sustainability

MDPI

Article

Completeness of Digitally Accessible Knowledge of Plants Across Africa and Priorities for Future Data Discovery

Jean Cossi Ganglo

Laboratory of Forest Sciences (LSF), Faculty of Agricultural Sciences (FSA), University of Abomey-Calavi (UAC), Cotonou P.O. Box 526, Benin; ganglocj@gmail.com

Abstract

Digitally accessible knowledge (DAK) is of the utmost importance for biodiversity conservation. The Global Biodiversity Information Facility (GBIF) is a mega data infrastructure with more than three billion and seventy million (3,070,000,000) occurrence records as of 4 March 2025. It is by far the largest initiative assembling and sharing DAK to support scientific research, conservation, and sustainable development. We analyzed plant data published on the GBIF site in Africa to highlight the contribution of the continent to the GBIF and thereby highlight data quality issues and data gaps across taxonomic groups and geographic space. Therefore, we downloaded data on 17 January 2023 from the Plantae kingdom in Africa. We performed data treatment and analysis via R using several packages and related functions. Although Africa is home to rich biodiversity with many hotspots, the global data contribution of the continent to the GBIF (61,176,994 as of 17 January 2023) is still extremely low (2.69%). Furthermore, there are large disparities between African countries, with South Africa contributing far more than 50% of the continent's data alone. The plant data from Africa (9,116,401 occurrence records) accounted for 14.90% of the data from the continent; this underlines enormous gaps between taxonomic groups. We noted important data loss during the process of data cleaning, clearly underlining the limited data quality from the continent; indeed, the data fitness for completeness analysis was only 50.94% of the total data records initially downloaded. Efforts toward quality checks before data publication on the GBIF site are still needed across African countries. Magnoliopsida was the dominant plant class with the highest number of records (71.07%) and the highest number of species (68.36%), followed by Liliopsida, with 22.80% of the records and 19.06% of the species. In geographic space, plant data gaps are also quite large across the continent; data completeness is greater in West Africa, Southern Africa, East Africa, and Madagascar. To account for the non-normal distribution of the data, robust correlation methods and robust mean comparison methods were used. According to the results, accessibility by rivers and roads and accessibility to protected areas are limiting factors for data completeness across the continent. The large multidimensional data gaps identified in this study and the important data loss noted during the data cleaning process should be prioritized in future data collections across the continent.

Keywords: digital accessible knowledge (DAK); biodiversity data; completeness analysis; data gaps; GBIF; Africa

Check for updates

Received: 18 June 2025
Revised: 8 August 2025
Accepted: 9 August 2025
Published: 30 October 2026

Copyright: © 2026 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Digitally accessible knowledge (DAK) consists of primary biodiversity data that are digitized, published, and made accessible on international repositories worldwide (Sousa-

Baena et al.) [1]. They are indispensable for informing decisions on biodiversity conservation, particularly in the context of climate and global changes [2–7]. DAK completeness across taxonomic groups and temporal and spatial patterns, along with data quality control, is indispensable for its usefulness in research [8–11], the effectiveness of data use and biodiversity conservation [8,11–13] and the design of new useful data collections to enhance data use to inform decisions. Data pattern assessment with respect to the basis of records is also quite informative of the reliability and quality of the nature of data records and can help design more quality-improved new data collections. Africa is rich in biodiversity. Indeed, with respect to plant species richness, the dataset of approximately 6000 plant species used in the redefinition of African hotspots of biodiversity represented only 10% to 15% of the flora of Africa (Küper et al. [14]), which can, therefore, be estimated at 40,000–60,000 plant species (Lebrun and Stork [15]) for tropical Africa and at approximately 80,000 accepted taxa for all of Africa (African Plant Database [16]).

In line with the present work, data completeness analyses were diversely conducted at national and regional scales across Africa. In Benin, data completeness analysis was performed on the DAK of plants to guide future inventory and data discovery efforts in the country [3]. Many spatial and environmental gaps existed in the completeness of the data obtained, particularly in the northernmost departments of the country. The completeness of knowledge of African palms was also analyzed with a focus on Benin to identify gaps in the knowledge of that plant group [17]. The authors reported a high value of inventory completeness across the country along with remote, poorly known and less accessible areas. In Ghana, inventory completeness was assessed on the DAK of the plants to prioritize future surveys and inventories [18]. The central and northern parts of the country contained the largest spatial gaps. In western Africa, completeness analysis was also performed on the DAK of birds [19] to guide future avian surveys across the subregion. The results revealed marked spatial, seasonal, and environmental information gaps in the data coverage. At larger scales, Küper et al. [14], in their redefinition of Africa hotspots of biodiversity, mapped the plant distribution to a one-degree grid resolution covering sub-Saharan Africa south of 17° north latitude. They noted that the redefined hotspots presented substantial overlap with the 25 hotspots identified worldwide by Myers et al. [20]. In their study of the patterns of biodiversity at the continental scale, La Ferla et al. [21] mapped 3563 species, approximately 9% of the known African angiosperm flora, to one-degree latitude–longitude grid cells of sub-Saharan Africa and reported that higher taxa, specifically at the genus level, can provide a powerful and accurate tool to predict large-scale patterns of species diversity in Sub-Saharan Africa, although some exceptions exist, such as the case of the rich Cape Region in South Africa. Vertebrate data, plant data, human population data, and human infrastructure data were mapped to one-degree latitude–longitude grid cells across three African tropical mountains and sub-Saharan Africa [21]. The authors then reported strong positive correlations between human density and species richness, endemism and the proportion of threatened species across the three tropical African mountains and all of sub-Saharan Africa.

All these studies have the merit of advancing the knowledge of data gaps at different scales across Africa. However, to our knowledge, no study has addressed plant data completeness analysis at the whole-continent scale, given the large amount of data considered in this study. Although data from Africa are still underrepresented in the more than three billion and seventy million occurrence records of GBIF (www.gbif.org), as of 4 March 2025, this study is based on more than nine million occurrence records across Africa. The objective of this paper is to answer the following research questions: (1) What are the gaps in plant data with respect to taxonomic groups and the basis of records? (2) What are the species richness patterns across the plant taxonomic groups? (3) What are the spatial

gaps in plant data across Africa, and what are the underlying factors? Answering these questions will surely help design new demand-driven and useful new data collections to enable and enhance research and data use to inform decisions on biodiversity conservation and sustainable use across the continent.

2. Methods

2.1. Data Used and Data Cleaning

To achieve our purpose, we downloaded data from the GBIF site [22] on 17 January 2023 from the Plantae kingdom in Africa. They are available at https://doi.org/10.15468/dl.p2n6um (accessed on 17 January 2023) and comprise 9,116,401 occurrence records. Data were, therefore, cleaned; in that process, we discarded data with zero or undefined geographic coordinates; data missing full dates (year, month, and day); data with undefined country codes; and data records not determined to species rank.

2.2. Data Completeness Analysis

Data analysis was performed in R (4.5.0) [23]. Using the "count" function of the package dplyr (1.1.4), cleaned data were grouped and summarized with respect to countries, the basis of records, and taxonomic groups. To achieve completeness analysis at the scale of Africa with respect to spatial grid cells of half a degree and one degree, adequate aggregation square shape file grids were created and displayed over Africa maps via QGIS [24]. The aggregated square shape files created were imported into R, and via the "over" function of the "sp" package (2.2.0), the identification codes of the aggregated shape files were retrieved and added to the occurrence data to enable completeness analysis. Using the "count" and "filter" functions of the "dplyr" package, we thereafter calculated several parameters for completeness analysis: the total number (N) of records per grid square; the total number of distinct species recorded per grid square (Sobs); the number of species detected on exactly one day (a); the number of species detected on exactly two days (b); and the species richness, which is the number of species across the taxonomic groups. We thereafter used the equations provided by Chao [25] to calculate the expected number of species per grid square (*Sexp*) as:

$$Sexp = Sobs + \frac{a^2}{2b} \quad (1)$$

More explicitly, on the basis of [25], we estimated the size of a closed population (here, in this study, the number of species in each grid cell) via a model or equation that takes into account the captures and recaptures (in our work, several plant collections from various collectors) of individuals (species) and, therefore, the heterogeneity of capture probability. From studies on different datasets, Ref. [25] reported that the estimator of the population size is less biased and more accurate when most of the captured individuals (here, species) are caught only once (a) or twice (b) in the sample.

Furthermore, Ref. [25] reported that an acceptable and more accurate estimator of the population size (*Sexp*) (here, the number of species per grid cell) with respect to the number of individuals caught (here, the number of species collected/observed in the grid cell (*Sobs*)) is as follows:

$$Sexp = Sobs + \frac{\mathrm{a}^2}{2\mathrm{b}} \quad (2)$$

The inventory completeness (C), which is the degree to which each grid square is known, was calculated as:

$$C = \frac{Sobs}{Sexp} \quad (3)$$

The calculated parameters were joined to the aggregated shape files via the "merge" function of the "sp" package. The rest of the analysis, including the design of the layout showing the different grid squares with respect to the range of C values, was performed in QGIS [24]. The remaining figures were created via the function "ggplot" of the package "ggplot2".

2.3. Analysis of the Impacts of Accessibility and Protected Areas on Data Completeness

We analyzed the impacts of rivers, roads, and protected areas on data completeness. To analyze the impact of rivers on data completeness, we downloaded the shape files of African rivers [26]; the road network we used was downloaded from the Africa–Road Network [27]; and the shape files of the protected areas were downloaded from Protected Planet [28]. The function "over" of the "sp" package was, therefore, used to retrieve the identification codes of the grid cells (half-degree and one-degree spatial grid cells); then, we added them to the respective data frames of the rivers, roads, and protected areas. The "count" and "filter" functions of the "dplyr" package were then appropriately used to calculate the lengths of rivers and roads and the number of records per grid cell (half a degree and one degree, respectively). For the protected areas, the "count" and "filter" functions were also used to calculate the number of records per grid cell in protected areas and non-protected areas, respectively. Parametric tests are not robust to non-normality (Kitchenham [29]); therefore, to account for the non-normality of data and successfully appraise the impact of protected areas on data completeness, we used the Yuen test and its bootstrap version (Mair et al. [30]), to compare the mean number of records per grid cell in protected areas and non-protected areas. To account for the non-normal data distribution in assessing the impacts of roads and rivers on data completeness, robust correlation methods such as percentage bend correlation and Winsorized correlation [30] were used.

3. Results

The appendices we refer to in the text are in the Supplementary Materials.

3.1. Data Loss During the Data Cleaning Process

In total, 9,116,401 plant records were downloaded from the GBIF site (Figure 1). They are available at https://doi.org/10.15468/dl.p2n6um (accessed on 17 January 2023). During the data cleaning process, the total number of data records identified to the species level was 7,570,244 and accounted for 83.04% of the initial data records; therefore, 16.96% of the plant records were still to be described to the species level to enable full taxon determination. The number of records with adequate coordinates was 6,461,460 and accounted for 70.88% of the initial data. The data fitness for completeness analysis (records determined up to the species level, with adequate coordinates, and full dates) was only 4,643,586, accounting for approximately 50.94% of the total data records initially downloaded from the GBIF site (Figure 2). Efforts for quality checks before data publication on the GBIF site are, therefore, still highly needed across African countries.

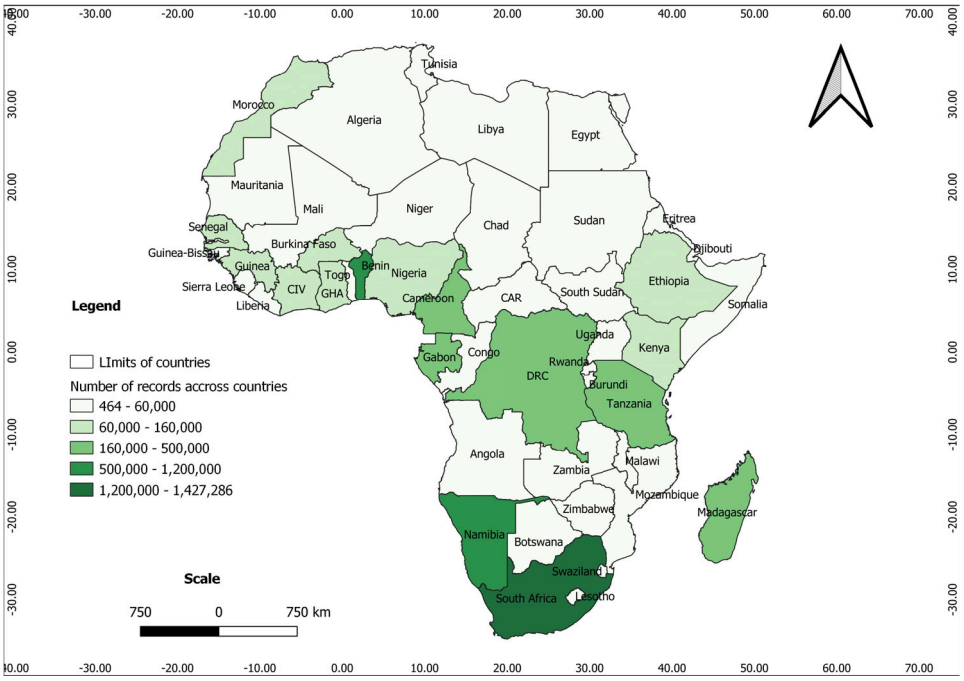


Figure 1. Terrestrial plant data records extracted from data downloaded from the GBIF site with adequate geographic coordinates (https://doi.org/10.15468/dl.p2n6um (accessed on 17 January 2023)) (the map was generated via QGIS 3.18.1 (https://api.qgis.org/api/3.18/index.html (accessed on 17 January 2023)) and WGS84 as the coordinate reference system).

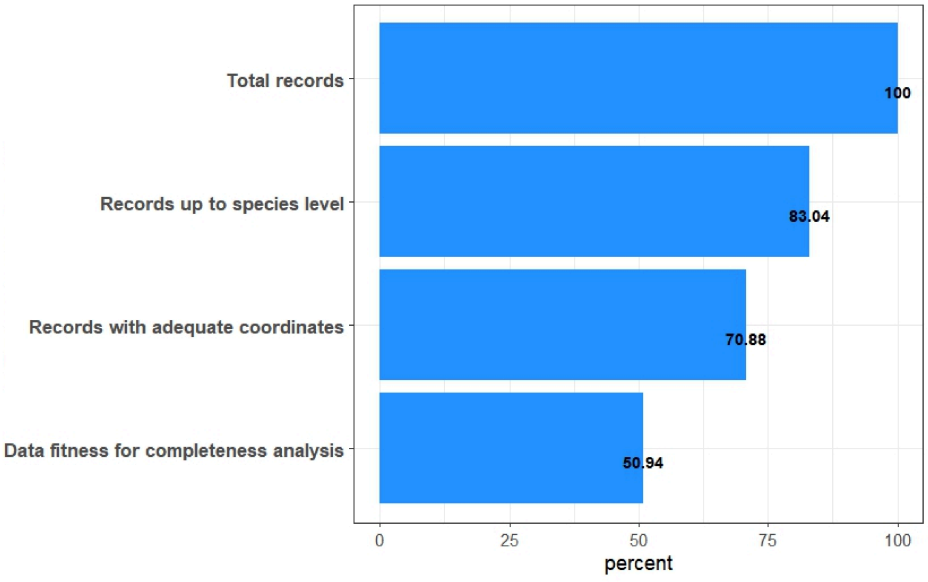


Figure 2. Data loss during the data cleaning process (the bar plots were generated via R version 4.1.3 (https://www.R-project.org/ accessed on 17 January 2023)).

3.2. Data Completeness and Species Richness Patterns Across Plant Classes

Among the 37 plant classes, Magnoliopsida had the highest number of records (71.07%), accounting for 68.36% of the species, followed by Liliopsida (22.80% of the records with 19.06% of the species) (Figure 3). This clearly highlighted that Magnoliopsida was given priority in data discovery, followed by Liliopsida; the other taxonomic groups represented about 6% of the records with about 12% of the species (Figure 3).

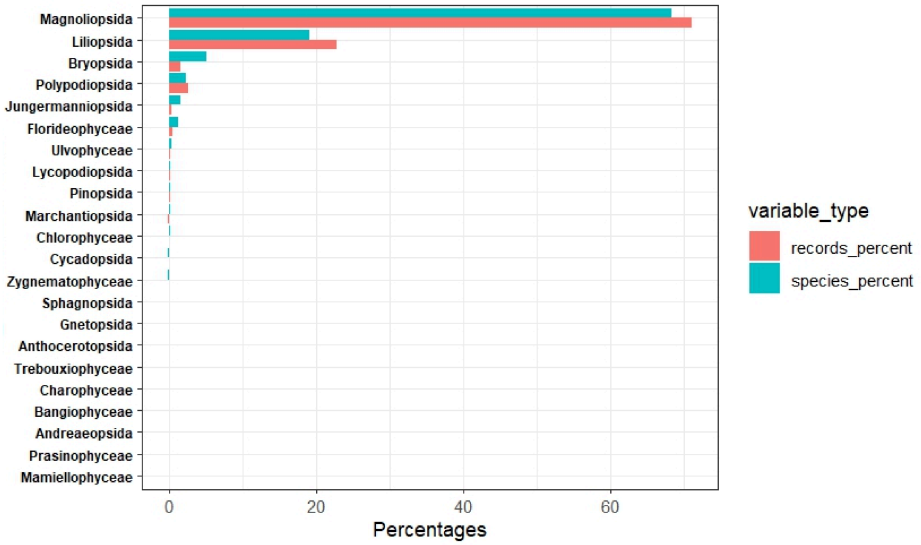


Figure 3. Diversity of plant classes sharing the maximum number of records and species (the bar plots were generated via R version 4.1.3 (https://www.R-project.org/)).

3.3. Data Completeness and Species Richness Patterns Across Plant Families

Twenty out of the 851 plant families derived from the downloaded data represented approximately 60% of the plant records from Africa and contained 53.84% of the species. The other families had less than 1% of plant data records each (Figure 4). With respect to species richness, the most dominant families were Fabaceae (8.01% of species with 12.20% of records), Asteraceae (7.66% of species with 6.47% of records), Rubiaceae (4.46% of species with 4.87% of records), and Poaceae (4.08% of species with 11.06% of records) (Figure 4).

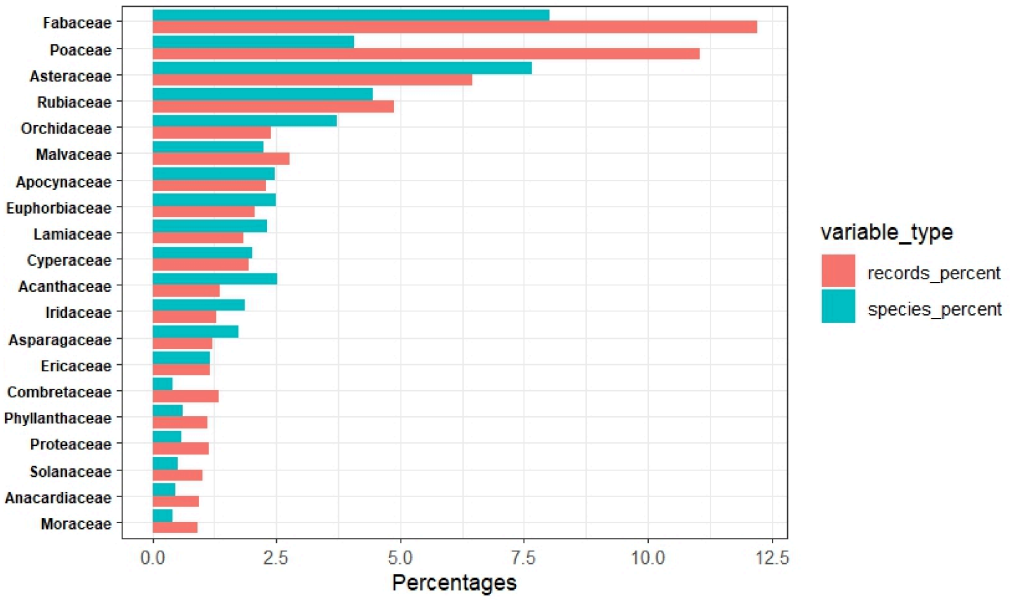


Figure 4. Species richness and record patterns across plant families (the bar plots were generated via R version 4.1.3 (https://www.R-project.org/)).

3.4. Data Completeness on the Basis of Records

The basis of records refers to the nature and method of data collection. In the data plant records from Africa, two bases of records were dominant: preserved samples (59.68%) and human observations (30.86%) (Figure 5).

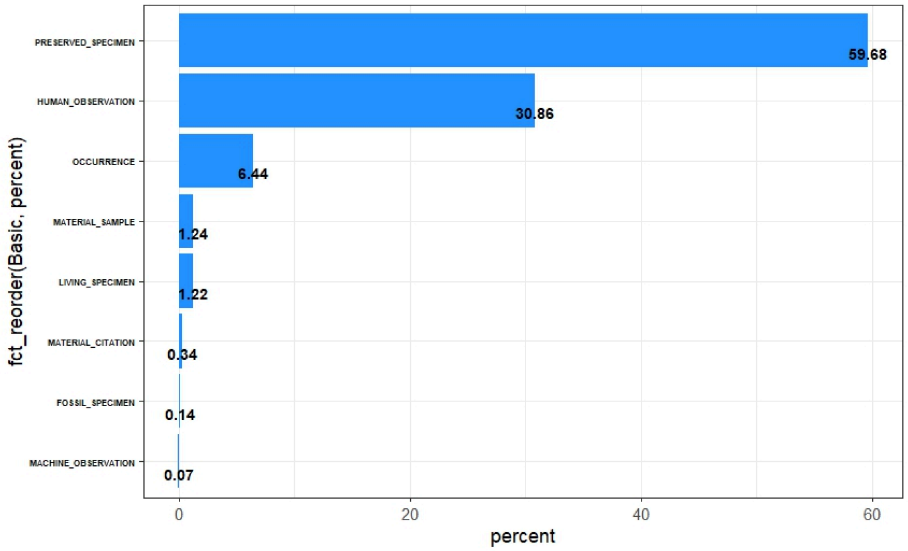


Figure 5. Diversity of the basis of records in plant records from Africa (the bar plots were generated via R version 4.1.3 (https://www.R-project.org/)).

3.5. Data Completeness and Species Richness in Geographic Space

From Figures 6 and S1, plant data gaps are quite large across the continent at either spatial resolution. Data completeness is greater in West Africa (Benin, Ghana, North Burkina Faso, South Liberia, and South Niger). In East Africa, Tanzania and Kenya are more successful with respect to data completeness. In central Africa, plant data completeness is greater in Cameroon, Chad, and Gabon. Data gaps are considerable in DRC, Sahelian, and North African countries. In southern Africa, data completeness is greater in South Africa. Madagascar is successful in plant data completeness acquisition. The species richness pattern across geographic space (Figure 7) revealed that southern Africa and northern Africa, including the Mediterranean zone, are the richest zones in terms of species. Indeed, in those two subregions, the number of records per species, which is an index of the species richness of the records, was among the lowest across the continent (22 and 29, respectively).

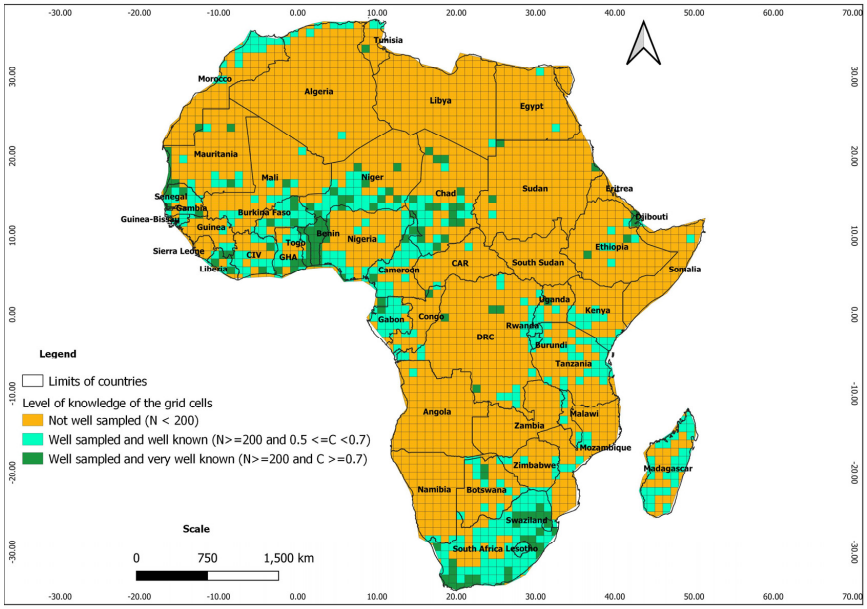


Figure 6. Geographic patterns of inventory completeness across Africa at a 01° spatial resolution (the map was generated via QGIS 3.18.1 (https://api.qgis.org/api/3.18/index.html) and WGS84 as the coordinate reference system).

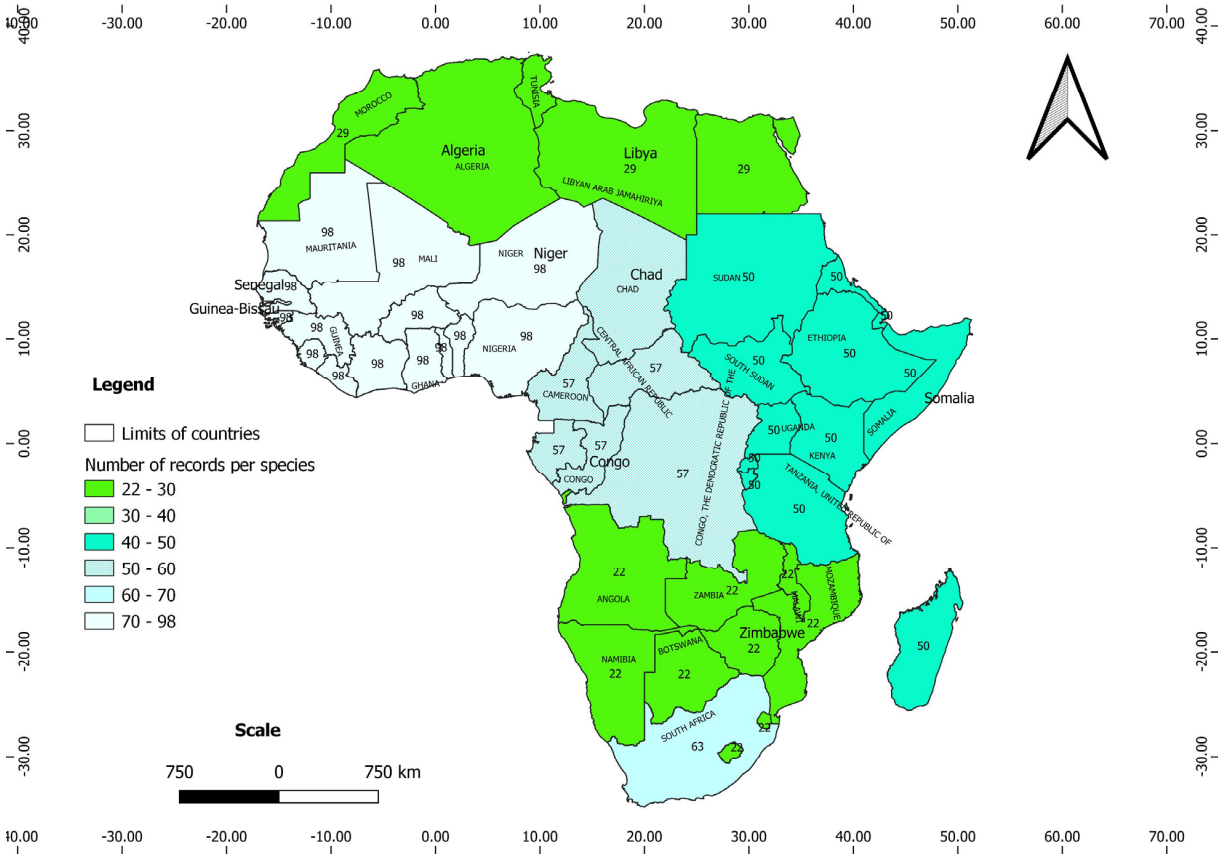


Figure 7. Geographic patterns of species richness across Africa (the map was generated via QGIS 3.18.1 (https://api.qgis.org/api/3.18/index.html) and WGS84 as the coordinate reference system).

3.6. Data Completeness Versus Accessibility and Protected Areas

Access to areas by roads significantly affects data completeness. Indeed, data completeness is correlated with road length at either a 0.5° grid cell spatial resolution or a 01° grid cell spatial resolution (Figures 8 and S2). For the 0.5° grid cell spatial resolution, with respect to the percentage bend correlation, the robust correlation coefficient = 0.4383, and the p value = 0; for the winsorized correlation, the robust correlation coefficient = 0.433, and the p value = 0. For the 01° grid cell spatial resolution, with respect to the percentage bend correlation, the robust correlation coefficient = 0.6321, and the p value = 0; for the winsorized correlation, the robust correlation coefficient = 0.6172, and the p value = 0. Data completeness is also correlated with waterway length either at a 0.5° grid cell spatial resolution or at a 01° grid cell spatial resolution (Figures 9 and S3). For the 0.5° grid cell spatial resolution, with respect to the percentage bend correlation, the robust correlation coefficient = 0.2635, and the p value = 0.0031; for the winsorized correlation, the robust correlation coefficient = 0.2592, and the p value = 0.00408. For the 01° grid cell spatial resolution, with respect to the percentage bend correlation, the robust correlation coefficient = 0.6869, and the p value = 0; for the winsorized correlation, the robust correlation coefficient = 0.6898, and the p value = 0.

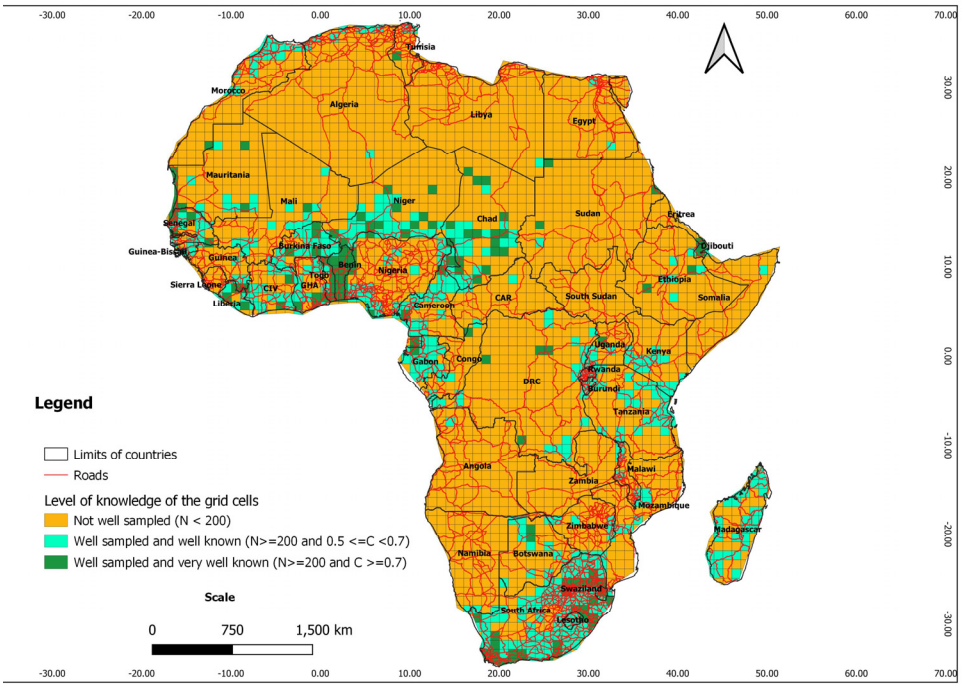


Figure 8. Geographic patterns of inventory completeness across Africa at a 01° spatial resolution versus the road layer (the map was generated via QGIS 3.18.1 (https://api.qgis.org/api/3.18/index.html) and WGS84 as the coordinate reference system).

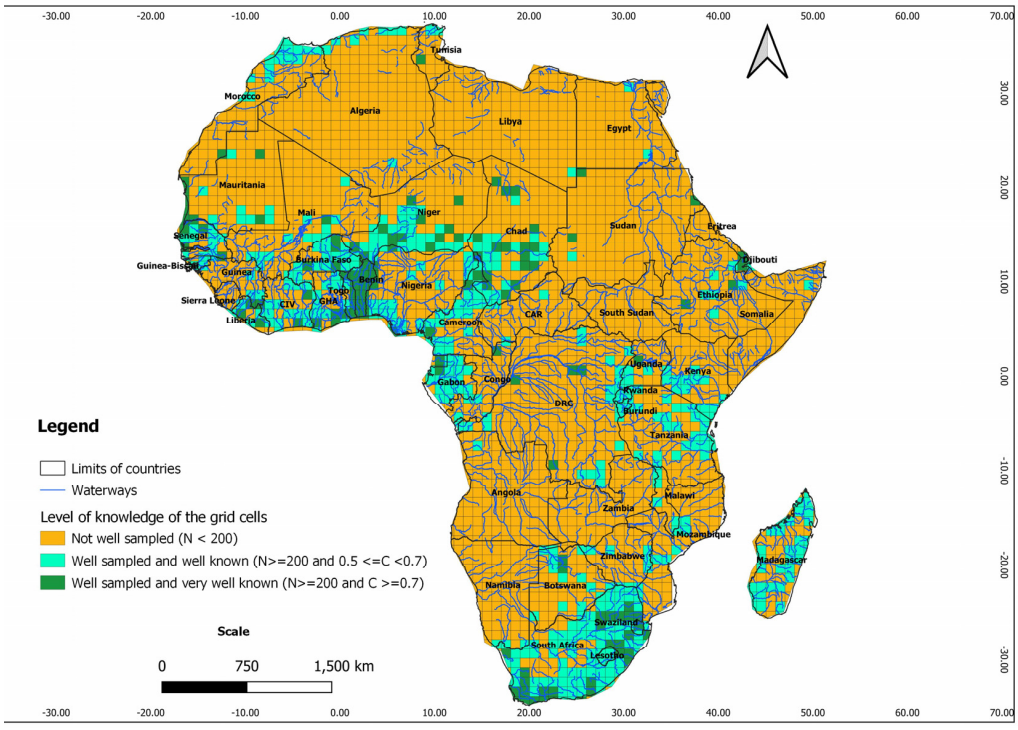


Figure 9. Geographic patterns of inventory completeness across Africa at a 01° spatial resolution versus the river layer (the map was generated via QGIS 3.18.1 (https://api.qgis.org/api/3.18/index.html) and WGS84 as the coordinate reference system).

Protected areas affect data completeness at either resolution. At the coarser resolution of the 01° spatial grid cells, the mean number of records in protected areas (50) is significantly smaller than the mean number of records in non-protected areas (97) (Figure 10) (two-sample Yuen test, trim = 0.2; t = 5.3825, df = 1204.2, p value = 8.823×10^{-8}; for the

bootstrap version of the Yuen test: tr = 0.2, nboot = 5000; test statistic: 5.3764 (df = NA), p value = 0). At the finer resolution of the 0.5° spatial grid cells (Figure S4), the mean number of records in protected areas (23) is significantly smaller than the mean number of records in non-protected areas (30) (two-sample Yuen test, trim = 0.2, t = 3.494, df = 1535.5, p value = 0.0004893; for the bootstrap version of the Yuen test: tr = 0.2, nboot = 5000; test statistic: 3.4943 (df = NA), p value = 4×10^{-4}).

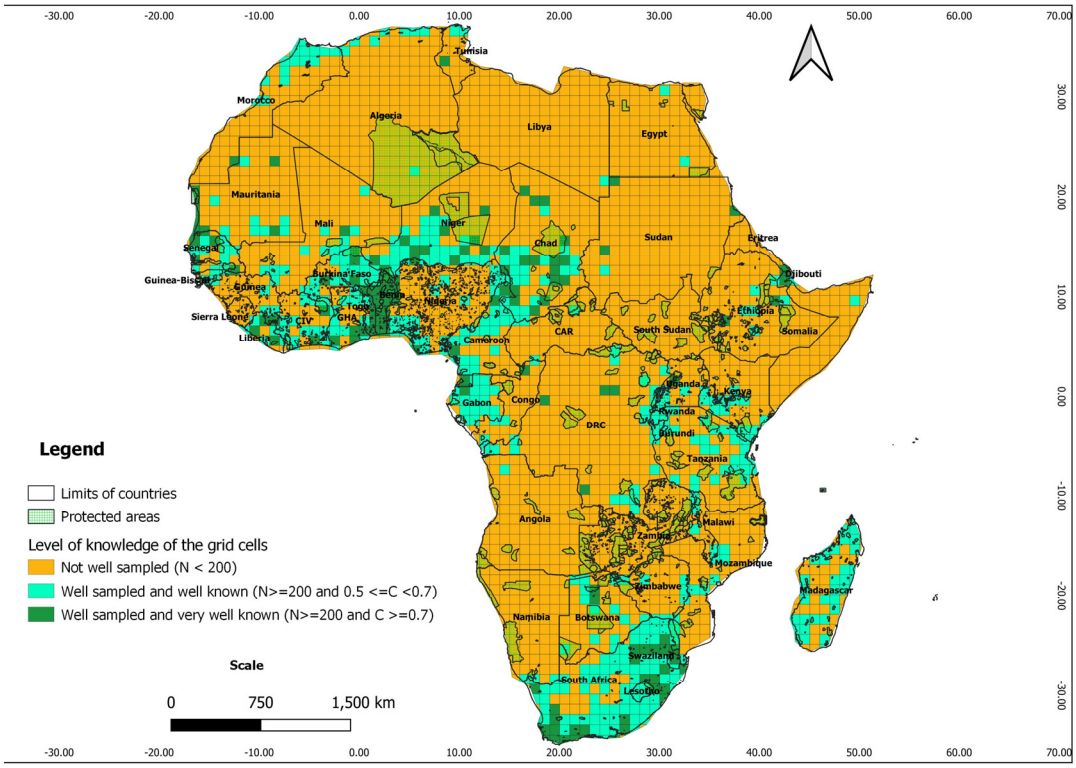


Figure 10. Geographic patterns of inventory completeness across Africa at a 01° spatial resolution versus the layer of protected areas (the map was generated via QGIS 3.18.1 (https://api.qgis.org/api/3.18/index.html) and WGS84 as the coordinate reference system).

Globally, protected areas are, therefore, less well inventoried at both resolutions. This implies that restrictions on accessibility to protected areas by forest guards are limiting factors for biodiversity knowledge.

4. Discussion

4.1. Plant Taxonomic Group Completeness, Species Richness Patterns, and Methods for New Data Collection

With respect to taxon rank completeness, the results revealed that taxons related to approximately 17% of the plant records were not determined to the species level. The information related to those records is incomplete and cannot be used in fine analyses such as ecological niche modeling or completeness analysis at the species level. However, the rate of plant taxon determination at the species level in this study was much better than what has been noted in some taxa of the animal kingdom. Indeed, Ref. [31] reported that, in Australia, only 60.9% of the taxa identified at the species level were terrestrial invertebrates. With respect to the plant species of Africa, my suggestion is to refer to specimen images or vouchers, where possible, to complement the determination of species.

Among the 37 plant classes reported by the GBIF site in Africa, two plant classes dominated the occurrence records by more than 90%. They are Magnoliopsida (71.07% of the records and 68.36% of the species) and Liliopsida (22.80% of the records and 19.06% of

the species). The remaining 35 plant classes represented less than 10% of the records and 13% of the species. This clearly indicated large data gaps across plant classes. Our findings are supported by Troia and McManamay [32], who reported that GBIF data coverage varies among taxonomic groups due to geographic, environmental, and institutional sampling biases.

Therefore, collecting data in deficient classes should be among the priorities for new collections. New data collections that integrate community science data with specimen-based documentation and the development of societal initiatives such as citizen science can help achieve that purpose [33–35]. Many medicinal plants, microalgae, and phytoplankton belong to underrepresented plant classes. For example, the class of Ginkgoopsida containing the species *Ginkgo biloba* L. has only 119 records; however, the medicinal properties of that species are recognized worldwide [36,37]. In the class Pyramimonadophyceae, phytoplankton are quite important in the marine food web [38], whereas microalgae belonging to other underrepresented classes (Ulvophyceae, Chlorophyceae, Trebouxiophyceae, etc.) play important roles in the food webs of streams [39]. Collecting more data records in these marginally represented classes will surely enable efficient data use and contribute to the conservation and sustainable use of their biodiversity. This will support the achievement of Sustainable Development Goal 3 (SDG 3), dedicated to ensuring healthy lives and promoting well-being for all at all ages. New data collection methods to fill data gaps across plant taxonomic groups should also address the representativeness of plant families. For example, the Meliaceae, Ebenaceae, and Moraceae families are rich in many valuable timber plant species; the Lamiaceae, Combretaceae, Sapotaceae, and Annonaceae families contain many medicinal plants and/or fruits of commercial interest [40–42]. They should also be among the priorities for new data collection across Africa. Efforts to fill these data gaps will also contribute to achieving SDG 15 related to the protection, restoration and promotion of sustainable use of terrestrial ecosystems.

Data published from Africa on the GBIF site are still scarce (less than 3% of the records, even nowadays). However, the total number of plant species we derived from the few records was 79,725. This number is much greater than the estimated number (40,000–60,000) of plant species from tropical Africa [14,15] and very close to the total number (80,000) of accepted names of plants from Africa as a whole [16]. The total number of plant species derived from the data published from Benin is 3898, as opposed to the 2807 species reported in the flora of Benin by [43]. GBIF-mediated data will, therefore, surely provide insights, in the near future, into the revision of the flora of tropical Africa and many African countries. This approach will also surely apply to countries on other continents and confirm that GBIF is a unique and leading mega infrastructure and data repository worldwide, a real unique reservoir and driver of research and knowledge on biodiversity. To achieve this, in line with Márquez et al. [44], we need more efforts directed towards data digitization, data management (data quality control), and data publication.

With respect to data loss (49%), because of a lack of quality, quality control of new records is of the utmost importance. This is a shared responsibility of data providers, their institutions, and the GBIF node managers. Indeed, in the implementation of the GBIF work plan, node managers are in charge of the capacity building of their country partners in data digitization, data management, including data cleaning, and data publishing. Therefore, they have a great shared responsibility for quality data publication on the GBIF site.

As records of preserved samples constitute approximately 60% of the data downloaded, there is room for data quality improvement in the process of data digitization from herbaria or natural historic plant collections because of the availability of vouchers. More capacity building of partners is needed across Africa, and more attention during data entry and

specimen imaging should be drawn to contribute to effective data quality improvement (Chapman [45]).

4.2. Data Completeness with Respect to the Basis of Records

The basis of a record refers to the type, nature, and method of data collection [46]. In the plant record data from Africa, two bases of records were dominant: preserved samples (59.68%) and human observations (30.86%). Globally, at the African level, efforts must be made to publish more preserved specimen data. Indeed, human observation data have some limitations in their validation in the absence of vouchers (Chapman [45]). The records of preserved specimens offer opportunities for data quality improvement in the process of data digitization because of the availability of vouchers. Human observations are not based on specimen vouchers. They are plant specimen observations, described and directly named in the field. With this method, bias can be introduced by misidentification and then in the name attributed to the observed specimen. Human observations, therefore, have limitations with respect to their reliability. However, human observations can also be reliable when the collectors of samples take vouchers of the specimens observed in the field and then hand them over to plant systematicians for their thorough determination based on preserved specimens in herbaria or the flora of the country. More capacity building of partners is needed across Africa, and more attention during data entry and specimen imaging should be drawn to contribute to effective data quality improvement. The data pattern of the basis of records, however, varies from country to country. For example, in Benin, human observation data represent 77.16% of the records, preserved specimen data represent 8.29%, and material samples represent 3.21% of the records published on the GBIF site; in contrast, in Kenya, preserved specimen data represent 59.04%, human observations represent 24.03%, and material sample data represent 5.60% of the records published on the GBIF site; in Madagascar, preserved specimens dominate the data published on the GBIF site, up to 92.26%; in Tanzania, preserved specimen data represent 75.63%, human observations represent 10.86%, and material sample records represent 2.09% of the records published on the GBIF site; and in South Africa, preserved specimens dominate the data published, up to 63.63%, followed by human observation data (34.12%). Therefore, we deduce that, from country to country, efforts will vary to increase the representativeness of preserved sample data as a first priority and then data related to other record bases.

4.3. Data Completeness Across Subregions of the Continent

Given the geographic patterns of data completeness, plant data gaps are quite large across the continent at both spatial resolutions. Data completeness is greater in West Africa (Benin, Ghana, and Southern Liberia). In East Africa, Tanzania and Kenya are more successful in terms of data completeness. In Central Africa, plant data completeness is more successful in Cameroon and Gabon. Data gaps are quite large in DRC, Sahelian, and North African countries. In southern Africa, the data gap is greater in South Africa and Mozambique. Madagascar is quite successful in achieving data completeness. Therefore, in the different subregions of Africa, efforts are still needed across countries to fill data gaps. Our results are in accordance with those of [18], who reported that data completeness was greater in the southern parts of Ghana. Plant data completeness was also analyzed in Benin [3]; the authors noted that data gaps were mostly found in the northernmost departments of Benin. The results of this study indicate that since then, appreciable efforts in data publication have been made to fill the data gaps identified so far in Benin.

Globally, Africa is quite late in data mobilization and publication on the GBIF site (less than 3% of the large amount of GBIF data, even nowadays). Efforts are, therefore, needed across the continent but need to be adapted to each situation across countries.

In the DRC, Sahelian, and North African countries, where data gaps are actually quite large, the countries concerned are not affiliated with the GBIF and, therefore, do not have enough incentive to publish data on the GBIF site; thus, the biodiversity of these countries is not well known and is, therefore, quite subject to hazards. Here, my suggestion is to reach out to the governments of those countries in order to raise their awareness of the possible risks their natural resources can face if efforts are not made toward biodiversity data discovery in their countries, and, therefore, the need for data publication on international repositories such as GBIF to enable data use by researchers to inform decisions on biodiversity conservation and sustainable use.

For other countries where data gaps are more limited, such as some countries in West Africa, Central Africa, East Africa, Southern Africa and Madagascar, data gap analyses across taxonomic groups are needed to identify taxonomic data gaps and, therefore, prioritize data collection to fill the gaps identified. This approach will be quite useful for balanced data use across taxonomic groups to inform more appropriate decision-making on biodiversity conservation.

4.4. Impact of Accessibility and Protected Areas on Data Completeness

Sampling bias is commonly addressed in plant and animal inventories [47,48]. According to our results, accessibility by roads significantly affects data completeness, and the coarser the resolution of spatial grid cells is, the stronger the correlation between road length and the number of records. Our results are supported by those of [34,49], who reported positive effects of traffic access, road density, and tourism on the inventory completeness of tropical insects in Sub-Saharan Africa. Our results are also supported by a previous study [3] on plant data completeness in Benin, where the authors reported that data completeness was positively impacted by the density of roads. Our results are also in line with those of [1,48] in their respective studies on plant data completeness in Brazil and species data in Israel, where they reported that road bias affected the distribution of data. Our results are also supported by those of [47] in their study on the spatial quality of biological databases, where they reported that the spatial pattern of the database in West Africa was strongly biased with respect to distance to cities, coasts, rivers and protected areas. Data completeness is also affected by the density of rivers. This is in line with the results of Araujo et al. [50], who determined the contributions of rivers to biodiversity data completeness in Amazonia. Our results showed that protected areas (PAs) also affect data completeness. Our results are also supported by Girardello et al. [35], who reported gaps in inventories of butterflies in areas with a low density of PAs; however, they provided no indication of the size of the PAs. Another dimension of bias in data completeness worthy of consideration is underlined by Escribano et al. [10] in their study on gap analysis of digitally accessible information on terrestrial vertebrates; they found that data completeness depends on distance to researchers, local research funding, and data publication rather than on transportation infrastructure. The lesson to learn from possible bias in data completeness is that new data collection priorities must address remote areas from roads (beyond 500–2000 m) (Kadmon et al. [48]), as well as remote areas of the largest protected areas (>2500 km^2). The opening of new roads, which are more representative of the ecological conditions of landscapes across African countries, can significantly contribute to less biased data collection. Addressing less charismatic species and developing societal initiatives such as citizen science can also contribute to less biased data collection [33]. Greater incentives for data publication and more intense connections between researchers can also contribute to filling gaps in information [10].

5. Conclusions

Africa is far behind in terms of data publications, accounting for less than 3% of the more than three billion occurrence data points available on the GBIF site even nowadays. South Africa alone has contributed more than 50% of data from Africa, and this is alarming with respect to the needed awareness and efforts from the rest of the continent toward indispensable investments in data mobilization and data cleaning to enable data use to inform decisions on biodiversity conservation and sustainable use. Data gaps are quite steep across taxonomic groups and geographic space. The results achieved here clearly set priorities for new data collection to fill the gaps identified. The priority areas of new data collections are the orange grid cells clearly indicated in Figures 6 and 8–10 and their related Figures S1–S4. Those in charge of such new data collections should take into account the limiting factors we underlined as impacting data completeness and design appropriate plans for new data collections. Data loss during the data cleaning process is quite disappointing, in that only approximately 50% of the original data downloaded from the GBIF site are adequate for data completeness analysis. This underlines the responsibilities of data providers, their institutions and GBIF node managers for collaborative data quality checks and improvements before data publication. Indeed, in the implementation of the GBIF work plan, node managers are in charge of capacity building for their country partners in data digitization and data management, including data cleaning, before data publication on the GBIF site. Node managers must check the quality of data before publication and then assist data providers in improving the quality of their data when needed before publication. Therefore, they are highly responsible for quality data publication on the GBIF site. In line with our results, taking into account the priorities identified for new data collections, along with more efforts in publishing more cleaned data on the GBIF site, is a prerequisite to meet the challenges of biodiversity conservation across Africa in the context of climate and global changes.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/su182111054/s1, Figure S1: Geographic patterns of inventory completeness across Africa at a 0.5° spatial resolution (the map was generated using QGIS 3.18.1 (https://api.qgis.org/api/3.18/index.html (accessed on 17 January 2023)) and WGS 84 as Coordinate Reference System); Figure S2: Geographic patterns of inventory completeness across Africa at a 0.5° spatial resolution versus the road layer (the map was generated using QGIS 3.18.1 (https://api.qgis.org/api/3.18/index.html (accessed on 17 January 2023)) and WGS 84 as Coordinate Reference System); Figure S3: Geographic patterns of inventory completeness across Africa at 0.5° spatial resolution versus river layer (the map was generated using QGIS 3.18.1 (https://api.qgis.org/api/3.18/index.html (accessed on 17 January 2023)) and WGS 84 as Coordinate Reference System); Figure S4: Geographic patterns of inventory completeness across Africa at 0.5° spatial resolution versus the layer of protected areas (the map was generated using QGIS 3.18.1 (https://api.qgis.org/api/3.18/index.html (accessed on 17 January 2023)) and WGS 84 as Coordinate Reference System).

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw data used during the current study are available at the following link: https://doi.org/10.15468/dl.p2n6um (accessed on 17 January 2023).

Acknowledgments: The author expressed his sincere gratitude to Peterson Townsend of the University of Kansas, USA, his former trainer in data completeness analysis. The author also tells his profound gratitude to the Secretariat of the Global Biodiversity Information Facility (GBIF, www.gbif.org) for offering him the opportunity for additional training on the DataCamp platform (https://app.datacamp.com/learn (accessed on 17 January 2023)). He also thanks Daniel Noesgaard from the GBIF secretariat for his useful comments on a previous version of the manuscript.

Conflicts of Interest: The author declares no conflicts of interest.

Abbreviations

DAK	Digitally Accessible Knowledge
GBIF	Global Biodiversity Information Facility
QGIS	Quantum Geographic Information System
ggplot	Quantum Geographic Information System
DRC	Democratic Republic of the Congo
WGS	world geodetic system

References

1. Sousa-Baena, M.S.; Garcia, L.C.; Peterson, A.T. Completeness of digital accessible knowledge of the plants of Brazil and priorities for survey and inventory. *Divers. Distrib.* **2014**, *20*, 369–381. [CrossRef]
2. Elith, J.; Phillips, S.J.; Hastie, T.; Dudík, M.; Chee, Y.E.; Yates, C.J. A statistical explanation of Maxent for ecologists. *Divers. Distrib.* **2011**, *17*, 43–57. [CrossRef]
3. Ganglo, J.C.; Kakpo, S.B. Completeness of digital accessible knowledge of plants of Benin and priorities for future inventory and data discovery. *Biodivers. Inform.* **2016**, *11*, 23–39. [CrossRef]
4. Ganglo, J.C. Ecological niche model transferability of the white star apple (*Chrysophyllum albidum* G. Don) in the context of climate and global changes. *Sci. Rep.* **2023**, *13*, 2430. [CrossRef]
5. Ganglo, J.C.; Djotan, G.K.; Gbètoho, J.A.; Kakpo, S.B.; Aoudji, A.K.; Koura, K.; Tessi, R.Y. Ecological niche modeling and strategies for the conservation of *Dialium guineense* Willd. (Black velvet) in West Africa. *Int. J. Biodivers. Conserv.* **2017**, *9*, 373–388. [CrossRef]
6. Djotan, A.K.; Aoudji, A.K.N.; Codjia, S.A.F.; Gbètoho, A.J.; Koura, K.; Ganglo, J.C. How far can climate changes help to conserve and restore Garcinia kola Heckel, an extinct species in the wild in Benin (West Africa). *Int. J. Biodivers. Conserv.* **2018**, *10*, 203–213. [CrossRef]
7. Kakpo, S.B.; Aoudji, A.K.N.; Gnanguènon-Guéssè, D.; Gbètoho, A.J.; Koura, K.; Djotan, G.K.; Ganglo, J.C. Spatial distribution and impacts of climate change on *Milicia excelsa* in Benin, West Africa. *J. For. Res.* **2021**, *32*, 143–150. [CrossRef]
8. Mora, C.; Tittensor, D.P.; Myers, R.A. The completeness of taxonomic inventories for describing the global diversity and distribution of marine fishes. *Proc. R. Soc. B Biol. Sci.* **2008**, *275*, 149–155. [CrossRef]
9. Huttunen, K.L.; Mykrä, H.; Muotka, T. Temporal variability in taxonomic completeness of stream macroinvertebrate assemblages. *Freshw. Sci.* **2012**, *31*, 423–441. [CrossRef]
10. Escribano, N.; Galicia, D.; Ariño, A.H. Completeness of Digital Accessible Knowledge (DAK) about terrestrial mammals in the Iberian Peninsula. *PLoS ONE* **2019**, *14*, e0213542. [CrossRef]
11. Huang, X.; Lin, C.; Ji, L. The persistent multidimensional biases of biodiversity digital accessible knowledge of birds in China. *Biodivers. Conserv.* **2020**, *29*, 3287–3311. [CrossRef]
12. Meyer, C.; Kreft, H.; Guralnick, R.; Jetz, W. Global priorities for an effective information basis of biodiversity distributions. *Nat. Commun.* **2015**, *6*, 8221. [CrossRef]
13. Neves, I.Q.; Mathias, M.D.L.; Bastos-Silveira, C. Mapping Knowledge Gaps of Mozambique's Terrestrial Mammals. *Sci. Rep.* **2019**, *9*, 18184. [CrossRef]
14. Küper, W.; Sommer, J.H.; Lovett, J.C.; Mutke, J.; Linder, H.P.; Beentje, H.J.; Rompaey, S.A.R.; Chatelain, C.; Sosef, M.; Barthlott, W. Africa's hotspots of biodiversity redefined. *Ann. Mo. Bot. Gard.* **2004**, 525–535.
15. Lebrun, J.-P.; Stork, A. *Enumération Des Plantes À Fleurs d'Afrique Tropicale 4Vols. Ville de Genève*; Editions des Conservatoire et Jardin Botaniques: Geneva, Switzerland; pp. 1991–1997.
16. African Plant Database. 2024. Available online: http://africanplantdatabase.ch (accessed on 4 June 2024).
17. Idohou, R.; Arino, A.; Assogbadjo, A.; Kakai, R.G.; Sinsin, B. Diversity of wild palms (Arecaceae) in the Republic of Benin: Finding the gaps in the national inventory combining field and digital accessible knowledge. *Biodivers. Inform.* **2015**, *10*, 45–55. [CrossRef]

18. Asase, A.; Peterson, A.T. Completeness of digitally accessible knowledge of Ghana's plants. *Biodivers. Inform.* **2016**, *11*, 1–11. [CrossRef]
19. Freeman, B.; Peterson, A.T. Completeness of digital accessible knowledge of the birds of western Africa: Priorities for survey. *The Condor* **2019**, *121*, duz035. [CrossRef]
20. Myers, N.; Mittermeier, R.A.; Mittermeier, C.G.; Da Fonseca, G.A.; Kent, J. Biodiversity hotspots for conservation priorities. *Nature* **2000**, *403*, 853–858. [CrossRef]
21. La Ferla, B.; Taplin, J.; Ockwell, D.; Lovett, J.C. Continental scale patterns of biodiversity: Can higher taxa accurately predict African plant distributions? *Bot. J. Linn. Soc.* **2002**, *138*, 225–235. [CrossRef]
22. GBIF. Occurrence Download. 2023. Available online: https://www.gbif.org (accessed on 17 January 2023).
23. R Core Team. *R. A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2020. Available online: https://www.R-project.org/ (accessed on 1 January 2025).
24. QGIS Development Team QGIS Geographic Information System. Open Source Geospatial Foundation Project. 2021. Available online: https://api.qgis.org/api/3.18/index.html (accessed on 17 June 2021).
25. Chao, A. Estimating the population size for capture recapture data with unequal catchability. *Stat. Sin.* **1987**, *10*, 227–246.
26. World Agroforestry Centre. 2022. Available online: https://data.humdata.org/dataset/africa-hydrorivers-vectorized-line-network-of-all-african-rivers (accessed on 26 August 2022).
27. Africa—Road Network. 2022. Available online: https://energydata.info/dataset/road-network (accessed on 26 August 2022).
28. Protected Planet. 2022. Available online: https://www.protectedplanet.net/region/AF (accessed on 27 August 2022).
29. Kitchenham, B. Robust statistical methods: Why, what and how: Keynote. In Proceedings of the 19th International Conference on Evaluation and Assessment in Software Engineering, Nanjing China, 27–29 April 2015; pp. 1–6.
30. Mair, P.; Wilcox, R.; Patil, I. A Collection of Robust Statistical Methods. 2026. Available online: https://mirrors.nics.utk.edu/cran/web/packages/WRS2/WRS2.pdf (accessed on 26 August 2022).
31. Mesaglio, T.; Callaghan, C.T.; Samonte, F.; Gorta, S.B.; Cornwell, W.K. Recognition and completeness: Two key metrics for judging the utility of citizen science data. *Front. Ecol. Environ.* **2023**, *21*, 167–174. [CrossRef]
32. Troia, M.J.; McManamay, R.A. Filling in the GAPS: Evaluating completeness and coverage of open-access biodiversity databases in the United States. *Ecol. Evol.* **2016**, *6*, 4654–4669. [CrossRef]
33. Chesshire, P.R.; Fischer, E.E.; Dowdy, N.J.; Griswold, T.L.; Hughes, A.C.; Orr, M.C.; Ascher, J.S.; Guzman, L.M.; Hung, K.-L.J.; Cobb, N.S.; et al. Completeness analysis for over 3000 United States bee species identifies persistent data gap. *Ecography* **2023**, *2023*, e06584. [CrossRef]
34. Troudet, J.; Grandcolas, P.; Blin, A.; Vignes-Lebbe, R.; Legendre, F. Taxonomic bias in biodiversity data and societal preferences. *Sci. Rep.* **2017**, *7*, 9132. [CrossRef]
35. Girardello, M.; Chapman, A.; Dennis, R.; Kaila, L.; Borges, P.A.; Santangeli, A. Gaps in butterfly inventory data: A global analysis. *Biol. Conserv.* **2019**, *236*, 289–295. [CrossRef]
36. Singh, B.; Kaur, P.; Singh, R.D.; Ahuja, P.S. Biology and chemistry of Ginkgo biloba. *Fitoterapia* **2008**, *79*, 401–418. [CrossRef]
37. Belwal, T.; Giri, L.; Bahukhandi, A.; Tariq, M.; Kewlani, P.; Bhatt, I.D.; Rawal, R.S. Ginkgo biloba. In *Nonvitamin and Nonmineral Nutritional Supplements*; Academic Press: Cambridge, MA, USA, 2019; pp. 241–250.
38. Jónasdóttir, S.H. Fatty acid profiles and production in marine phytoplankton. *Mar. Drugs* **2019**, *17*, 151. [CrossRef]
39. Hodač, L.; Brinkmann, N.; Mohr, K.I.; Arp, G.; Hallmann, C.; Ramm, J.; Spitzer, K.; Friedl, T. Diversity of microscopic green algae (Chlorophyta) in calcifying biofilms of two karstic streams in Germany. *Geomicrobiol. J.* **2015**, *32*, 275–290. [CrossRef]
40. Britannica, T. Editors of Encyclopaedia "Moraceae". Argon. In Encyclopedia Britannica. 2022. Available online: https://www.britannica.com/plant/Moraceae (accessed on 3 October 2022).
41. Houessou, L.G.; Lougbegnon, T.O.; Gbesso, F.G.; Anagonou, L.E.; Sinsin, B. Ethno-botanical study of the African star apple (*Chrysophyllum albidum* G. Don) in the Southern Benin (West Africa). *J. Ethnobiol. Ethnomed.* **2012**, *8*, 40. [CrossRef]
42. Luteyn, J.L.; Stevens, P. "Ericales". Encyclopedia Britannica. 2022. Available online: https://www.britannica.com/plant/Ericales (accessed on 3 October 2022).
43. Akoègninou, A.; Van der Burg, W.J.; Van der Maesen, L.J.G. *Flore Analytique Du Bénin (No. 06.2)*; Backhuys Publishers: Kerkwerve, The Netherlands, 2006; p. 1034.
44. Márquez, J.R.G.; Dormann, C.F.; Sommer, J.H.; Schmidt, M.; Thiombiano, A.; Da, S.S.; Chatelain, C.; Dressler, S.; Barthlott, W. A methodological framework to quantify the spatial quality of biological databases. *Biodivers. Ecol.* **2012**, *4*, 25–39. [CrossRef]
45. Chapman, A.D. *Principles and Methods of Data Cleaning—Primary Species and Species-Occurrence Data*, version 1.0. Report for the Global Biodiversity Information Facility; Global Biodiversity Information Facility: Copenhagen, Denmark, 2005.
46. GBIF. 2022. Available online: https://www.gbif.org/data-quality-requirements-occurrences#dcBasis (accessed on 3 October 2022).
47. Ball-Damerow, J.E.; Brenskelle, L.; Barve, N.; Soltis, P.S.; Sierwald, P.; Bieler, R.; LaFrance, R.; Ariño, A.; Guralnick, H. Research applications of primary biodiversity databases in the digital age. *PLoS ONE* **2019**, *14*, e0215794. [CrossRef]

48. Kadmon, R.; Farber, O.; Danin, A. Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. *Ecol. Appl.* **2004**, *14*, 401–413. [CrossRef]
49. Ballesteros-Mejia, L.; Kitching, I.J.; Jetz, W.; Nagel, P.; Beck, J. Mapping the biodiversity of tropical insects: Species richness and inventory completeness of African sphingid moths. *Glob. Ecol. Biogeogr.* **2013**, *22*, 586–595. [CrossRef]
50. de Araujo, M.L.; Quaresma, A.C.; Ramos, F.N. GBIF information is not enough: National database improves the inventory completeness of Amazonian epiphytes. *Biodivers. Conserv.* **2022**, *31*, 2797–2815. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.