Perspective

Drowning in data, thirsty for information and starved for understanding: A biodiversity information hub for cooperative environmental monitoring in South Africa

Sandra MacFadyen[a,ah,*], Nicky Allsopp[b], Res Altwegg[c], Sally Archibald[d], Judith Botha[e], Karen Bradshaw[f], Jane Carruthers[g], Helen De Klerk[h], Alta de Vos[i], Greg Distiller[c], Stefan Foord [j], Stefanie Freitag-Ronaldson [k], Richard Gibbs [a], Michelle Hamer [l], Pietro Landi[a], Duncan MacFadyen[m], Jeffrey Manuel[n], Guy Midgley[o], Glenn Moncrieff[b,c], Zahn Munch[h], Onisimo Mutanga[p], Sershen[q,r], Rendani Nenguda[m], Mzabalazo Ngwenya[c], Daniel Parker[s,t], Mike Peel[u,v,w], John Power[x], Joachim Pretorius[y], Syd Ramdhani[z], Mark Robertson[aa], Ian Rushworth[ab], Andrew Skowno[n,ac], Jasper Slingsby[ac,b], Andrew Turner [ad,ae], Vernon Visser [c,ah,ai], Gerhard Van Wageningen [af], Cang Hui [a,ag,ah]

[a]Mathematical Biosciences Hub, Department of Mathematical Sciences, Stellenbosch University, Matieland 7602, South Africa
[b]South African Environmental Observation Network, Fynbos Node, South Africa
[c]Centre for Statistics in Ecology, the Environment and Conservation, Department of Statistical Sciences, University of Cape Town, South Africa
[d]Centre for African Ecology, School of Animal, Plant and Environmental Sciences, University of the Witwatersrand, Johannesburg 2050, South Africa
[e]Scientific Services, South African National Parks, South Africa
[f]Computer Science, Rhodes University, South Africa
[g]Department of History, University of South Africa, South Africa
[h]Geography and Environmental Studies, Stellenbosch University, South Africa
[i]Department of Environmental Science, Rhodes University, South Africa
[j]Department of Zoology, University of Venda, South Africa
[k]Garden Route and Frontier Research Unit, SANParks, South Africa
[l]South African Research Infrastructure Roadmap, SANBI, South Africa
[m]Research and Conservation, Oppenheimer Generations, South Africa
[n]South African National Biodiversity Institute, South Africa
[o]Global Change Biology Group, Stellenbosch University, South Africa
[p]Department of Environmental Sciences, University of KwaZulu-Natal, South Africa
[q]Institute of Natural Resources, South Africa
[r]University of the Western Cape, South Africa
[s]School of Biology and Environmental Sciences, University of Mpumalanga, South Africa
[t]Wildlife and Research Management Research Group, Department of Zoology and Entomology, Rhodes University, Makhanda, South Africa
[u]ARC-Animal Production Institute, Rangeland Ecology Group, Nelspruit, South Africa
[v]School for Animal, Plant and Environmental Sciences, University of the Witwatersrand, Johannesburg, South Africa
[w]Applied Behavioural Ecology and Ecosystem Research Unit, University of South Africa, South Africa
[x]North West Provincial Government, Department of Economic Development, Environment, Conservation & Tourism, Directorate of Biodiversity Management, South Africa
[y]Data Analysis, ABSA, South Africa
[z]School of Life Sciences, University of KwaZulu-Natal, South Africa
[aa]Department of Zoology and Entomology, University of Pretoria, South Africa
[ab]Scientific Services, Ezemvelo KZN Wildlife, South Africa
[ac]Department of Biological Sciences and Centre for Statistics in Ecology, Environment and Conservation, University of Cape Town, Rondebosch, South Africa
[ad]Cape Nature, South Africa
[ae]Department of Biodiversity and Conservation Biology, University of the Western Cape, Cape Town, South Africa
[af]High Performance Computing, Stellenbosch University, South Africa
[ag]Biodiversity Informatics Group, African Institute for Mathematical Sciences, Cape Town 7945, South Africa

[ah]National Institute for Theoretical and Computational Sciences (NITheCS), Stellenbosch University, Matieland 7602, South Africa
[ai]African Climate and Development Initiative, University of Cape Town, Rondebosch 7701, Cape Town, South Africa

*Corresponding author at: Biodiversity Informatics Hub, Department of Mathematical Science, Stellenbosch University, Matieland 7602, South Africa. Email: macfadyen@sun.ac.za

Highlights

- Challenges and opportunities for Biodiversity Informatics in South Africa
- Misgivings around data sharing and multidisciplinary collaboration to leverage combined expertise
- Culture of cooperation, collaboration, interoperability to establish operational workflows for biodiversity data synthesis
- Local systems to link with global networks for an interconnected digital biodiversity knowledgebase
- SA-BioInfo-Hub: Free, user friendly, functional, stable, integrative system with varied access agreement levels

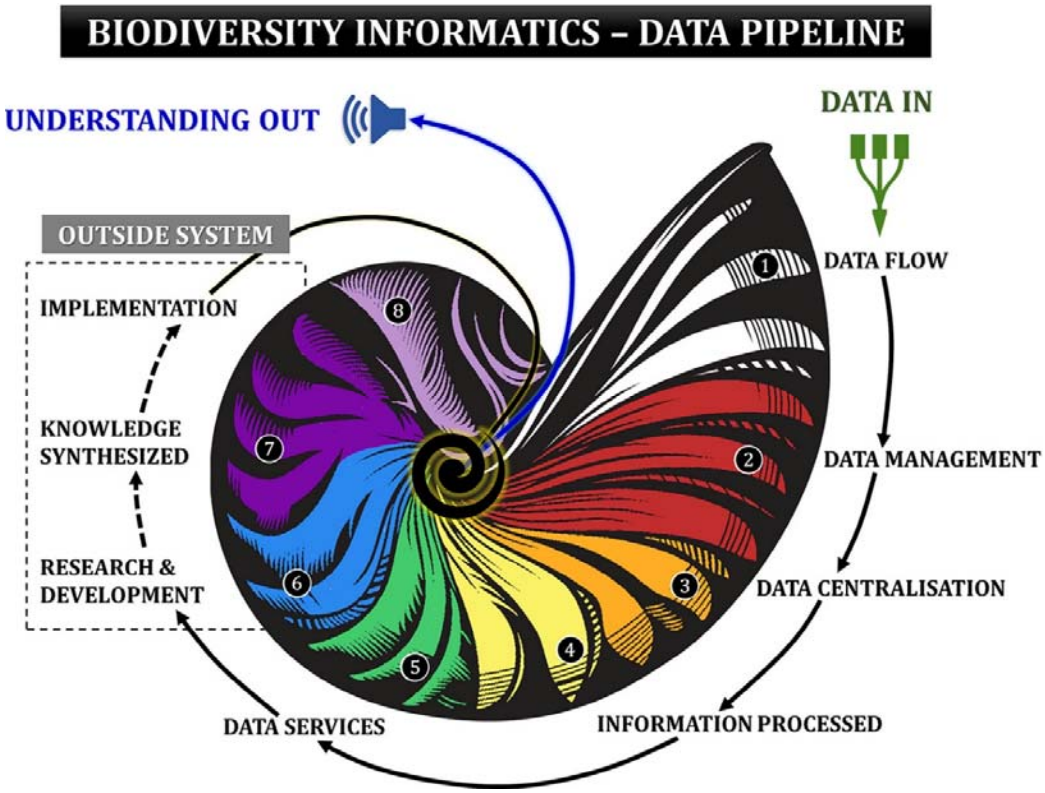
Abstract

The world is firmly cemented in a *notitian* age (Latin: *notitia*, meaning data) – drowning in data, yet thirsty for information and the synthesis of knowledge into understanding. As concerns over biodiversity declines escalate, the volume, diversity and speed at which new environmental and ecological data are generated has increased exponentially. Data availability primes the research and discovery engine driving biodiversity conservation. South Africa (SA) is poised to become a world leader in biodiversity conservation. However, continent-wide resource limitations hamper the establishment of inclusive technologies and robust platforms and tools for biodiversity informatics. In this perspectives piece, we bring together the opinions of 37 co-authors from 20 different departments, across 10 SA universities, 7 national and provincial conservation research agencies, and various institutes and private conservation, research and management bodies, to develop a way forward for biodiversity informatics in SA. We propose the development of a SA Biodiversity Informatics Hub and describe the essential components necessary for its design, implementation and sustainability. We emphasise the importance of developing a culture of cooperation, collaboration and interoperability among custodians of biodiversity data to establish operational workflows for data synthesis. However, our biggest challenges are misgivings around data sharing and multidisciplinary collaboration. We recommend a system that is free, user friendly, functional, stable, integrative and designed to cater for different data access agreement levels. Sharing data through this pipeline will directly advance the science and practice of conservation, giving multiple stakeholders and decision-makers access to valuable biodiversity data to support research and biodiversity conservation.

Keywords

Biodiversity informatics
Data collection
Data synthesis
Data sharing
Multidisciplinary science
Information systems

Graphical abstract



1. Introduction

1.1. Drowning in data

Forty years ago, Naisbitt (1982) wrote: “We are drowning in information but starved for knowledge”. Today that phrase resonates 100-fold with data engineers, scientists, ecologists and conservation and environmental managers alike (Kosta, 2017). Amid global predictions of continued biodiversity declines (Butchart et al., 2010; Díaz et al., 2019; IPCC, 2022), the volume, diversity and speed at which new environmental and ecological data, in particular, are being generated are increasing exponentially (Escamilla Molgora et al., 2020; Ivanova and Shashkov, 2021). More recently, successful citizen science and extensive public participation and community research platforms, advances in Geographical Information Systems (GIS), satellite and aerial remote sensing, camera trap and acoustic technology inter alia, have further escalated data production, availability and accessibility of both historical and almost-real-time information (Heberling et al., 2021). Concerns over spatial data bias, repeatability, accuracy and reliability do, however, also grow concurrently with these advances (Hugo and Altwegg, 2017). This exponential increase in the availability of biosphere-scale data is accompanied by a growing need for data repositories and data management systems capable of catering for a variety of data types, analytical applications and discipline-specific needs, while also ensuring recognised data standards.

Driven by the need to successfully detect and measure biodiversity change at various scales as environmental change accelerates worldwide, ecologists are continuously developing new approaches to process, analyse, and represent biodiversity data. Accordingly,

ecologists are becoming more reliant on larger, more complex, diverse, real-time, long-term and macroscale datasets (Geller et al., 2020). However, identifying and consolidating appropriate information resources from the deluge of available data is a considerable challenge for ecologists and the research world at large (Cornford et al., 2020). Nevertheless, those involved in the collection, management, analysis and use of biodiversity data have made great strides towards synthesising useable information from raw data through the theory and practice of biodiversity informatics (Bingham et al., 2017; Gadelha et al., 2020; Heberling et al., 2021). Biodiversity Informatics applies techniques from Information Technologies (IT) to improve the "management, presentation, discovery, exploration, integration and analysis of biodiversity data" (Martellos and Attorre, 2012). While numerous online biodiversity databases already exist (Table S2), heterogeneous data integration – i.e. diverse data types from a variety of sources, comprised of potentially varying scales (spatial, temporal, taxon-level), formats and accuracies – remains a challenge that can limit our ability to detect and understand the full spectrum of biodiversity change and its socio-ecological implications (Noss, 1990; La Salle et al., 2016; Enquist et al., 2016).

Despite the considerable efforts made to describe and quantify biodiversity, researchers also need to be aware of potential shortfalls in biodiversity data resources that can restrict their efficacy in helping answer pivotal research and/or management questions (Hortal et al., 2015). These include recognised knowledge gaps in species taxonomy (Coleman and Radulovici, 2020), distribution, abundance and evolutionary patterns, abiotic tolerances of species, species traits and biotic interactions (Jucker et al., 2018) among others. For example, haphazard, rather than probabilistic, sampling designs can lead to strong spatial biases in the data (Tulloch et al., 2013; Hugo and Altwegg, 2017). Imperfect detection can also lead to biased estimates of demographic parameters, abundances, species richness and distribution patterns (Yoccoz et al., 2001). In addition, a range of resource and technical limitations, data accuracy uncertainties, ethical and legislative (e.g. permitting) challenges still exist, especially in the developing regions of the world (Stephenson et al., 2017a). Finally, coordination and collaboration between the many actors in the sector is crucial. The existing structures and forums do not adequately address biodiversity monitoring and informatics, and there is scope to improve this regionally, nationally and globally.

What follows is a synthesis of the challenges facing the development of biodiversity informatics in South Africa (SA). We outline the opportunities that exist for its broader application and suggest the development of formal pathways to collaboration and networking to facilitate improved data sharing and more inclusive data use. Using SA as a case study, we illustrate how biodiversity informatics cannot always be successfully applied in areas with high (or unestimated) biodiversity value but few technical and financial resources. Furthermore, we propose a solutions framework to aid other megadiverse, developing countries to expand national or even continent-wide biodiversity informatics applications.

1.2. Global data needs for monitoring biodiversity

The 'big data revolution' and the rise of Information and Communication Technologies (ICT) have transformed many research fields. However, the application of biodiversity informatics in ecology has grown especially quickly (Osawa, 2019). This growth is unsurprising, as the need to monitor and evaluate global scale environmental change is rapidly expanding in the face of climate change and the unprecedented growth in human population size (Ceballos et al., 2015; Navarro et al., 2017). Recognising this need, the United Nations Convention on Biological Diversity developed the 20 Aichi Biodiversity Targets as part of the Strategic Plan for Biodiversity 2011–2020. To report on progress towards these targets, the Group on Earth Observations Biodiversity Observation Network (GEO BON) later proposed 22 Essential Biodiversity Variables (EBVs) to measure target achievements using different biodiversity indicators (Pereira et al., 2013). Proença et al. (2017) linked these EBVs with available data

sources and demonstrated that very few datasets could in fact be readily consolidated into representative and measurable indicators. They further highlighted the need for more intensive global monitoring programmes to build datasets with sufficient coverage to enhance the utility of EBVs (Proença et al., 2017). Similarly, Kissling et al. (2015) warned that data available from global research infrastructures are not always sufficiently standardised to build effective EBV data products. In response, Hardisty et al. (2019) developed the Bari Manifesto, consisting of 10 principles of best practice for EBV-focused biodiversity informatics, illustrating the multiplicity needed to produce relevant, repeatable, and fit for purpose EBV datasets.

The failings of the Aichi targets a decade later (six partially and zero fully achieved by 2020; CBD, 2020) have in part been attributed to weak implementation strategies, underdeveloped knowledge management plans, inadequate programmes for building human capacity (Xu et al., 2021), and their lack of utility (Anonymous, 2020). In other words, they failed to translate into real-world or applied measures by which progress could be evaluated and goals realistically achieved (Anonymous, 2020). These challenges are likely to persist, even as the world shifts focus to the new post-2020 goals, if the obstacles to generating effective indicators are not addressed (CBD, 2021). These obstacles are exacerbated when indicators are derived from composites of imperfect data, amassed from disparate sources (e.g. historic – published and grey – literature, field surveys, biological collections, molecular data, automated sensors) (Proença et al., 2017; Hansen et al., 2021; Heberling et al., 2021).

To filter and synthesise such diverse data into reliable information, well documented (metadata), large-scale datasets need to be openly available, useable, scalable and easily interpreted (Stephenson et al., 2017a; Gadelha et al., 2020). A culture of cooperation, collaboration and interoperability among custodians of biodiversity data is, therefore, an essential component to establish operational workflows of trans-national and cross-infrastructure and/or cross-platform biodiversity data synthesis (Hardisty et al., 2019). SA has already made significant progress towards achieving an open data society by virtue of various institutional biodiversity data portals (Table S2). Most notable of these include: Biodiversity Advisor (http://biodiversityadvisor.sanbi.org/), FBIS (freshwaterbiodiversity.org), BODATSA (newposa.sanbi.org), SAEON (2021) (catalogue.saeon.ac.za) and *E*-GIS (egis.environment.gov.za). Nonetheless, locating, navigating and consolidating data from multiple individual, disconnected local and global systems, confounds and constrains the timeliness of data availability for biodiversity management and long-term monitoring of status or trends (Ball-Damerow et al., 2019; Blair et al., 2020). The demand for adaptive management and evidence-based conservation strategies also requires more rapid data integration, analysis and knowledge extraction to develop iterative learning feedbacks that inform decisions and enhance conservation best practice (Gillson et al., 2019; Raymond et al., 2022). Likewise, the flood of different data and metadata formats, and contrasting scales and accuracies of data from distinct portals and diverse scientific disciplines also complicates the data integration and wider interoperability that are required for effective national and international research and monitoring efforts.

While using biodiversity informatics principles to address these and other challenges, it is similarly important to recognise inherent inequalities between the Global North and South (Kuras et al., 2020). For example in the Global South, biodiversity-related knowledge and technology, derived from work conducted by resource-rich countries, are often not effectively transferred to their developing country hosts (Vanhove et al., 2017). This is further inhibited by “parachute”, “helicopter” or “colonial” science where scientists from the global North do research in the global South without collaborating or sharing data and knowledge or skills with local scientists and authorities (Pettorelli et al., 2021; Stefanoudis et al., 2021). Successful partnerships between the Global North and South thrive, however, when the latter has strong in-house capacity that can direct efforts into building internal capacity and

help actualise research that is relevant to solving real-world problems. We believe SA is strategically positioned and has the potential for more multi-disciplinary and inclusive data sharing by developing a national system, which links portals and people (including policy and decision-makers), while also encouraging improved and ethical data integration for biodiversity monitoring both nationally and globally.

1.3. The need for and pitfalls of data sharing

Data sharing is important in any field as it can lead to new and more robust insights when studies are expanded across both spatial and temporal scales, knowledge bases are widened, and different disciplines are brought together to promote innovative and transdisciplinary thinking (Thessen et al., 2018). However, data sharing is unequivocally vital for expanding ecological research as ecologists need data from a variety of fields to understand the complexity of ecosystems (Hortal et al., 2015). For example, these data may represent different *environmental components* such as topography, altitude, geology, climate, fauna and flora, *patterns of ecological processes* like plant phenology, pollination, herbivory, fire and decomposition, and *anthropogenic factors* that embody many socio-economic, cultural, and sustainability issues (Noss, 1990; Shin et al., 2020). To capture the complexity and dynamics of *socio-ecological systems*, more qualitative data are often needed to better understand the complex causal relationships between human societies and ecosystems (Biggs et al., 2018; De Vos et al., 2019; Cox et al., 2021).

'Data sharing' among researchers and other data custodians (e.g. communities) can, however, be contentious, and is often underpinned by complex power dynamics. The 2020 State of Open Data report identified "trust" (or the lack thereof) as a formidable, albeit intangible barrier to data sharing (Digital Science Report, 2020). Addressing issues of trust, along with institutional policies on ethics, intellectual property rights, data ownership agreements and academic reward systems, is a challenge that requires innovative approaches that recognise local attitudes as well as global inequalities. For instance, SA scientists are often custodians of valuable datasets that attract international collaborators (e.g. Schurr et al., 2012; Smit et al., 2013; Smith et al., 2013). However, due to analytical or technical disadvantages, capacity limitations and/or inadequate succession planning or "parachute" science, SA scientists are often peripheral to the emerging research. This only serves to further undermine trust and limits our ability to improve expertise and capacity (Hudson et al., 2020; Ambler et al., 2021). Secondly, some data are only commercially available, while others need to be requested and new data user agreements signed for each separate use case; for example, same researcher, different project (e.g. South African Weather Services, SANParks). The sector is also rife with poor data capture and management practices, resulting in poorly documented data (no or inadequate metadata), lack of quality control and loosely applied data standards (Bayraktarov et al., 2019). However, data management has historically never formally formed part of any curriculum linked to biodiversity researcher development in SA (Ball-Damerow et al., 2019). Rather than simply linking data providers and data users, SA scientists need to equip themselves with the necessary skills and means to source secondary data to integrate with their own primary data to answer relevant ecological questions and build meaningful 'data' partnerships that will advance scientific, developmental and/or conservation goals. This includes developing skills in and encouraging more reproducible research to ensure repeatable and more rigorous science.

In general, the younger generation of researchers is more open to data sharing (Tenopir et al., 2011; Stieglitz et al., 2020). Due to academic or funder-imposed time constraints of M.Sc. and Ph.D. degrees, students are often unable to collect extensive or long-term datasets. Instead, these students may use secondary or derivative data from previous studies and/or data downloaded from the many online databases (Table S2). In contrast,

ecologists of the past often spent years collecting data towards specific management, monitoring or research goals. While some of those goals may remain unfulfilled (e.g. unpublished) due to the aforementioned computational and/or analytical limitations, these researchers may be reluctant to openly share their data and/or metadata. Even where data are shared, embargoes of two, five or more years are not uncommon to help safeguard intellectual properties and support the publication of results (Michener, 2015). However, this concession can also create a lag in research discoveries (and associated impact), heighten the risks of data being poorly indexed, referenced or stored, underutilised and/or even lost (Heidorn, 2008).

Again, we believe the solution lies in supporting and encouraging collaborative networks and building communities of practice around biodiversity data, using established and agreed upon workflows based on secure and dependable infrastructure, rather than simply sharing it through an anonymous pipeline. Ideally, the system must allow users to choose the level of data package security and accessibility. For instance, 'private' (only data custodian has access, and is available offline), 'meta-data only' (data in progress), 'user-restricted' (access restricted to user-defined group), 'collaborative' (click to connect with data custodian to develop collaborative network) and 'open' data (data available to anyone provided data package is cited). While some of the existing systems do cater for varying levels of access, we argue that a completely 'private' or offline option, in which custodians may upload incomplete data packages even if they choose to never release them, might make researchers more willing to ultimately share. To achieve this at a national level, SA needs to build an inclusive network of key biodiversity stakeholders. In this way, a more connected and collaborative landscape of biodiversity data can be established to better engage biodiversity science for evidenced-based decision making (Musvuugwa et al., 2021). Table S1 lists what we believe to be the key stakeholders of biodiversity data in SA that may benefit from collaborative partnerships in the future.

2. The benefits of a South African Biodiversity Informatics Hub (SA-BioInfo-Hub)

The importance of a cooperative network for biodiversity observation or monitoring has already been well articulated by GEOBON in their call for the establishment of national and regional Biodiversity Observation Networks (BON) connected to a global system (Scholes et al., 2012). The benefit of a national (local rather than global) system lies in stakeholder engagement. Many SA biodiversity stakeholders (Table S1) either already work together or are aware of each other's research, which lays a solid groundwork to encourage data sharing, inspire partnerships and ultimately help strengthen SA's research profile and its biodiversity conservation. A national platform can also be more flexible and dynamic, and can potentially support more applied research needs of public, private and provincial conservation agencies. This includes support for the continuation of longer-term monitoring programmes in, for instance, South African National Parks, South African National Biodiversity Institute, Ezemvelo KZN Wildlife, Mpumalanga Tourism and Parks Agency, North West Parks Board, CapeNature, Agricultural Research Council and other private research and conservation groups. The proposed SA-BioInfo-Hub should, together with existing structures, also be linked to Departments of Forestry, Fisheries and the Environment (DFFE) and Science and Innovation (DSI), and form the backbone of a South African BON. Initially, the development of the SA-BioInfo-Hub will require the construction of interdependent data pipelines from the ground-up, aligned with global systems or programmes and international conventions (e.g. CBD), that will acknowledge and incorporate existing local efforts, and leverage the transfer of knowledge and skills before they are lost. For instance, researchers from the Smithsonian Institute estimate that "up to 80 percent of raw scientific data collected in the early 1990s are gone forever, mostly because no one knows where to find it" (Vines et al., 2013). Along with addressing the obvious challenges of a *notitian* age (from the Latin word *notitia*, meaning data), the SA-

BioInfo-Hub will also afford SA scientists many opportunities to share, integrate and synthesise relevant biodiversity knowledge through the theory and practice of biodiversity informatics (Escamilla Molgora et al., 2020; IPBES, 2020; Heberling et al., 2021). Biodiversity informatics combines information technology with ecological and biodiversity sciences, helping decision-makers harness and synthesise useable information from raw data (Bingham et al., 2017; Osawa, 2019; Gadelha et al., 2020). It should also include the development of standards (e.g. Darwin Core; Wieczorek et al., 2012), methods, and tools for capturing, digitising, storing, managing, accessing, displaying and analysing biodiversity data through a standard, reproducible 'pipeline' (Bingham et al., 2017; Ivanova and Shashkov, 2021). Wilkinson et al. (2016) established the FAIR principles for scientific data management and stewardship to improve the **F**indability, **A**ccessibility, **I**nteroperability, and **R**euse (i.e. repeatability or ability to regenerate data) of digital data. These principles form the backbone of many data management platforms including the Global Biodiversity Information Facility (GBIF, 2021).

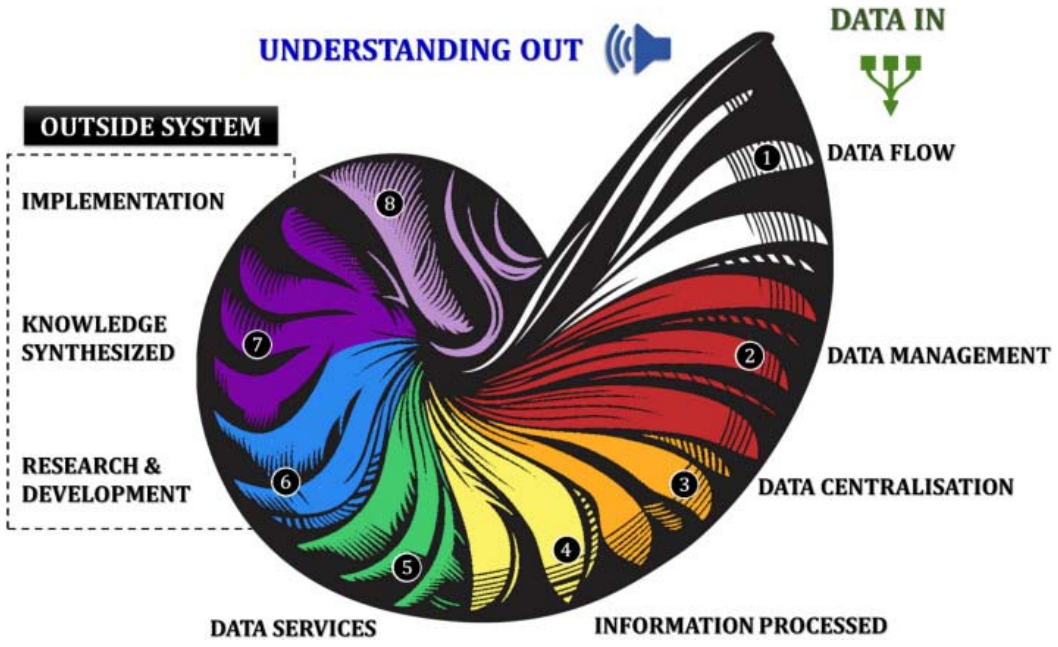


Fig. 1. Proposed data pipeline for Biodiversity Informatics in South Africa, indicating ❶ the flow of data into the system, to develop ❷ a data management plan (incl., design and implement data and metadata standards, data structures and formats, conditions of use and resource requirements) that will facilitate ❸ the centralisation of data (incl., standardised data collection, automated data capture, database management, storage and service systems), that can be ❹ processed into information (i.e. data is processed into information, directly linked to decision needs) and shared via ❺ data services (i.e. services that allow users to navigate, search or query, display, analyse and export outputs) that will filter data products out of the system to end-users, who will use these data to ❻ research and develop new ideas (e.g. advanced analysis, wider data integration and export outputs) that can be ❼ synthesised into new knowledge. The ❽ implementation or application of this new knowledge (e.g. solutions based application for monitoring and management of biodiversity) leads to 〉〉 ⇒ true understanding and is the pathway to change (i.e. the pathway to change percolates through awareness → agreement → acceptance → standard practice → implementation → and ultimately behavioural change). The figure is designed with the Nautilus Shell in mind because it exemplifies the Golden Ratio spiral where the first two numbers in a series of numbers add up to the succeeding number. This pattern is repeatedly found in nature and illustrates how ecosystems and biodiversity are not as random or irregular as they may appear, but can often be explained in the logic of mathematics. At the same time the spiral epitomises how all biodiversity data are intrinsically linked and should form part of a greater whole for global biodiversity monitoring, research and conservation.

GBIF (including the South African Biodiversity Information Facility, SABIF), the Botanical Information and Ecology Network (BIEN), Map of Life (MOL), and other effective global systems, are growing rapidly, providing invaluable data for global ecosystem monitoring (Stephenson and Stengel, 2020). For areas where these global systems are limited (Yesson et al., 2007; Boakes et al., 2010; Zizka et al., 2020), especially in Africa (Siddig, 2019), local systems help bridge the geographic and taxonomic divide (Maldonado et al., 2015). The vast majority of both global and local systems are focused on species distributions, including GBIF (Petersen et al., 2021), with very few that combine occurrence records with underlying abundance and/or environmental data (Stephenson and Stengel, 2020). We argue that a national system will help streamline biodiversity research and stimulate more ecosystem-level inquiry (Fig. 1). For example, by acting as a repository for environmental layers (vector) and surfaces (raster), not already included in existing systems, which allows users to query existing databases through Application Programming Interfaces (APIs), programmable code (e.g. R or Python) and/or existing cloud-based technologies like Google Earth Engine (Gorelick et al., 2017), data cube libraries like Geospatial Data Abstraction Library (GDAL), gdalcubes (Appel and Pebesma, 2019) and sits (Simoes et al., 2021). Further incorporation of Big Data analytical tools, including machine learning and artificial intelligence (AI), will facilitate extraction and selection of relevant features to solve complex biodiversity problems.

2.1. Understanding and refurbishing data pipeline functionality

In biodiversity informatics (or any information system), data generally runs through different stages of a pipeline, from raw data collected in the field or laboratory, according to agreed collection and metadata standards → to data captured into digital formats, including standardisation and additional metadata capture → data storage or warehousing → to end-users for basic analytics, including data querying and visualisation → to end-users and more advanced data analyses → to research, management or policy outputs (Fig. 1). Each stage in the data lifecycle requires extensive manpower and multiple skillsets spanning a variety of disciplines. Thus, without more human capacity and a multi-disciplinary approach, many valuable datasets will remain uncaptured, lack important metadata standards, and/or preclude wider data sharing and integration. In 2013, the Global Biodiversity Informatics Outlook stressed the importance of cooperative networks between researchers, policymakers, and other stakeholders to encourage data sharing, integration and synthesis to support better decisions in conservation management (Hobern et al., 2013). We share these sentiments and further explore the key challenges and related opportunities facing SA biodiversity informatics using a data pipeline framework depicted in Fig. 1. Each challenge, need and associated opportunity is described in more detail in Table S3.

- (1) Raw data flows in: In this information age we are inundated by data from a myriad of sources, comprising numerous formats, degrees of accuracy and regulated conditions of use. For biodiversity scientists, the challenge lies in finding, accessing, and consolidating data for reuse. Therefore, a sound plan for managing data in a way that will conquer these challenges and make the most of opportunities is of paramount importance.
- (2) Data management plan: Our data management should provide a consensus of data and metadata standards, approved data structures and formats, statements of legal conditions of use, and resource requirements. Resource requirements can be further unpacked into:
    - i) *People (manpower), training (skills) and building more inclusive collaborative research networks:* Human capital will rely more heavily on technology in the future. However, rather than encouraging more students to enrol in computer science, data science, information systems and IT degrees, which may lead to these fields becoming saturated with graduates, we suggest a more interdisciplinary approach to train new and upskill existing SA

scientists. For example, ecologists are often expected to be experts in their fields while having knowledge or skills in information science, including information security, copyright laws, librarianship and archiving, data management, statistics and engineering (Digital Science Report, 2020). This knowledge can be gained by incorporating elements of data science into existing syllabuses (botany, zoology, entomology, ecology, conservation etc.) and/or by facilitating links among ecologists, bioinformaticists and Big Data and IT experts. Cross-faculty courses that develop basic data literacy or applied data science skills may be integrated into existing graduate programmes, and should offer capacity development support to research postgraduates, faculty and practicing ecologists. Moreover, we propose that collaborative networks should begin to be encouraged from the university level (senior under- and post-graduate). For instance, different faculties might collectively design a 'data pipeline project' where students from various disciplines including, botany, zoology, ecology, environmental science and anthropology work together with students from mathematics, information technology and data science, to create a working database management system (DBMS) to collect, capture, store, process, query, share and use biodiversity data.

    - ii) *Services (*e.g. *network infrastructure):* A crucial step for SA biodiversity informatics is the expansion of existing network infrastructure and modernisation of ICT systems, particularly in rural and low income areas (ICASA, 2020). While this is a national development prerogative, it affects biodiversity informatics in that reliable access to information is crucial for ecologists, scientists, managers and land-owners in isolated formal and informal protected areas to make effective management decisions. While SA has made some progress in this regard, conservation bodies require more consideration. A case in point is the National Research and Education Network of SA (SANReN) which aims to develop a high-speed network dedicated to science, research, education and innovation (SANReN, 2021). By March 2019, SANReN had already connected 236 universities, science or research councils, national facilities and institutions, academic hospitals and museums. However, none of the provincial or national conservation agencies (e.g. SANParks, Cape Nature, Ezemvelo KZN-Wildlife) are currently part of this group (TENET, 2021). This lack of inclusion may in part be due to confusion around the legislation of these agencies as formal research institutions (National Research Foundation Act [No. 23 of 1998] 1998).
    - iii) *Systems (computer infrastructure); iv) Software; v) Repositories (data warehousing facilities); and vi) Maintenance schedules and/or contracts* are all additional resource requirements that should be predetermined.

- (3) Data centralisation: In line with Wilkinson et al.'s (2016) FAIR principles, a single national data pipeline for biodiversity data in SA needs to centralise data and help integrate different systems. In this way we can, 1) avoid duplication of efforts; 2) simplify data discovery (**F**AIR, **F**indability); 3) make data more accessible (F**A**IR, **A**ccessibility); 4) expedite data synthesis (FA**I**R, **I**nteroperability) and; 5) reform attitudes towards data sharing (FAI**R**, **R**euse). Naturally, this is not a trivial task with different systems optimised for distinct disciplines and diverse data formats, e.g. species occurrence records, satellite or aerial remote sensing products, and camera trap images, among others (Stephenson et al., 2017b). The need for data is the common thread running through disciplines, even though analytical methods and techniques can differ widely. With this in mind, we suggest the SA-BioInfo-Hub should initially focus on developing a sound foundational framework for standardised data and metadata collection, capture, storage or warehousing and basic data querying and visualisation. That is, academic end-users may export these data products for more advanced data analyses outside the system (Fig. 1). However, we

encourage the development of a ‘sandbox’ where users can test and share code, access learning tutorials and/or different tools. Any useful data pipeline needs to start somewhere and should cover at least three functional requirements: i) Standardised data collection; ii) Automated data capture; iii) Database management, storage and service systems:

  - i) The first step is to standardise data collection protocols, preferably using free, easy to use, established systems like: CyberTracker (Kruger and MacFadyen, 2011); Survey123 (2021); Open Data Kit (2021); or KoBoToolbox (2021). In this manner, we can eliminate much of the data cleaning associated with field data recording, nomenclature, data collection errors (e.g. typographical errors) and other capture errors. The backbone of some of these systems even includes the design or setup of appropriate data architectures and/or database systems/structures needed to store/warehouse data (Fig. 1). These standards will need to incorporate many data types, e.g. tabular, spatial, remotely-sensed and other imagery, audio and more.
  - ii) A paradox of modern scientific research is that everyone needs clean, well-annotated, longterm datasets to generate accurate and reliable information that feeds into research initiatives focused on filling recognised knowledge gaps, but few want to capture and/or ‘clean’ the data. Indeed, some highly valuable, long-term datasets remain uncaptured, stored away on hardcopy datasheets, while others remain stored on old *floppy* or *stiffy* discs. Technologies already exist in the Librarian and ICT fields to extract or capture such data using, for example, Optical Character Recognition (OCR), Natural Language Processing (NLP) and Advanced Analytics (Owen et al., 2020). Making these technologies known and accessible to non-data scientists or non-ICT specialists is another key addition to the pipeline. The paradox continues where long-term datasets are essential for monitoring environmental change (e.g. climate change) but longstanding monitoring programmes are shut down, have data gaps or are scaled-back due to funding restrictions, institutional changes and/or lack of succession planning (Slingsby et al., 2021).
  - iii) A DBMS handles the storage, retrieval, and updating of data in the pipeline (Sreenivasaiah and Kim, 2010). DBMS software functions as an interface between the end-user and the database, simultaneously managing the data, the database engine, and the database schema in order to facilitate the organisation and processing of data (Vargas-Solar et al., 2017). Similarly, data as a service (DaaS) is typically a set of cloud-based software tools used for managing, analysing and sharing data in a data warehouse. In this way, end users can access ‘cleaned’/standardised data while data security and copyright strategies safeguard data ownership rights. Moreover, data packages and APIs that link these systems with popular statistical platforms like R, Jupyter notebook, and Python, could help advance SA biodiversity informatics. Here new systems can learn from or expand upon existing ones like SAEON's e-catalogue, for example.

- (4) Information processing: Before data can be effectively used, it needs to be processed into information using standardised structures and formats that can be shared, analysed and presented. Amidst the myriad of heterogeneous data, the Darwin Core data standard offers a common language to facilitate biodiversity data sharing (Wieczorek et al., 2012). For spatial data, the Spatial Data Infrastructure Act 54 of 2003 outlines standards for the South African Spatial Data Infrastructure (SASDI) implementation by the National GeoSpatial Information directorate (NGI) through the National Spatial Information Forum (NSIF).
- (5) Data services: To facilitate effective data use, the system should be designed in a way that professionals can easily navigate the platform, access data, and avoid the trap of trying to design an all-encompassing, super-system comprised of all the tools,

techniques or analytical methods any scientists could possibly want. The development of APIs or tools that promise to satisfy the needs of researchers across taxonomic, disciplinary, geographical and socio-economic boundaries, are a common feature of modern day biodiversity science literature (Heberling et al., 2021). Whereas much less attention is given to developing sound, foundational data stewardship plans, which includes crucial foundational system structure design and data architecture planning. Considering that there are many existing software platforms (e.g. R, Jupyter notebook, Python, Matlab), along with numerous methods/techniques that are so varied and wide-ranging (e.g. statistical or mathematical modelling, incl. AI and machine learning), we believe it impractical to try to combine all of these into a single system. Therefore, a set of easy to use data access tools that are of immediate benefit to researchers may encourage data use more readily across disciplines.

Actions outside the system: The remaining three nodes of the biodiversity data pipeline occur outside of the system in its most basic form. We briefly discuss these in the context of practical applications for conservation research and management and future development pathways to guide solutions based tool construction.

- (6) Research and development: Any biodiversity informatics platform should be dynamic, encouraging collaborations to grow and produce novel ways to monitor and protect biodiversity. For instance, a teaching component (e.g. online course material) or a suite of application-specific tools for conservation managers (e.g. fire decision support system or stocking-rate manager) would be a valuable addition. Importantly, while 'data use' has always been seen as the exit point from a data pipeline, we affirm that 'exiting' data products or results should always feed back into the data pipeline. This highlights another key challenge for biodiversity informatics in SA, i.e., that end products or results from most environmental research projects are never fed back into any larger body of biodiversity knowledge.
- (7) Synthesised knowledge: Once data have been analysed by researchers in relevant fields, results should be synthesised into actionable knowledge. Synthesis takes place when research outputs are presented in a digestible format, e.g. scientific article, research report, policy brief or public presentation. However, the derivative layers should also be fed back into the data pipeline so that results can be synthesised into actionable knowledge for conservation managers to implement local monitoring strategies.
- (8) Implementation strategy: That is, this knowledge can now be applied to central research and/or management questions to provide user defined solutions and policy updates. For example, visualisation of surface water dynamics for predictive species distribution models (https://www.glad.umd.edu/dataset/global-surface-water-dynamics by Pickens et al., 2020).

〉〉 Understanding: Understanding is the pathway to change but needs to pass through awareness → agreement → acceptance → before it can become standard practise → be implemented and → effect actual change. Different disciplines may also develop different understandings from the same input which can run in parallel. Understanding is a fluid, dynamic and adaptive process that comes from developing knowledge in context.

3. Concluding remarks and recommendations

In a global context, new biodiversity targets are being drafted (CBD, 2021) and currently comprise four goals and 21 targets for 2050, and 10 milestones to achieve by 2030. Meanwhile, focus has shifted towards the United Nations' (UN) 2030 Agenda for Sustainable Development and its 17 Sustainable Development Goals (Hoskins et al., 2020). Recognising

the mismatch between ambition and achievement in the Aichi targets, the UN adopted a System of Environmental Economic Accounting (SEEA) as the new global standard for collecting and reporting environmental data (Anonymous, 2020). Thus, an urgent need to integrate macro-scale biodiversity knowledge clearly exists (Heberling et al., 2021). However, for global biodiversity monitoring to truly be successful, national initiatives addressing local user needs (especially those in under resourced countries), need to be encouraged and supported. For example, despite the clear need and importance of a national data pipeline being recognised in the past, a national system is yet to be developed in SA. In 2011, the National Integrated Cyber Infrastructure System (NICIS) initiated the Data Intensive Research Initiative of SA to develop a national data portal for SA research, providing various data management services that included User Subscriptions, Data Management Planning, Data Repositories and DOI Minting (DIRISA, 2021). Unfortunately, all development halted on this initiative before it could come to fruition. In 2013 the Foundational Biodiversity Information Programme (FBIP) was established by the Department of Science and Innovation (DSI), the National Research Foundation (NRF) and SANBI to help generate funds for Biodiversity Informatics systems in SA (Coetzer and Hamer, 2019). National-level requirements, like the need for metadata and data collection, generation, management, dissemination and reuse of standards for biodiversity information, still require more attentive resolve. While many local, national and international systems continue to exist and grow in SA, they could benefit from improved platform connectivity and data synthesis (stakeholders listed in Table S1 and systems listed Table S2). The Freshwater Biodiversity Information System (freshwaterbiodiversity.org) is a compelling example of how different forms of ecological data ranging from ecosystem state data to distribution data can be presented and shared. Another is the Atlas of Living Australia (Belbin et al., 2021) and the Biodiversity National Network of Mozambique (https://bionomo.openscidata.org/bionomo). However security of funding is always a concern in such initiatives.

We believe that SA might promote more multi-disciplinary and inclusive data sharing by developing a national system that links portals and people, and encourages improved data integration for biodiversity monitoring. Within the diverse SA data landscape, exists the potential for constructive linkages, mutually beneficial relationships, and functional complementarity (Bingham et al., 2017). We identified several key challenges for biodiversity informatics in SA and offered ideas for possible solutions or opportunities (Table S3). The importance of multi-stakeholder engagement to identify stoppages in the data pipeline and find common solutions, should not be overlooked during the design and implementation phases of building a national system. It is also clear that in this Big Data and ICT era, a critical first step for biodiversity informatics in SA is the development of meaningful partnerships among data stakeholders. We list potential stakeholders (Table S1) as well as commonly used online databases (Table S2) to encourage future network building. We also highlighted the importance of funding to complete vital network infrastructure and ICT system upgrades, especially within protected areas to support conservation agencies and organisations often nested in rural landscapes. Human capital development is also emphasised as an essential requirement to boost multi-disciplinary skills.

We call for a national pipeline for biodiversity data and described the essential components required to design and implement the SA-BioInfo-Hub. The expansion hereof includes the development of standards, methods, tools and infrastructure for capturing, digitising, storing, managing, accessing and analysing biodiversity data through a structured, secure biodiversity data pipeline. We strongly advocate for the integration of such a national system into existing global initiatives. For example, SABIF is the South African node of the Global Biodiversity Information Facility (GBIF), which strives to empower a global network of ecological data stakeholders to develop an interconnected digital knowledgebase for biodiversity data. We believe the SA-BioInfo-Hub will help promote African science by

African scientists, and reshape the way we engage with global biodiversity scientists and research programmes. In doing so, we hope to adopt a more holistic approach to biodiversity monitoring by combining the extensive species distribution records of GBIF, iNaturalist, GEO-BON and others with local and/or regional patterns of relevant environmental variables and processes. Furthermore, we expect stronger stakeholder participation and data sharing opportunities within a national framework, because many stakeholders may have worked together in diverse past contexts and already established some degree of trust.

South Africa is a known hotspot of biodiversity, comprising almost 2 % of the world's recognised biodiversity hotspots (Newbold et al., 2016), making it the 18th most biodiverse country in the world. Regionally, Africa comprises more than 18 % of these hotspots and is the third most diverse continent. As such, it is paramount that SA scientists recognise and address the challenges facing biodiversity monitoring and data management in line with global standards. These include lags in technology and skills transfer, limited manpower and succession planning, and a lack of goal directed initiatives focused on developing more rigorous biodiversity informatics across, between and connecting multiple disciplines. Rather than toiling in an unfamiliar data science domain, we believe that researchers need to establish transdisciplinary, collaborative networks across Africa – preferably early in their careers – that bring together the expertise of computer scientists and information technologists, librarians and historians, statisticians and mathematicians, ecologists, social scientists, local communities and conservation managers to make data readily and sensibly accessible. These different disciplines and knowledge domains already have the necessary tools and expertise to complete separate tasks in the pipeline, but we can only begin to benefit from an open data society when we are able to bring all biodiversity stakeholders and relevant expertise together. At the same time, we recognise the need to strengthen regional collaboration for environmental data synthesis across Africa. As a whole, the continent requires support to meet ambitious conservation targets especially given its unique biodiversity across diverse biomes, biogeographical gradients and disparate development trajectories and demands. Efforts like the Southern African Science Service Centre for Climate Change and Adaptive Land Management (SASSCAL) give prominence to wider data synthesis needs. Ultimately, we hope the perspectives synthesised here can be expanded to include an intrinsically African Biodiversity Informatics Hub.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. Sandra MacFadyen, Richard Gibbs, Pietro Landi and Cang Hui report financial support was provided by National Research Foundation of South Africa (grant no. 89967) as well as the National Institute for Theoretical and Computational Sciences (NITheCS) research programme: Advancing Biodiversity Informatics and Ecological Modelling. Jasper Slingsby, Glenn Moncrieff and Vernon Visser report financial support was provided by National Research Foundation of South Africa (grant no. 118593).

Acknowledgments

We thank Dr. Mervyn Lötter, Dr. Pierre-Cyril Renaud and Guin Zambatis for helpful discussions. SM, RG, PL and CH were supported by the National Research Foundation of South Africa (NRF, grant 89967); JS, GM and VV were supported by NRF (grant 118593). SM, VV and CH were also supported by the National Institute for Theoretical and Computational Sciences (NITheCS) research programme: Advancing Biodiversity Informatics and Ecological Modelling.

Appendix A. Supplementary data

Table S1: List of all biodiversity data stakeholders in SA, collated to encourage collaborative partnerships that can leverage combined expertise and datasets. Table S2: List of online biodiversity databases highlighting the magnitude of the challenge behind heterogeneous data integration. Table S3: Challenges (red) and opportunities (green) for biodiversity informatics in South Africa illustrated in Figure 1.

Data availability

No data was used for the research described in the article.

References

Ambler, J., Diallo, A.A., Dearden, P.K., Wilcox, P., Hudson, M., Tiffin, N., 2021. Including digital sequence data in the Nagoya protocol can promote data sharing. Trends Biotechnol. 39 (2), 116–125. https://doi.org/10.1016/j.tibtech.2020.06.009.

Anonymous, 2020. New biodiversity targets cannot afford to fail. Nature 78 (7795), 337–338. https://doi.org/10.1038/d41586-020-00450-5.

Appel, M., Pebesma, E., 2019. On-demand processing of data cubes from satellite image collections with the gdalcubes library. Data 4 (3), 1–16. https://doi.org/10.3390/data4030092.

Ball-Damerow, J.E., Brenskelle, L., Barve, N., Soltis, P.S., Sierwald, P., Bieler, R., et al., 2019. Research applications of primary biodiversity databases in the digital age. PLoS ONE 14 (9), e0215794. https://doi.org/10.1371/journal.pone.0215794.

Bayraktarov, E., Ehmke, G., O'Connor, J., Burns, E.L., Nguyen, H.A., McRae, L., Possingham, H.P., Lindenmayer, D.B., 2019. Do big unstructured biodiversity data mean more knowledge? Front. Ecol. Evol. 6, 239. https://doi.org/10.3389/fevo.2018.00239.

Belbin, L., Wallis, E., Hobern, D., Zerger, A., 2021. The Atlas of Living Australia: history, current state and future directions. Biodivers. Data J., e65023 https://doi.org/10.3897/BDJ.9.e65023.

Biggs, R., Peterson, G.D., Rocha, J.C., 2018. The regime shifts database: a framework for analyzing regime shifts in social-ecological systems. Ecol. Soc. 23 (3), 9. https://doi.org/10.5751/ES-10264-230309.

Bingham, H., Doudin, M., Weatherdon, L., Despot-Belmonte, K., Wetzel, F., Groom, Q., et al., 2017. The biodiversity informatics landscape: elements, connections and opportunities. Res. Ideas Outcomes 3, e14059. https://doi.org/10.3897/rio.3.e14059.

Blair, J., Gwiazdowski, R., Borrelli, A., Hotchkiss, M., Park, C., Perrett, G., Hanner, R., 2020. Towards a catalogue of biodiversity databases: an ontological case study. Biodivers. Data J. 8, e32765 https://doi.org/10.3897/BDJ.8.e32765.

Boakes, E.H., McGowan, P.J.K., Fuller, R.A., Chang-qing, D., Clark, N.E., O'Connor, K., Mace, G.M., 2010. Distorted views of biodiversity: spatial and temporal bias in species occurrence data. PLoS Biol. 8 (6), e1000385 https://doi.org/10.1371/journal.pbio.1000385.

Butchart, S.H.M., Walpole, M., Collen, B., van Strien, A., Scharlemann, J.P.W., Almond, R.E.A., 2010. Global biodiversity: indicators of recent declines. Science 328, 1164. https://doi.org/10.1126/science.1187512.

CBD, 2020. Global Biodiversity Outlook 5. Secretariat of the Convention on Biological Diversity. Montreal, Canada.

CBD, 2021. Press Release: A New Global Framework for Managing Nature Through 2030: 1st Detailed Draft Agreement Debuts. URL: shorturl.at/cltP9.

Ceballos, G., Ehrlich, P.R., Barnosky, A.D., García, A., Pringle, R.M., et al., 2015. Accelerated modern human-induced species losses: entering the sixth mass extinction. Science 373 (6550), 56–60. https://doi.org/10.1126/science.abh0945.

Coetzer, W., Hamer, M., 2019. Managing South African biodiversity research data: meeting the challenges of rapidly developing information technology. S. Afr. J. Sci. 115 (3/4) https://doi.org/10.17159/sajs.2019/5482.

Coleman, C.O., Radulovici, A.E., 2020. Challenges for the future of taxonomy: talents, databases and knowledge growth. Megataxa 001 (1), 028–034. https://doi.org/10.11646/megataxa.1.1.5.

Cornford, R., Deinet, S., De Palma, A., Hill, S.L.L., McRae, L., Pettit, B., et al., 2020. Fast, scalable, and automated identification of articles for biodiversity and macroecological datasets. Glob. Ecol. Biogeogr. 30, 339–347. https://doi.org/10.1111/geb.13219.

Cox, M., Gurney, G.G., Anderies, J.M., Coleman, E., Darling, E., Epstein, G.B., Frey, U., Nenadovic, M., Schlager, E., Villamayor-Tomas, S., 2021. Lessons learned from synthetic research projects based on the ostrom workshop frameworks. Ecol. Soc. 26 (1), 17. https://doi.org/10.5751/ES-12092-260117.

De Vos, A., Biggs, R., Preiser, R., 2019. Methods for understanding social-ecological systems: a review of place-based studies. Ecol. Soc. 24 (4), 16. https://doi.org/10.5751/ES-11236-240416.

Díaz, S., Settele, J., Ngo, E.S., Agard, H.T., Arneth, J., Balvanera, A., et al., 2019. Pervasive human-driven decline of life on Earth points to the need for transformative change. Science 366 (6471), eaax3100. https://doi.org/10.1126/science.aax3100.

Digital Science Report, 2020. The State of Open Data 2020: The Longest-running Longitudinal Survey and Analysis on Open Data. Digital Science and Figshare, London, UK info@figshare.comdoi:10.6084/m9.figshare.13227875.

DIRISA, 2021. Data Intensive Research Initiative of South Africa (DIRISA). Accessed August 2021. National Integrated Cyberinfrastructure System. https://www.dirisa.ac.za.

Enquist, B.J., Condit, R., Peet, R.K., Schildhauer, M., Thiers, B.M., 2016. Cyberinfrastructure for an integrated botanical information network to investigate the ecological impacts of global climate change on plant biodiversity. PeerJ 4, e2615v2. https://doi.org/10.7287/peerj.preprints.2615v2. Preprints.

Escamilla Molgora, J.M., Sedda, L., Atkinson, P.M., 2020. Biospytial: spatial graph-based computing for ecological big data. GigaScience 9, 1–25. https://doi.org/10.1093/gigascience/giaa039.

Gadelha Jr., L.M.R., de Siracusa, P.C., Dalcin, E.C., Estev˜ao da Silva, L.A., Augusto, D.A., Krempser, E., et al., 2020. A survey of biodiversity informatics: concepts, practices, and challenges. WIREs Data Min. Knowl. Discovery 11, e1394. https://doi.org/10.1002/widm.1394.

GBIF, 2021. GBIF Occurrence Download. GBIF.org. https://doi.org/10.15468/dl.htygre. Accessed 26 May 2021.

Geller, G.N., Cavender-Bares, J., Gamon, J.A., McDonald, K., Podest, E., Townsend, P.A., Ustin, S., 2020. Epilogue: toward a global biodiversity monitoring system. In: Cavender-Bares, J., Gamon, J.A., Townsend, P.A. (Eds.), Remote Sensing of Plant Biodiversity. Springer Open, pp. 519–526. https://doi.org/10.1007/978-3-030-33157-3_20.

Gillson, L., Biggs, H., Smit, I.P.J., Virah-Sawmy, M., Rogers, K., 2019. Finding common ground between adaptive management and evidence-based approaches to biodiversity conservation. Trends Ecol. Evol. 34 (1), 31–44. https://doi.org/10.1016/j.tree.2018.10.003.

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., Moore, R., 2017. Google earth engine: planetary-scale geospatial analysis for everyone. Remote Sens. Environ. 202, 18–27. https://doi.org/10.1016/j.rse.2017.06.031.

Hansen, A.J., Noble, B.P., Veneros, J., East, A., Goetz, S.J., Supples, C., et al., 2021. Toward monitoring forest ecosystem integrity within the post-2020 global biodiversity framework. Conserv. Lett., e12822 https://doi.org/10.1111/conl.12822.

Hardisty, A.R., Michener, W.K., Agosti, D., García, E.A., Bastin, L., Belbin, L., et al., 2019. The Bari manifesto: an interoperability framework for essential biodiversity variables. Eco. Inform. 49, 22–31. https://doi.org/10.1016/j.ecoinf.2018.11.003.

Heberling, J.M., Miller, J.T., Noesgaard, D., Weingart, S.B., Schigel, D., 2021. Data integration enables global biodiversity synthesis. PNAS 118 (6), e2018093118. https://doi.org/10.1073/pnas.2018093118.

Heidorn, P.B., 2008. Shedding light on the dark data in the long tail of science. Libr. Trends 57 (2), 280–299. https://doi.org/10.1353/lib.0.0036.

Hobern, D., Apostolico, A., Arnaud, E., Bello, J.C., Canhos, D., Dubois, G., 2013. Global Biodiversity Informatics Outlook: Delivering Biodiversity Knowledge in the Information Age. Global Biodiversity Information Facility Secretariat, Copenhagen.

Hortal, J., de Bello, F., Diniz-Filho, J.A.F., Lewinsohn, T., Lobo, J.M., Ladle, R.J., 2015. Seven shortfalls that beset large-scale knowledge of biodiversity. Annu. Rev. Ecol. Evol. Syst. 46, 523–549. https://doi.org/10.1146/annurev-ecolsys-112414-054400.

Hoskins, A.J., Harwood, T.D., Ware, C., Williams, K.J., Perry, J.J., Ota, N., et al., 2020. BILBI: supporting global biodiversity assessment through high-resolution macroecological modelling. Environ. Model. Softw. 132, 104806 https://doi.org/10.1101/309377.

Hudson, M., Nanibaa, A.G., Sterling, R., Caron, N.R., Fox, K., Yracheta, J., Anderson, J., Wilcox, P., Arbour, L., Brown, A., Taualii, M., 2020. Rights, interests and expectations: indigenous perspectives on unrestricted access to genomic data. Nat. Rev. Genet. 21 (6), 377–384. https://doi.org/10.1038/s41576-020-0228-x.

Hugo, S., Altwegg, R., 2017. The second southern African bird atlas project: causes and consequences of geographical sampling bias. Ecol. Evol. 7, 6839–6849. https://doi.org/10.1002/ece3.3228.

ICASA, 2020. The State of the ICT Sector Report in South Africa. Independent Communications Authority of South Africa. March 2020.

IPBES, 2020. Summary for Policymakers of the Global Assessment Report on Biodiversity and Ecosystem Services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services. https://doi.org/10.5281/zenodo.3553579.Accessed 15 July 2021.

IPCC, 2022. In: P¨ortner, H.-O., Roberts, D.C., Tignor, M., Poloczanska, E.S., Mintenbeck, K., Alegría, A., Craig, M., Langsdorf, S., L¨oschke, S., M¨oller, V., Okem, A., Rama, B. (Eds.), Climate Change 2022: Impacts, Adaptation, and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press. In Press.

Ivanova, N.V., Shashkov, M.P., 2021. The possibilities of GBIF data use in ecological research. Russ. J. Ecol. 52 (1), 1–8. https://doi.org/10.1134/S1067413621010069.

Jucker, T., Wintle, B., Shackelford, G., Bocquillon, P., Geffert, J.L., Kasoar, T., et al., 2018. Ten-year assessment of the 100 priority questions for global biodiversity conservation. Conserv. Biol. 32 (6), 1457–1463. https://doi.org/10.1111/cobi.13159.

Kissling, W.D., Hardisty, A., García, E.A., Santamaria, M., De Leo, F., Pesole, G., et al., 2015. Towards global interoperability for supporting biodiversity research on essential biodiversity variables (EBVs). Biodiversity 16 (2–3), 99–107. https://doi.org/10.1080/14888386.2015.1068709.

KoBoToolbox, 2021. KoBoToolbox. Accessed August 2021. Harvard Humanitarian Initiative. https://www.kobotoolbox.org/.

Kosta, A., 2017. Age of Information: A New Concept, Metric, and Tool. Hanover, Massachusetts. ISBN 978-1-68083-361-4.

Kruger, J.M., MacFadyen, S., 2011. Science support within the South African National Parks adaptive management framework. Koedoe 53 (2), 1–7. https://doi.org/10.4102/koedoe.v53i2.1010.

Kuras, E.R., Warren, P.S., Zinda, J.A., Aronson, M.F.J., Cilliers, S., Goddard, M.A., Nilon, C.H., Winkler, R., 2020. Urban socioeconomic inequality and biodiversity often converge, but not always: a global meta-analysis. Landsc. Urban Plan. 198, 103799 https://doi.org/10.1016/j.landurbplan.2020.103799.

La Salle, J., Williams, K.J., Moritz, C., 2016. Biodiversity analysis in the digital era. Philos. Trans. R. Soc. B 371, 20150337. https://doi.org/10.1098/rstb.2015.0337.

Maldonado, C., Molina, C.I., Zizka, A., Persson, C., Taylor, C.M., Alb´an, J., et al., 2015. Estimating species diversity and distribution in the era of Big Data: to what extent can we trust public databases? Glob. Ecol. Biogeogr. 24 (8), 973–984. https://doi.org/10.1111/geb.12326.

Martellos, S., Attorre, F., 2012. New trends in biodiversity informatics. Plant Biosyst. 146 (4), 749–751. https://doi.org/10.1080/11263504.2012.740092.

Michener, W.K., 2015. Ecological data sharing. Eco. Inform. 29, 33–44. https://doi.org/10.1016/j.ecoinf.2015.06.010.

Musvuugwa, T., Dlomu, M.G., Adebowale, A., 2021. Big data in biodiversity science: a framework for engagement. Technologies 9 (60). https://doi.org/10.3390/technologies9030060.

Naisbitt, J., 1982. Megatrends: Ten New Directions Transforming Our Lives. Warner Books, New York. ISBN: 0446512516.

Navarro, L.M., Fern´andez, N., Guerra, C., Guralnick, R., Kissling, W.D., et al., 2017. Monitoring biodiversity change through effective global coordination. Curr. Opin. Environ. Sustain. 29, 158–169. https://doi.org/10.1016/j.cosust.2018.02.005.

Newbold, T., Hudson, L.N., Arnell, A.P., et al., 2016. Has land use pushed terrestrial biodiversity beyond the planetary boundary? A global assessment. Science 353, 288–291. https://doi.org/10.1126/science.aaf2021.

Noss, R., 1990. Indicators for monitoring biodiversity: a hierarchical approach. Conserv. Biol. 4 (4), 355–364. http://www.jstor.org/stable/2385928.

Open Data Kit, 2021. Open Data Kit. Accessed August 2021. https://opendatakit.org/

Osawa, T., 2019. Perspectives on biodiversity informatics for ecology. Ecol. Res. 34, 446–456, 0.1111/1440-1703.12023.

Owen, D., Livermore, L., Groom, Q., Hardisty, A., Leegwater, T., van Walsum, M., Wijkamp, N., Spasi´c, I., 2020. Towards a scientific workflow featuring natural language processing for the digitisation of natural history collections. Res. Ideas Outcomes 6, e55789. https://doi.org/10.3897/rio.6.e55789.

Pereira, H.M., Ferrier, S., Walters, M., Geller, G.N., Jongman, R.H.G., Scholes, R.J., et al., 2013. Essential biodiversity variables. Science 339 (6117), 277–278. https://doi.org/10.1126/science.1229931.

Petersen, T.K., Speed, J.D.M., Grøtan, V., Austrheim, G., 2021. Species data for understanding biodiversity dynamics: the what, where and when of species occurrence data collection. Ecol. Solutions Evid. 2 (1), e12048 https://doi.org/10.1002/2688-8319.12048.

Pettorelli, N., Barlow, J., Nu˜nez, M.A., Rader, R., Stephens, P.A., Pinfield, T., Newton, E., 2021. How international journals can support ecology from the global south. J. Appl. Ecol. 58 (1), 4–8. https://doi.org/10.1111/1365-2664.13815.

Pickens, A.H., Hansen, M.C., Hancher, M., Stehman, S.V., Tyukavina, A., Potapov, P., Marroquin, B., Sherani, Z., 2020. Mapping and sampling to characterize global inland water dynamics from 1999 to 2018 with full Landsat time-series. Remote Sens. Environ. 243, 111792 https://doi.org/10.1016/j.rse.2020.111792.

Proença, V., Martin, L.J., Pereira, H.M., Fernandez, M., McRae, L., Belnap, J., et al., 2017. Global biodiversity monitoring: from data sources to essential biodiversity variables. Biol. Conserv. 213, 256–263. https://doi.org/10.1016/j.biocon.2016.07.014.

Raymond, C.M., Cebrián-Piqueras, M.A., Andersson, R., Andrade, R., Schnell, A.A., Romanelli, B.B., et al., 2022. Inclusive conservation and the Post-2020 Global Biodiversity Framework: Tensions and prospects. One Earth 5 (3), 252–264. https://doi.org/10.1016/j.oneear.2022.02.008.

SAEON, 2021. South African Environmental Observation Network (SAEON) data portal. Accessed August 2021. National Research Foundation. http://www.saeon.ac.za/data-portal-access.

SANReN, 2021. South African National Research Network (SANReN). Accessed August 2021. National Integrated Cyberinfrastructure System. https://sanren.ac.za/.

Scholes, R.J., Walters, M., Turak, E., Saarenmaa, H., Heip, C.H.R., Tuama,´E.´O., Faith, D.P., Mooney, H.A., Ferrier, S., Jongman, R.H.G., Harrison, I.J., Yahara, T., Pereira, H.M., Larigauderie, A., Geller, G., 2012. Building a global observing system for biodiversity. Curr. Opin. Environ. Sustain. 4 (1), 139–146. https://doi.org/10.1016/j.cosust.2011.12.005.

Schurr, F.M., Esler, K.J., Slingsby, J.A., Allsopp, N., 2012. Fynbos proteaceae as model organisms for biodiversity research and conservation. S. Afr. J. Sci. 108 (11–12), 12–16. https://doi.org/10.4102/sajs.v108i11/12.1446.

Shin, N., Shibata, H., Osawa, T., Yamakita, T., Nakamura, M., Kenta, T., 2020. Toward more data publication of long-term ecological observations. Ecol. Res. 35, 700–707. https://doi.org/10.1111/1440-1703.12115.

Siddig, A.A.H., 2019. Why is biodiversity data-deficiency an ongoing conservation dilemma in Africa? J. Nat. Conserv. 50, 125719 https://doi.org/10.1016/j.jnc.2019.125719.

Simoes, R., Camara, G., Queiroz, G., Souza, F., Andrade, P.R., Santos, L., Carvalho, A., Ferreira, K., 2021. Satellite image time series analysis for big earth observation data. Remote Sens. 13, 1–20. https://doi.org/10.3390/rs13132428.

Slingsby, J.A., Buys, A., Simmers, A.D.A., Prinsloo, E., Forsyth, G.G., Glenday, J., Allsopp, N., 2021. Jonkershoek: Africa's oldest catchment experiment - 80 years and counting. Hydrol. Process. 35 (4), 1–7. https://doi.org/10.1002/hyp.14101.

Smit, I.P.J., Riddell, E.S., Cullum, C., Petersen, R., 2013. Kruger National Park research supersites: establishing long-term research sites for cross-disciplinary, multiscaled learning. Koedoe 55 (1), a1107. https://doi.org/10.4102/koedoe.v55i1.1107.

Smith, M.D., van Wilgen, B.W., Burns, C.E., Govender, N., Potgieter, A.L.F., Andelman, S., Biggs, H.C., Botha, J., Trollope, W.S.W., 2013. Long-term effects of fire frequency and season on herbaceous vegetation in savannas of the Kruger National Park, South Africa. J. Plant Ecol. 6 (1), 71–83. https://doi.org/10.1093/jpe/rts014.

Sreenivasaiah, P.K., Kim, D.H., 2010. Current trends and new challenges of databases and web applications for systems driven biological research. Front. Physiol. 1, 147. https://doi.org/10.3389/fphys.2010.00147.

Stefanoudis, P.V., Licuanan, W.Y., Morrison, T.H., Talma, S., Veitayaki, J., Woodall, L.C., 2021. Turning the tide of parachute science. Curr. Biol. 31 (4), 184–185. https://doi.org/10.1016/j.cub.2021.01.029.

Stephenson, P.J., Stengel, C., 2020. An inventory of biodiversity data sources for conservation monitoring. PLoS ONE 15 (12), e0242923. https://doi.org/10.1371/journal.pone.0242923.

Stephenson, P.J., Bowles-Newark, N., Regan, U., Stanwell-Smith, D., Diagana, M., H¨oft, R., et al., 2017a. Unblocking the flow of biodiversity data for decision-making in Africa. Biol. Conserv. 213, 335–340. https://doi.org/10.1016/j.biocon.2016.09.003.

Stephenson, P.J., Brooks, T.M., Butchart, S.H.M., Fegraus, E., Geller, G.N., Hoft, R., et al., 2017b. Priorities for big biodiversity data. Front. Ecol. Environ. 15 (3), 124–125. https://doi.org/10.1002/fee.1473.

Stieglitz, S., Wilms, K., Mirbabaie, M., Hofeditz, L., Brenger, B., L´opez, A., et al., 2020. When are researchers willing to share their data? Impacts of values and uncertainty on open data in academia. PLoS ONE 15 (7), e0234172. https://doi.org/10.1371/journal.pone.0234172.

Survey123, 2021. ArcGIS Survey123. Accessed August 2021. Environmental Systems Research Institute (ESRI). https://survey123.arcgis.com/.

TENET, 2021. Tertiary Education and Research Network of South Africa (TENET). Accessed August 2021. https://graphs.tenet.ac.za/iris/api2/tenet/home.

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A.U., Wu, L., Read, E., et al., 2011. Data sharing by scientists: practices and perceptions. PLoS One 6 (6), 1–21. https://doi.org/10.1371/journal.pone.0021101.

Thessen, A.E., Poelen, J.H., Collins, M., Hammock, J., 2018. 20 GB in 10 minutes: a case for linking major biodiversity databases using an open sociotechnical infrastructure and a pragmatic, cross-institutional collaboration. PeerJ Comput. Sci. 4, e164 https://doi.org/10.7717/peerj-cs.164.

Tulloch, A., Mustin, K., Possingham, H.P., Szabo, J.K., Wilson, K.A., 2013. To boldly go where no volunteer has gone before: predicting volunteer activity to prioritize surveys at the landscape scale. Divers. Distrib. 19, 465–480. https://doi.org/10.1111/j.1472-4642.2012.00947.x.

Vanhove, M.P., Rochette, A.J., de Bisthoven, L.J., 2017. Joining science and policy in capacity development for monitoring progress towards the Aichi biodiversity targets in the global south. Ecol. Indic. 73, 694–697. https://doi.org/10.1016/j.ecolind.2016.10.028.

Vargas-Solar, G., Zechinelli-Martini, J.L., Espinosa-Oviedo, J.A., 2017. Big data management: what to keep from the past to face future challenges? Data Sci. Eng. 2, 328–345. https://doi.org/10.1007/s41019-017-0043-3.

Vines, T.H., Albert, A.Y.K., Andrew, R.L., D´ebarre, F., Bock, D.G., Franklin, M.T., Gilbert, K.J., Moore, J.-S., Renaut, S., Rennison, D.J., 2013. The availability of research data declines rapidly with article age. Curr. Biol. 24, 94–97. https://doi.org/10.1016/j.cub.2013.11.014.

Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., D¨oring, M., Giovanni, R., et al., 2012. Darwin Core: an evolving community-developed biodiversity data standard. PLoS ONE 7 (1), e29715. https://doi.org/10.1371/journal.pone.0029715.

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., et al., 2016. The FAIR guiding principles for scientific data management and stewardship. Sci. Data 3, 160018. https://doi.org/10.1038/sdata.2016.18.

Xu, H., Cao, Y., Yu, D., Cao, M., He, Y., Gill, M., Pereira, H.M., 2021. Ensuring effective implementation of the post-2020 global biodiversity targets. Nat. Ecol.Evol. 5 (4), 411–418. https://doi.org/10.1038/s41559-020-01375-y.

Yesson, C., Brewer, P.W., Sutton, T., Caithness, N., Pahwa, J.S., Burgess, M., 2007. How global is the global biodiversity information facility? PLoS ONE 2 (11), e1124. https://doi.org/10.1371/journal.pone.0001124.

Yoccoz, N.G., Nichols, J.D., Boulinier, T., 2001. Monitoring of biological diversity in space and time. Trends Ecol. Evol. 16, 446–453. https://doi.org/10.1016/S0169-5347(01)02205-4.

Zizka, A., Antunes Carvalho, F., Calvente, A., Rocio Baez-Lizarazo, M., Cabral, A., Coelho, J.F.R., et al., 2020. No one-size-fits-all solution to clean GBIF. PeerJ 8, e9916. https://doi.org/10.7717/peerj.9916.