

4. Data use

Biological recording around the world has generated hundreds of millions of observations of thousands of species, and the rate at which observations are being made is ever increasing. This wealth of data has great potential for research, education and engagement which technology is helping to realise.

4.1 Analyses

The increasing demand for policy relevant information from data collected through citizen science projects (Procter, in this special issue), has driven the development of new methods to extract meaningful information from these datasets.

The unstructured nature of much biological recording can add noise to data, hiding true changes in species' distribution and abundance. This has been a major criticism to date of using these data for analyses. Biases include variation in recorder effort per site visit, species detectability, and recorder effort, both spatially and temporally (Powney, in this special issue). Despite these challenges, opportunistic species occurrence data have been used to provide insights into the impact of climate change on species range (Hickling *et al.*, 2006), the spread of invasive species (Roy *et al.*, 2012), and assessing species extinction risk (Maes, in this special issue), amongst others.

Methods developed for analysing opportunistic species occurrence data take on two main forms. The first attempts to model the bias that is present in the data and the second tries to subsample the data, leaving only unbiased records. Recently these various methods have been reviewed (Isaac *et al.*, 2014) showing considerable variation in the ability for these methods to detect a change in species occurrence in the presence of realistic recording behaviour. In support of concerns over the use of citizen science data, Isaac *et al* showed that simple methods fail under almost any deviation from even recording. However, a number methods show great potential, being able to detect realistic changes in species occurrence in the face of significant changes in recorder behaviour across time. These methods provide opportunities to produce valuable information for the assessment of extinction risk (Maes, in this special issue), to measure progress towards conservation targets (Procter, in this special issue), and to assess drivers of change (Mason, in this special issue). Many of these methods are being made freely available in the statistical programming language R, allowing anyone to perform these analyses on their own datasets (https://github.com/BiologicalRecordsCentre/sparta). Some of the

most promising of these methods are extremely computationally intensive and have only become viable thanks to advances in high performance computing. This does, however, mean that some methods become impractical for those who do not have access to such computer systems. This may be alleviated in the future by the continued improvement of computing power and by the move to cloud based computing.

4.2 Feedback

While the motivations of recorders is varied, it is generally accepted that feedback is motivational, and the sooner it is given after a record is submitted, the better. Feedback can come in many forms, traditionally being face-to-face conversations with fellow volunteers or written feedback, new technologies have opened up the possibilities for new mechanisms for providing feedback which utilise advances in web design and computer generated text.

Online recording and identification systems, as well as the use of social media sites, have made it possible for experts and record verifiers to communicate instantaneously with a wider range of people than was possible through direct contact alone. This places demands on those people who do have the skills and experience to answer the many questions posed by beginners. This can be especially problematic for taxonomic groups that require specialist input to identify species correctly. In these cases there may be relatively few "experts" compared to "beginners", and the experts that choose to engage may feel that they unable to meet demand.

Community feedback is an important feature of recording websites that rely on identifications that are achieved by consensus. On these websites users can aid each other's learning by providing feedback on identifications, adding a social element to the online recording experience. Some sites such as iSpot encourage this behaviour by offering incentives (in the form of points and online 'badges') to users who provide feedback.

Automated, computer generated, feedback has been an area of research in computer sciences and web design for some time. Applications that use this technology range from targeted advertising and film suggestions, to computer generated news articles and stock reports. Automation of this kind allows computers to turn data into information that is engaging to users but monotonous for a human to create. Graphical feedback is perhaps the most common form and can be seen on many biological recording websites. This usually takes the form of a map with the submitted observation in the context of all other recordings of that species (e.g. Sealife tracker; www.brc.ac.uk/sealife_tracker/home), but

other examples include up-to-date phenologies (e.g. BirdTrack; http://blx1.bto.org/birdtrack/main/data-home.jsp), and current recording activity (e.g. eBird).

More complicated than generating graphical feedback is creating human readable text, termed natural language generation (NLG). This technique takes underlying data and transforms it into text that reads as though it has been written by a human. This method has already been utilised by some biological recording schemes. BeeWatch (www.bumblebeeconservation.org/get-involved/surveys/beewatch) uses NLG to provide feedback to users who submit photos. When the user gets the identification of a photograph wrong, as deemed by an expert verifier, the system automatically creates a response that thanks the user for their record and highlights the characteristics of the bees that were misidentified. Additionally the system is able to provide contextual information about the record. In a study comparing how this feedback affected the behaviour of recorders there was evidence that NLG feedback improved identification skills over time as well as increasing the number of records submitted per user (Blake *et al.*, 2012). In the future this concept could be developed to provide information relevant to the experience of the recorder, to their geographic location, or even suggest other locations or species that they might want to record.

4.3 Gamification

Games can be seen as rapid feedback environments in which the users' actions effect the feedback they get over short timescales. Some biological recording platforms have seen potential benefits in applying design elements from games in an effort to make the user experience more enjoyable and engaging. The most common elements deployed are league tables which show the 'top users' (e.g. BirdTrack), typically defined by the number of observations contributed, and badges which serve to identify that a user has accomplished a task or goal (e.g. Project Noah). Careful consideration should be given to the effect that gamification has on users' behaviour, for example game elements such reputation badges in iSpot encourage users to help each other assign species identities to photographs, whereas league tables may prompt users to try and increase the volume of data contributed at the cost of quality .

Successful gamification requires a detailed understanding of the motivation of users. It can work well when used to amplify intrinsic motivations, such as concern for the environment, or provide rewards that are desired by users, but cannot be expected to get users to undertake tasks that they do not

want to do (Deterding *et al.*, 2011). Indeed game elements may dissuade participation, particularly by experts, as the platform becomes more game-like and less tool-like (Prestopnik & Crowston, 2012).

4.4 Dissemination of information

4.4.1 Visualisation

Free access to non-sensitive records, particularly via real-time results maps, is identified as an important motivation for participants in biological recording projects (Tweddle *et al.*, 2012). Enabling volunteers to visualise and explore submitted records in the context of other data, such as previous biological records, environmental data, conservation action or interpreted data such as modelled distribution maps, may be even more motivational but is not yet widely practised. Map of Life (www.mol.org) provides such visualisations, displaying species occurrence records alongside processed data such as models of species range and checklists for ecoregions and protected areas. The portal includes the ability to discover the data behind the interpreted maps, including links to the original sources of the data. This provides an easy to use interface for members of the public to explore the available data while also having sufficient data for NGOs and researchers.

A challenge for websites wishing to allow users to explore data is to make this process as unrestricted as possible. The NBN Gateway's interactive mapping tool gives users control over what is presented and how. Users can select and display multiple layers of data about species, sites and habitats, and change the colour, transparency and resolution of these layers. However, in order to analyse data the user needs to download the records or access them via an R package (cran.r-project.org/web/packages/rnbn/index.html).

The rapid flow of data through recording systems as a result of online data recording allows real-time analysis and visualisations. The movement of migrant birds (e.g. British Trust for Ornithology; www.bto.org/volunteer-surveys/birdtrack) and butterflies (e.g. Butterfly Conservation; www.butterfly-conservation.org/612/Migrant-watch.html) are tracked using live interactive mapping systems that can display records as soon as they are contributed. Other phenological events are tracked to assess how species respond to climatic variables, and climate change more broadly (e.g. Nature's Calendar; www.naturescalendar.org.uk).

4.4.2 Open access

Data that is 'open access' is freely accessible to all at minimal cost, usually via the internet and with no restrictions on use. Calls for biological records data to be open access are not new, but open access is becoming increasingly accepted in the broader scientific community and technologies are being developed which make open access more easily achievable.

In 1995 the Coordinating Commission for Biological Recording (CCBR) estimated that there were over 2,000 organisations in the UK involved in collecting or curating biological records, but three quarters of all data were used only within the original collecting/collating organisation (Burnett *et al.*, 1995). The majority of records were paper-based, and those that were digitised were held in a range of different software and formats. Clearly the technology available today means that there is no longer any need for data to be physically isolated in 'information silos' as it was twenty years ago, yet there is still a risk that technology can be used to create barriers to data access and use, resulting in 'virtual silos'. These silos can exist when data are shared but excessive limitations are placed on their use. GBIF announced in 2014 that all data they hold will be under one of three creative commons licenses. This allows much more freedom for users of the data while still allowing data providers to prevent commercial use of data if needed. However, there is some debate over the meaning of 'commercial' and some uses of data under a non-commercial use license by charities may not be allowed (Hagedorn *et al.*, 2011).

Technology also has a crucial role to play in enabling data collators and publishers to gather statistics on how their data are being used and in ensuring that users can cite data correctly. Practical data citation protocols and mechanisms, incorporating persistent identifiers such as DOIs, are vital for encouraging more individuals and organisations to make their data openly available, as it provides a means to secure recognition and reward for their efforts (Costello, 2009; CODATA-ICSTI Task Group on Data Citation Standards and Practices, 2013). These developments indicate a move from data 'sharing' via the internet to data 'publication' in perpetuity with a citation mechanism, and ideally peer review. This is identified as an important step in providing the motivation to make biodiversity data publically available (Costello *et al.*, 2012, 2013). Data journals, such as the Biodiversity Data Journal, already provide a peer-reviewed publication mechanism for biodiversity data, facilitating access to the data and offering a means of tracking its re-use via DOI-based citation system. Additionally, Scratchpads (www.scratchpads.eu) offers anyone the ability to set up an online system for creating, collating, and publishing data using standard taxonomic dictionaries and data formats (DwC-A) to facilitate interoperability and open access to data (Smith *et al.*, 2009, 2011).

Open access should not only apply to data but also to tools and applications for biological recording. For example incorporating analytical methods within portals to biological records data greatly increases the value of the underlying data. This allows users to analyse the data in a way that answers their specific questions, which could prove valuable for local conservation practitioners, as well as national and international conservation managers. Currently there are few systems that integrate biological records data and tools for their analysis. openModeller (http://openmodeller.sourceforge.net/) is one such system, designed as a platform for environmental niche modelling it offers a range of analytical tools, as well as the ability to read in data from the GBIF (de Souza Muñoz *et al.*, 2009). Building on openModeller and many other web services including GBIF and the Catalogue of Life, BioVeL (www.biovel.eu) offers a suite of free tools for those studying biodiversity. While BioVeL and openModeller are targeted at a technical audience The Atlas of Living Australia has the ability to model species' distribution using occurrence data held about Australian species in a user-friendly web interface, more suited for members of the public. These systems required significant investments of time and funding but should be seen as the current standard for open data and analysis in biological recording.

5. Discussion

Technologies have always played an important part in documenting the world around us, from the invention of the microscope to the microprocessor. However it is only in recent times, on the back of the explosion of digital technologies of all kinds at the end of the 20th century (Figure 1), that technology has changed the landscape of biological recording at such a rate. Technologies are now challenging all elements of biological recording, from how data is collected and verified, through to data analysis and dissemination.

5.1 Revolution versus evolution

Technologies that are applied to biological recording fall on a continuum between revolutionary, creating an entirely new process, and evolutionary, using technology to better an existing process. Most technologies that we have encountered are the latter which asks the question "Are there more

opportunities to apply technologies to existing rather than new processes; or are we missing opportunities for revolutionary approaches?"

Websites and apps that allow data entry are an evolution of the traditional paper and pencil recording form, and many online recording forms still retain many of the characteristics of these paper precursors. Such developments have undoubtedly increased the accuracy and speed of data capture. Revolutionary data capture methods, for example the automated detection of cicada's using a smartphone app (The New Forest Cicada Hunt), can be harder to place into the standard biological recording pathway, and the data generated may be difficult to analyse in conjunction with data collected by more traditional methods. However, using technologies in novel ways is likely to open up opportunities to answer new questions (Lawson-Handley, in this special issue). We suggest that revolutionary ideas are likely to emerge from the interface of biodiversity science and other disciplines such as computer science, engineering and social sciences and collaborations of this nature should be encouraged.

5.2 Removing barriers to public engagement with science

Amateur naturalists have been undertaking scientific research for hundreds of years and it is perhaps only in modern times that 'science' has become the preserve of professional researchers. Technologies can provide access to literature (e.g. online keys and guides) as well as individuals and communities (e.g. scheme websites and discussion forums) that allow anyone to gain the skills and knowledge required to record the natural environment. Furthermore, technologies have recently opened the doors to those without any skills or knowledge of biological recording. For example, smart phone apps such as iRecord ladybirds (www.ladybird-survey.org/recording.aspx) do not require any existing knowledge of the taxa to participate. These apps have a great potential to engage, educate, and enthuse the public with the natural world and to raise awareness of environmental issues and active areas of scientific research. The low cost of electronics and software now also allows non-professionals to build their own equipment, or develop their own platforms, to explore areas of science that are of personal interest. While data collected by people with little background in biological recording may be more ad-hoc and of lower quality than that collect by experienced taxonomists, more widespread engagement of the public is key to influencing the public's attitude to science and nature, and to inspiring the next generation of naturalists.

5.3 Data quality and flow

The widening of participation in biological recording and the increased ease with which records can be generated has seen an increase in the number of records submitted in recent years (Tulloch *et al.*, 2013). In tandem, tools have been developed to ensure that data quality is as high as possible. This includes developments during data collection, such as GPS and image capture, and post submission, such as automated validation and expert review. Combining this flow of data with centralised data centres across the world we now have more access to high quality, high resolution (both temporal and spatial) than ever before. These data are invaluable for addressing many of the pressing questions that face humanity in the 21st Century such as the effects of climate change, the spread of invasive species, and the effect of anthropogenic activities on species, communities and ecosystem services. The immediacy with which biological data are shared is of particular importance to the study of invasive species (Roy, in this special issue). Emerging EU regulations on Invasive Alien Species will require member states to set up surveillance and rapid alert systems for agreed lists of invasive species by December 2015 (European Parliament, 2014). Online reporting of potentially invasive non-native species can help prompt swift action to prevent them becoming established (e.g. GB Species alerts, www.nonnativespecies.org/alerts/index.cfm).

5.4 Enhancement or replacement

The affordability of new technologies along with increased desire to include the wider community in research poses a challenge to the traditional roles of expert data providers (Wilson & Graham, 2013a). Biological information is a significant component in the 'democratisation' of geographic knowledge (Warf & Sui, 2010) which creates new possibilities for biological recording and ecological politics since data can be proved by anyone rather than only professionals and experts. Whilst these processes seem set to alter the relationship between expert and amateur (Freenberg, 1999) it is unlikely to mean the end of the expert. It is suggested that the complexities of technological systems are such that power and expertise will remain within small groups (Haklay, 2013; Wilson & Graham, 2013b). Indeed the same is likely to be true of taxonomic expertise which is a crucial element of biological recording that cannot be replaced and will instead grow in demand as the amount of records requiring verification increases.

Not everything can be done online of course – many species identifications still depend on physical specimens being passed to experts to check, and there are many benefits to learning about fieldwork and recording wildlife from face-to-face meetings with people in the field. However, the range of tools and resources being developed are undoubtedly extending the reach and speed of the sorts of human interaction and collective sharing of information that have always been such a feature of biological recording

There are also areas where technologies can replace humans. Checks of record validity, such as ensuring records of terrestrial animals are not in the sea, are monotonous for humans but can be quickly and effectively carried out by computers. Even more complex tasks such as identifying species from sounds, images or environmental DNA, or providing written feedback can be completed by computers. While these methods are currently in their infancy they show the potential to work at least with more straightforward scenarios. This will allow experts to focus on the more challenging aspects of recording.

5.5 User motivation versus academic ambition

No matter the implementation of technology it is important to recognise that the motivations of the end-users and the creator are unlikely to be the same. A study of expert opinion recently found the properties of a citizen science project identified as important varied greatly depending on if the experts questioned were end-users or participants (Pocock *et al,* In Press). The motivation of participants may vary greatly, from shared beliefs with the project to a desire to increase one's reputation in the community (Nov, Arazy, & Anderson, 2014). These motivations must be considered and understood when developing a project. Once the key motivations are understood technologies can be used to target these. For example, league tables and badges can be implemented for users who are motivated by reputation and competition while apps can be loaded with identification guides and quizzes for users who are motivated by increasing their skills.

5.6 Conclusions

There are few aspects of society that have not been influenced by technologies developed in the post-industrial information revolution. These technologies shape the way we communicate, work, and socialise, fundamentally changing the way we live our lives in comparison to previous generations.

Technological advances are also changing the landscape of biological recording: websites and mobile technologies are streamlining data gathering, ensuring data quality and engaging a wider audience with nature; automation and crowd-sourcing are improving verification and validation systems; data are becoming better connected, more open and re-usable, allowing meaningful analyses at policy relevant scales; and data contributors are being rewarded with data visualisation tools, feedback and game like elements.

While there are undoubtedly challenges associated with the adoption of emerging technologies, they are set to significantly enhance biological recording, granting us a greater understanding and appreciation of the natural world.

Acknowledgements

We thank two anonymous reviewers for their helpful comments. The Biological Record Centre receives support from the Joint Nature Conservation Committee and the Natural Environment Research Council (via National Capability funding to the Centre for Ecology & Hydrology, project NEC04932).

Glossary:

Apps – An abbreviation of application, 'app' is typically used to describe a piece of software that can be downloaded and used on a mobile device such as a smartphone or tablet.

Big Data – Extremely large datasets, often a result of increased data gathering or new technology, which require modern, powerful computers and data processing to analyse.

Cloud Computing – The use of servers on the internet to store data and run software instead of a personal computer

Crowd Sourcing – Using a large number of people, often online, to perform a service often without pay. The term is a blend of 'crowd' and 'outsourcing'.

Data Warehouse – A repository of data, often coming from a number of different sources.

HTML5 – Hyper Text Markup Language is the language used to create webpages, defining how they look and function. HTML5 is the latest version of this language released in October 2014.

Machine Learning – The use of computational probabilistic methods for the detection ('learning') of patterns and structures in available ('training') data and the construction of algorithms that can be used to make predictions about previously unseen data.

Open Access – Freely accessible to all at minimal cost, usually via the internet and with no restrictions on use. The term is usually applied to data.

Open Source – Applied to software this term indicates that the source code is freely accessible.

Smartphone – A mobile phone that has functionality of a computer. Smartphones usually have access to the internet, GPS, cameras, a touch screen interface, and the ability to run 3rd party applications.

Social Networks/Media – Websites and applications that allow users to interact socially, typically online. These platforms often allow users to share messages, images and videos with one another

Anderson SE, Dave AS, Margoliash D. **1996**. Template-based automatic recognition of birdsong syllables from continuous recordings. *Journal of the Acoustical Society of America* **100**: 1209–1219.

Ansine J. **2013**. Reaching the public through iSpot: your place to share nature. Science Communication – a practical guide for scientists. Oxford: Wiley Blackwell, 257–259.

Blake S, Siddharthan A, Nguyen H, Sharma N, Robinson AM, O'Mahony E, Darvil B, Mellish C, van der Wal R. **2012**. Natural Language Generation for Nature Conservation: Automating Feedback to help Volunteers identify Bumblebee Species. Proceedings of COLING 2012.311–324.

Briggs F, Lakshminarayanan B, Neal L, Fern XZ, Raich R, Hadley SJK, Hadley AS, Betts MG. **2012**. Acoustic classification of multiple simultaneous bird species: A multi-instance multi-label approach. *Journal of the Acoustical Society of America* **131**: 4640–4650.

Burkmar R. **2014**. *The Shifting Paradigm of Biological Identification*. Shewsbury.

Burnett J, Copp C, Harding P. **1995**. *Biological recording in the United Kingdom - present practice and future development*. Department for the Environment.

Chesmore ED, Ohya E. **2004**. Automated identification of field-recorded songs of four British grasshoppers using bioacoustic signal recognition. *Bulletin of entomological research* **94**: 319–330.

CODATA-ICSTI Task Group on Data Citation Standards and Practices. **2013**. Out of Cite, Out of Mind: The Current State of Practice, Policy, and Technology for the Citation of Data. *Data Science Journal* **12**: CIDCR1–CIDCR75.

Comont R. **2013**. The Garden Bioblitz: citizen science meets social media. *National Forum for Biological Recording Newsletter* **46**: 19–20.

Costello MJ. **2009**. Motivating Online Publication of Data. *BioScience* **59**: 418–427.

Costello MJ, Michener WK, Gahegan M, Zhang Z, Bourne P, Chavan V. **2012**. *Quality assurance and Intellectual Property Rights in advancing biodiversity data publication October 2012*. Copenhagen: Global Biodiversity Information Facility.

Costello MJ, Michener WK, Gahegan M, Zhang ZQ, Bourne PE. **2013**. Biodiversity data should be published, cited, and peer reviewed. *Trends in ecology & evolution* **28**: 454–61.

Deloitte. **2013**. *The Deloitte Consumer Review Beyond the hype: The true potential of mobile*.

Demeritt D. **2001**. Scientific forest conservation and the statistical picturing of nature's limits in the Progressive-era United States. *Environment and Planning D: Society and Space* **19**: 431–459.

Department for Environment Food and Rural Affairs. **2012**. *Interim Chalara Control Plan*.

Department for Environment Food and Rural Affairs. **2013**. *UK Biodiversity Indicators in Your Pocket 2013*.

Deterding S, Sicart M, Nacke L, O'Hara K, Dixon D. **2011**. Gamification. using game-design elements in non-gaming contexts. *Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems - CHI EA '11*: 2425.

Doupe AJ, Kuhl PK. **1999**. Birdsong and human speech: Common themes and mechanisms. *Annual Review of Neuroscience* **22**: 567–631.

European Parliament. **2014**. *Legislative resolution of 16 April 2014 on the proposal for a regulation of the European Parliament and of the Council on the prevention and management of the introduction and spread of invasive alien species (COM(2013)0620 – C7-0264/2013 – 2013/0307(COD)*. European Parliament.

Freenberg A. **1999**. *Questioning Technology*. Routledge.

Galton F. **1907**. Vox populi. *Nature* **75**: 450–1.

Hagedorn G, Mietchen D, Morris R a, Agosti D, Penev L, Berendsohn WG, Hobern D. **2011**. Creative Commons licenses and the non-commercial condition: Implications for the re-use of biodiversity information. *ZooKeys* **149**: 127–49.

Haklay M. **2013**. Citizen Science and Volunteered Geographic Information – overview and typology of participation. In: Sui D, Elwood S, Goodchild M, eds. Crowdsourcing Geographic Knowledge: Volunteered Geographic Information (VGI) in Theory and Practice. Dordrecht: Springer Netherlands, 105–122.

Hickling R, Roy DB, Hill JK, Fox R, Thomas CD. **2006**. The distributions of a wide range of taxonomic groups are expanding polewards. *Global Change Biology* **12**: 450–455.

Isaac NJB, van Strien AJ, August TA, de Zeeuw MP, Roy DB. **2014**. Statistics for citizen science: extracting signals of change from noisy ecological data. *Methods in Ecology and Evolution* **5**: 1052–1060.

James T. **2011**. *Improving Wildlife Data Quality*. National Biodiversity Network.

Kumar N, Belhumeur PN, Biswas A, Jacobs DW, Kress WJ, Lopez IC, Soares JVB. **2012**. Leafsnap: A computer vision system for automatic plant species identification. Computer Vision – ECCV 2012, Lecture Notes in Computer Science.502–516.

Lintott CJ, Schawinski K, Slosar A, Land K, Bamford S, Thomas D, Raddick MJ, Nichol RC, Szalay A, Andreescu D, *et al.* 2008. Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society* **389**: 1179–1189.

Maclean IMD, Wilson RJ. **2011**. Recent ecological responses to climate change support predictions of high extinction risk. *Proceedings of the National Academy of Sciences of the United States of America* **108**: 12337–12342.

Martin J, Edwards HH, Burgess MA, Percival HF, Fagan DE, Gardner BE, Ortega-Ortiz JG, Ifju PG, Evers BS, Rambo TJ. **2012**. Estimating distribution of hidden objects with drones: From tennis balls to manatees. *PLoS ONE* **7**.

Mellinger D, Stafford K, Moore S, Dziak R, Matsumoto H. **2007**. An Overview of Fixed Passive Acoustic Observation Methods for Cetaceans. *Oceanography* **20**: 36–45.

Morrison D. **2011**. Why is taxonomy still presented to the world as books? *Australasian Systematic Botany Society Newsletter* **149**: 17–22.

Nimis PL, Lebbe R V. 2010. Tools for Identifying Biodiversity: Progress and Problems. Proceedings of the International Congress, Paris, September 20-22, 2010.

Nov O, Arazy O, Anderson D. **2014**. Scientists@Home: what drives the quantity and quality of online citizen science participation? *PloS one* **9**: e90375.

Prestopnik N, Crowston K. **2012**. Purposeful Gaming & Socio-Computational Systems: A Citizen Science Design Case. *Proceedings of the 17th ACM international conference on Supporting group work*: 75–84.

Roy HE, Adriaens T, Isaac NJB, Kenis M, Onkelinx T, Martin GS, Brown PMJ, Hautier L, Poland R, Roy DB, *et al.* 2012. Invasive alien predator causes rapid declines of native European ladybirds. *Diversity and Distributions* **18**: 717–725.

Saylor M. **2012**. *The Mobile Wave: How Mobile Intelligence Will Change Everything*. Perseus Books/Vanguard Press.

Scanlon E, Woods W, Clow D. **2014**. Informal Participation in Science in the UK: Identification, Location and Mobility with iSpot. *Educational Technology & Society* **17**: 58–71.

Shirky C. **2010**. *Cognitive Surplus: Creativity and Generosity in a Connected Age*. Penguin Group.

Silvertown J, Harvey M, Greenwood R, Dodd M, Rosewell J, Rebelo T, Ansine J, McConway K. **2015**. Crowdsourcing the identification of organisms: A case-study of iSpot. *ZooKeys* **480**: 125–146.

Sliwa J, Benoist E. **2011**. Pervasive Computing - The Next Technical Revolution. 2011 Developments in E-systems Engineering. IEEE, 621–626.

Smith VS, Rycroft SD, Harman KT, Scott B, Roberts D. **2009**. Scratchpads: a data-publishing framework to build, share and manage information on the diversity of life. *BMC bioinformatics* **10 Suppl 1**: S6.

Smith VS, Rycroft SD, Brake I, Scott B, Baker E, Livermore L, Blagoderov V, Roberts D. **2011**. Scratchpads 2.0: a Virtual Research Environment supporting scholarly collaboration, communication and data publication in biodiversity science. *ZooKeys* **70**: 53–70.

Snaddon J, Petrokofsky G, Jepson P, Willis KJ. **2013**. Biodiversity technologies: tools as change agents. *Biology letters* **9**: 20121029.

De Souza Muñoz ME, De Giovanni R, de Siqueira MF, Sutton T, Brewer P, Pereira RS, Canhos DAL, Canhos VP. **2009**. openModeller: a generic approach to species' potential distribution modelling. *GeoInformatica* **15**: 111–135.

Stowell D, Plumbley MD. **2011**. *Birdsong and C4DM: A survey of UK birdsong and machine recognition for music researchers*. Centre for Digital Music, Queen Mary, University of London.

Stowell D, Plumbley MD. **2014**. Automatic large-scale classification of bird sounds is strongly improved by unsupervised feature learning. *PeerJ* **2**: e488.

Van Strien AJ, 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 (V Devictor, Ed.). *Journal of Applied Ecology* **50**: 1450–1458.

Taxonomic Databases Working Group. **2011**. *GUID and Life Sciences Identifiers Applicability Statements*.

Trolliet F, Huynen M claude, Vermeulen C, Hambuckers A. **2014**. Use of camera traps for wildlife studies . A review. **18**: 446–454.

Tulloch AIT, Possingham HP, Joseph LN, Szabo J, Martin TG. **2013**. Realising the full potential of citizen science monitoring programs. *Biological Conservation* **165**: 128–138.

Tweddle JC, Robinson LD, Pocock MJO, Roy HE. **2012**. *Guide to citizen science: developing, implementing and evaluating citizen science to study biodiversity and the environment in the UK*. Natural History Museum and NERC Centre for Ecology & Hydrology for UK-EOF.

Walters CL, Freeman R, Collen A, Dietz C, Fenton MB, Jones G, Obrist MK, Puechmaille SJ, Sattler T, Siemers BM, *et al.* 2012. A continental-scale tool for acoustic identification of European bats. *Journal of Applied Ecology* **49**: 1064–1074.

Warf B, Sui D. **2010**. From GIS to neogeography: ontological implications and theories of truth. *Annals of GIS* **16**: 197–209.

Water Information System for Europe. **2008**. *Water Note 7 - Intercalibration: A common scale for Europe's waters*. European Commission.

Wieczorek J, Bloom D, Guralnick R, Blum S, Döring M, Giovanni R, Robertson T, Vieglais D. **2012**. Darwin Core: an evolving community-developed biodiversity data standard. *PloS one* **7**: e29715.

Wilson MW, Graham M. **2013a**. Situating neogeography. *Environment and Planning A* **45**: 3–9.

Wilson MW, Graham M. **2013b**. Neogeography and volunteered geographic information: a conversation with Michael Goodchild and Andrew Turner. *Environment and Planning A* **45**: 10–18.

Zwart MC, Baker A, McGowan PJK, Whittingham MJ. **2014**. The use of automated bioacoustic recorders to replace human wildlife surveys: an example using nightjars. *PloS one* **9**: e102770.

Table 1:

Table 1 – A summary of exemplar platforms and websites

Name	Description	URL
Atlas of Living Australia	An online platform for exploring and analysing data on the flora and fauna of Australia	www.ala.org.au
BeeWatch	A website that crowd sources records of bees, integrating natural language generation	www.bumblebeeconservation.org/get-involved/surveys/beewatch
BirdTrack	A website and smartphone app for recording bird observations in the UK	http://blx1.bto.org/birdtrack/main/data-home.jsp
BRC	The Biological Records Centre is the focus of recording for freshwater and terrestrial species in the UK	www.brc.ac.uk
eBird	An online system for recording observations of birds and for exploring the observations of others.	www.ebird.org
GBIF	The Global Biodiversity Information Facility is a repository of free to access global biodiversity data	www.gbif.org
Indicia	A free and open source toolkit for building biological recording websites	www.indicia.org.uk
iRecord	A website, built using Indicia, for collating and verifying biological records	www.brc.ac.uk/irecord
iSpot	A website designed to allow users to share images and learn about identification through crowd sourcing	www.ispotnature.org
NBN	The National Biodiversity Network is a partnership of organisations involved in biological recording in the UK. It is the UK node of GBIF	www.nbn.org.uk
The New Forest Cicada Hunt	A smartphone app for recording and identifying calls by the New Forest Cicada	http://www.newforestcicada.info/
Zooniverse	A collection of online projects that crowd source the interpretation of images and sounds	www.zooniverse.org

Figure 1. Milestones in biological recording (above) and technologies (below) in the context of the five waves of computing (bottom)

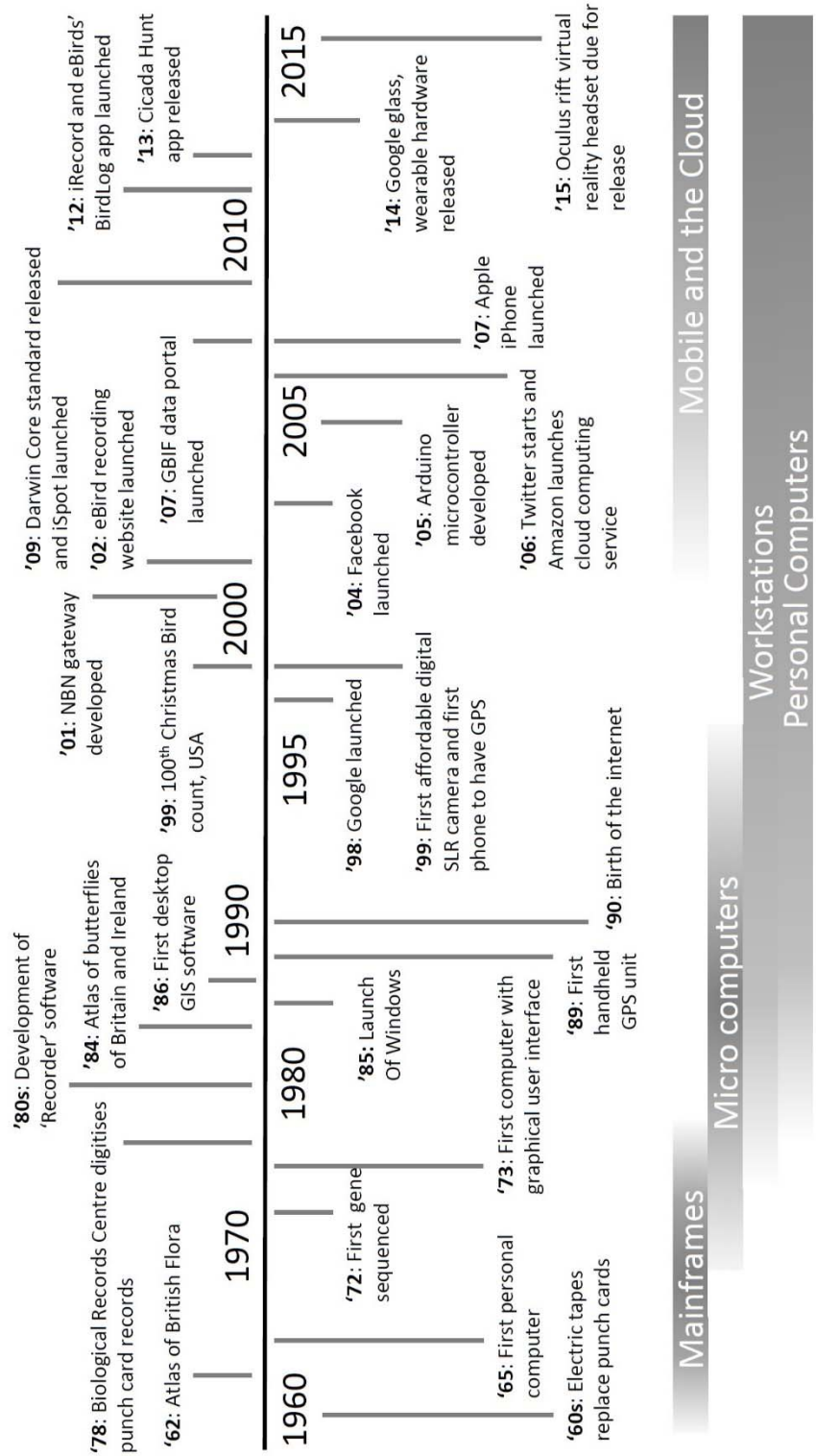

Figure 2. The number of observations per species for each species that has been observed on iSpot and that has been given an 'ID difficulty' category in the NBN Record Cleaner rulesets. iSpot data are based on all observations made in Britain up to 8 September 2014 that have received a 'Likely ID' (Silvertown *et al*., 2015). The number of species included in the analysis is given in parentheses after the taxon group name. Birds and Butterflies & moths, only use categories 1–4. Information on the NBN rulesets is available at http://www.nbn.org.uk/Tools-Resources/Recording-Resources/NBN-Record-Cleaner.aspx

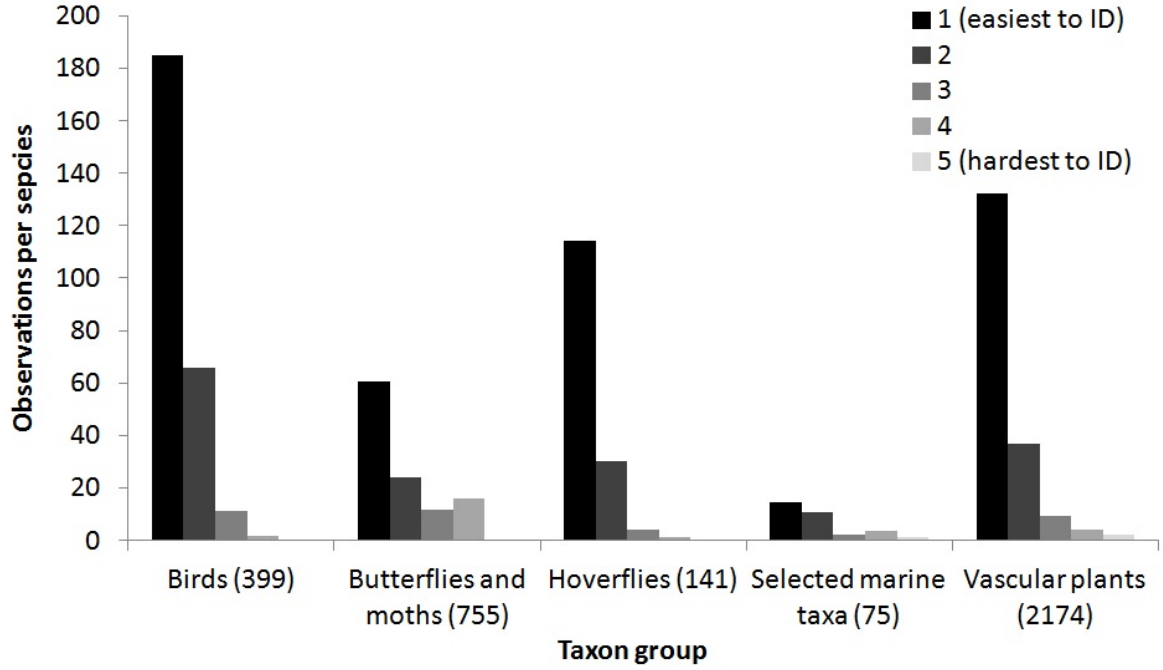

Figure 3. Schematic representation of data flow using the indicia toolkit. Data from many sources feed into the same data warehouse making verification and data sharing more straightforward. Solid arrows show data capture processes whereas dashed arrows show the flow of data post-collection.