SANBI-GBIF & GBIF-SPAIN TRAINING WORKSHOP 2023 REPORT

Analytical Techniques in Biodiversity Big Data using GBIF: Making an Impact

Venue: Iziko Museums of South Africa, Cape Town, South Africa

26 – 30 June 2023

Contents

Executive summary

SANBI-GBIF collaborated with GBIF-Spain and embarked on a five-day training workshop on big data in South Africa. The workshop focused on topics such as artificial intelligence (AI) and 'big data'. The use of the R programming language in Jupyter Notebook, which is a web based interactive computing platform, was also explored. Other aspects also included advances in machine learning and deep learning leading to the ability to decipher the content of natural images. This can provide new insight for researchers and make difficult analyses of natural images a routine task.

The coordinating team from SANBI-GBIF and GBIF-Spain, composed of the Node Managers and staff. Additional experts from various institutions (University of of Cape Town [SEEC], University of Cantabria [Spain], University of Free State, Sol Plaatje University) were also invited to create an exciting programme of work, to enable the development of the training course, aimed at big data approaches to effectively analyse, mine and interrogate the rich GBIF mediated biodiversity resource, which is at the disposal of the scientific community. The workshop was attended by a further 22 participants from the SANBI-GBIF node community, and included regional participation from Malawi.

This report describes the preparation of the workshop, provides links to websites, material, and social media content, and a summary of the workshop evaluation by participants.

Workshop Details

SANBI-GBIF and GBIF-SPAIN partnered on a project entitled “**Cross-continental partnership to investigate data mining approaches for impactful data use cases and stories**”, funded through the Capacity Enhancement Support Programme (CESP) of the Global Biodiversity Information Facility (GBIF). This joint venture included training and information exchange on the GBIF-Spain eLearning platform from 12th-16th December 2022 (Spain), and from 6th-10th March 2023 (South Africa). The overall aim was to optimise efforts and grow national capacity through the use and functionality of the eLearning Platform. Additionally, SANBI-GBIF and GBIF-Spain aimed to explore ways to do analyses to enable data mining and the identification of key tools, techniques and approaches that can answer pertinent research questions related to space, time, and taxonomy.

During the visit to Spain (December 2022), the course curriculum was drafted and the following topics were decided on: Artificial intelligence, deep learning, and data cleaning techniques were discussed; particularly enhancing the quality of data with further discussion about aspects like the quality index for data publishing. Data reduction techniques applied to digital accessible knowledge (DAK) was also addressed, especially as this relates to aspects like time, geographic space, and the taxonomy of biodiversity data. Further additional aspects were also considered and included mobilisation of data through eDNA, which has been identified as the **new frontier for increasing biodiversity data mobilised by orders of magnitude**. This is similar to the shift that citizen science has brought about to the data mobilisation arena in the past decades.

The agenda was developed with the aim of looking at a series of tools and analysis that was very relevant to current day researchers, and **spanned the data value chain from data access to data use**. The Partners (SANBI-GBIF and GBIF Spain) also identified that opportunities to showcase exciting use cases and stories with the data, would be important for participants to **understand the value of GBIF mediated data**. This human capital development opportunity presented by this big data world, enables stakeholders themselves to play a part down the line, in the conservation of our precious biodiversity.

The GBIF Spain team Dr. Franscisco Pando and Dr. Katia Cezon introduced participants to the GBIF API, and downloading data from GBIF as well as looking at different sources of biodiversity data, and some new data visualisation applications. Data Science expertise from the Spanish National Research Council included Dr. Fernando Aguilar, who demonstrated image segmentation techniques to satellite imagery. These machine learning techniques can be applied to map layers or remote sensing images, which may be relevant for biodiversity management or species distribution modelling.

The last few modules included basic steps and tools to create Species Distribution Models (SDM) looking at the **potential distribution of invasive species into new territories**, and changes in species distribution under different climate scenarios.

The course also provided the opportunity to delve into mechanisms of National Biodiversity Assessments, providing very clear indications of the data-driven nature of conservation assessments, and how occurrence data, such as that mobilised through our stakeholders, can make a difference to conservation planning work on the ground. Here, experts from SANBI presented the National Biodiversity Assessment (NBA) efforts taking place in South Africa, as well as delving into both the species and ecosystem workflows to generate the outputs in the NBA. Following on from the session on the assessments, systematic conservation planning was also discussed and Marxan was introduced as a conservation software tool. This demonstrated a practical example of how biodiversity data is used for on the ground planning, informing conversation plans of reserve managers.

High Level Workshop programme (detailed agenda below in Appendix 2)

PROGRAMME

ANALYTICAL TECHNIQUES IN BIODIVERSITY BIG DATA USING GBIF: MAKING AN IMPACT

Venue: The Iziko South African Museums, Biodiversity Lab, Cape Town

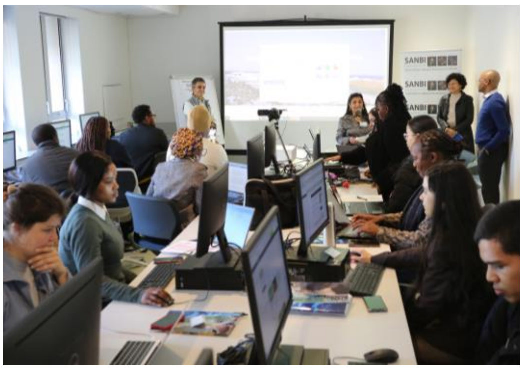
Date: 26th-30th June 2023

Time	Day 1 - 26th June	Day 2 - 27th June	Day 3 - 28th June	Day 4 - 29th June	Day 5 - 30th June
	Session 1	**Session 5**	**Session 8**	**Session 10**	**Session 13**
9:00- 9:30	Registration opens	Jupyter Notebooks (Vernon Visser, SEEC-UCT)	Molecular data mobilization - eDNA (Morne Du Plessis, UFS)	Distribution models using Maxent and Machine Learning Techniques (Francisco Pando, GBIF-Spain)	Conversation on search engines-ChatGPT (Francisco Pando, GBIF-Spain)
9:30-10:00	Welcome and Opening Dr. Bongani Ndhlovu (Iziko Museum of SA) & Ms. Carmel Mbizvo (SANBI)				
10:00-10:30	Introduction (Fatima Parker-Allie, SANBI-GBIF)				
10:30 -11:00	Tea break				
	Session 2	**Session 6**	**Session 8**	**Session 10**	**Session 14**
11:00-11:30	Concepts (Francisco Pando, GBIF-Spain)	Data Visualisation (Katia Cezon, GBIF-Spain)	Molecular data mobilization - eDNA (Morne Du Plessis, University of Free State)	Distribution models using Maxent and Machine Learning Techniques (Francisco Pando, GBIF-Spain)	Course Evaluation Wrap Up and Closing (Fatima Parker-Allie, SANBI-GBIF)
11:30-12:00					
12:00-12:30	software installation				
12:30-14:00	Lunch time				
	Session 2, 3	**Session 7**	**Session 9**	**Session 11**	
14:00-14:30	software installation	Finding Outliers / Data Reduction Techniques – Time, Space, Name (Francisco Pando, GBIF-Spain)	Image Segmentation – Machine Learning using EO browser/ Sentinel playground (Fernando Aguilar, CSIC)	National Biodiversity Assessment: Species & Ecosystem Assessment (Andrew Skowno, Dewidine van der Colff & Maphale Monyeki, SANBI)	
14:30-15:00					
15:00-15:30	Data Sources (Katia Cezon, GBIF-Spain)				
15:30-15:45	Tea break				
	Session 4	**Session 7**	**Session 9**	**Session 12**	

15:45-16:30	GBIF API (Francisco Pando, GBIF-Spain)	Finding Outliers / Data Reduction Techniques – Time, Space, Name - (Francisco Pando, GBIF-Spain)	Image Segmentation – Machine Learning using EO browser/ Sentinel playground (Fernando Aguilar, CSIC)	Conservation planning - (Douglas Harebottle, Sol Plaatje University)	
16:30-17:00	Wrap Up	Wrapping Up	Wrap Up	Wrap Up	

Workshop highlights

A talented group of stakeholders and experts came together and took part in an eventful week of biodiversity informatics training. The much anticipated training workshop was hosted at the beautiful Iziko Museums in Cape Town, South Africa. Here, the Executive Director of Iziko Museums welcomed Participants to the workshop and Ms. Carmel Mbizvo, the Deputy Director General conducted the official opening to the workshop. A contextualisation of the SANBI-GBIF Node, the Capacity Enhancement Support Programme of GBIF, and the Biodiversity Informatics Landscape was conducted by the SANBI-GBIF Node Manager. The hosting of the workshop at Iziko Museums, also provided the perfect opportunity to explore the Natural History Collections of this flagship institution, which contains some 700 million year old fossils, socio-cultural artefacts, and some very expansive insect and marine invertebrate collections of the country.



Participants **were exposed to a variety of scientific topics**, **and to many new and relevant tools and applications**, including Jupyter Notebooks, Data Visualization using Keppler.gl, Molecular Data Mobilization (eDNA), Image Segmentation – Machine Learning using EO browser and Sentinel playground, Distribution models using Maxent and Machine Learning Techniques, and Systematic Conservation Planning. The final day looked at the power of Artificial Intelligence (AI) through the application of ChatGPT, and left us pondering: “How to harness the potential of AI and the 4th Industrial Revolution for Biodiversity Conservation and the protection of our species and ecosystems".

Many sessions included exercises or showcases, which allowed participants to gain a greater understanding on the topic, as well as allowed for more interaction between the experts/trainers and the participants. Sessions were designed to cover a range of topics, to provide an exposure to a broad scope of subject areas in the area of big data analytics and the opportunities presented by GBIF data. Many of these topics have the opportunity to be explored at more depth, with more time allocation to each.

Workshop materials, e-Learning platform and social media

For the workshop we made use of GBIF-Spain's e-Learning platform for easy access to presentations and training materials; which can also be reused. Access to the course is through the creation of a user account on the e-Learning platform, which was setup for Participants. In future, this course is accessible to the public, through a very simple subscription process. Access to the course entitled "**Analytical Techniques in Biodiversity Big Data using GBIF: Making an Impact**" is from the course catalogue. We will continue to update the materials available on the e-Learning platform with videos as they become available.

The course can also be accessed via the SANBI-GBIF elearning page, https://www.sanbi-gbif.org/e-learning, and subscription at the following link https://elearning.gbif.es/course/138/about

Daily updates and photos of the workshop were tweeted on our SANBI-GBIF Twitter page, the GBIF-Spain Twitter page and individual Node Managers twitter page.

All participants were provided with a certificate of achievement from SANBI-GBIF and GBIF-Spain).

Workshop evaluation

Workshop participants were invited to submit evaluations via Google Forms, the link of which was found on the eLearning platform. Twenty-one responses were submitted, representing 95% of the workshop participants. Participants were asked to rate their level of satisfaction with the workshop, based on a number of aspects (Figure 1) on a scale of 1 (needs work) to 10 (excellent). Overall participant responses were very positive, with a big appreciation for the subject knowledge of the experts/instructors. Participants also expressed an appreciation for the subject areas selected with a sense that it was very progressive in terms of the contents offered and what was learnt and gained from the training. Despite the fact that this big data field might have a steep learning curve for many, participants managed the complexity of the course and indicated an average difficulty rating of 4.90, on a scale of 1(very easy) to 10 (extremely difficult). The course also provided excellent opportunities to interact and network (8.29) and an overall rating of 8.24 out of 10.

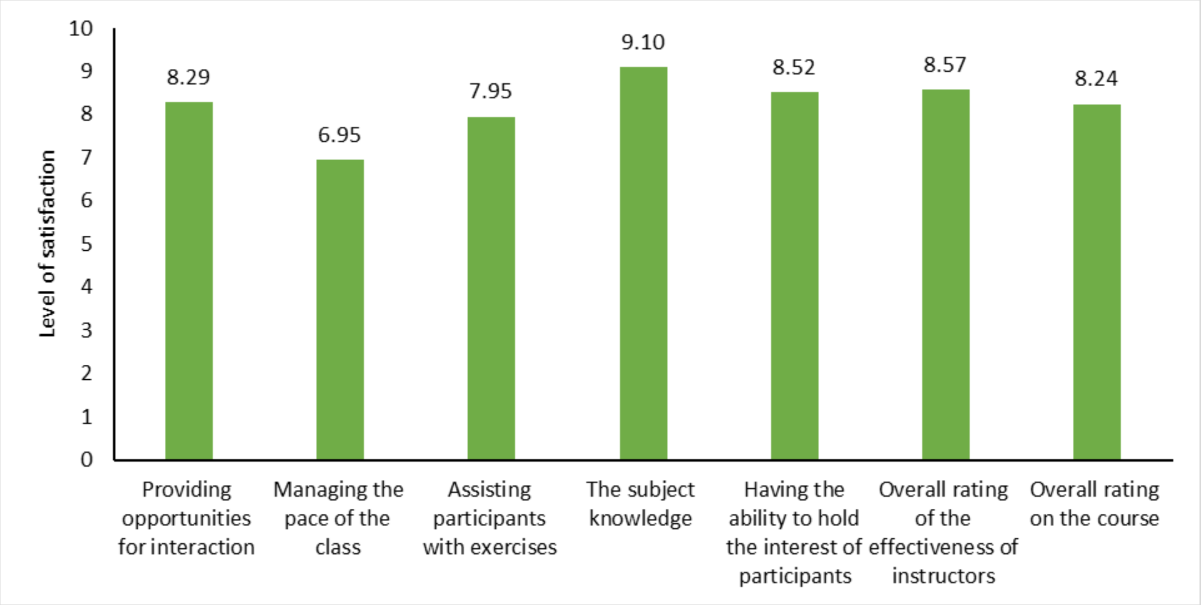


Figure 1: Average effectiveness of instructors and overall rating of the course (n=22)

Comments

1. **Instructors**

- *Their presentations were so informative. They made sure that we understood everything before moving to the next task.*
- *Good subject knowledge and preparation but needed more time for "set up" and going around checking everyone had the correct software etc.*
- *Large range of skills so would help to have additional assistants*

2. **Content**

- *Exactly what I needed since I am working on big data.*
- *The content was very forward thinking and gave new ideas on how to effectively use GBIF as a science resource.*
- *Covered many aspects but would have liked to get more detail in many of them so perhaps there should be longer dedicated sessions in the future on some of these.*
- *Interesting, wide variety of topics made the workshop interesting and engaging. If it were the same topics one after the other there may have been a lack of interest, but the timing, and spread of topics made it easy to maintain focus, as each session was refreshing.*
- *Too much content and so little time, we rushed through certain important topics but the content itself was very relevant for me as a Data technician.*

3. **Exercises**

- *Fun and interactive, easy to follow. I particularly enjoyed the image segmentation coding.*
- *There are sessions that could be expanded.*
- *The exercises were well designed and used practical examples. Instructors moved around the class to assist participants when necessary.*

4. **eLearning platform**

- *The platform was very user-friendly; perhaps applying a better interface between the presenter and the audience to share links more efficiently.*
- *I enjoyed it, it was straightforward and logical!*
- *The platform worked well for easily distributing information on the course outline, presentation slides, and links to exercises or further information*
- *Yes it was user friendly. However, the sessions would always timeout and require you to sign in again.*

5. **What Participants enjoyed most about the course? (Some responses consolidated)**

- *Enjoyed everything. Every presenter delivered a very informative training and information.*
- *Introduction to new tools and techniques relevant to my work. I will continue to teach myself these tools and techniques*
- *Paco's sense of humor and the variety of exposure sessions*
- *The practical exercises inspired me to actually start a 'programming for dummies' course :)*
- *Interactions and learning from others, exposure to some topics that I haven't worked with. Interacting with the diverse array of participants & exposure to topics not previously used. The diverse range of talks and exercises*
- *course materials easy to access, well structured and organized*
- *Knowing more about our SA platforms and initiatives; I was not aware of a lot of it.*
- *Running scripts to calculate land cover and accuracy assessment using satellite imagery and how to use ChatGPT.*
- *I enjoyed that most of the topics linked to each other in terms of programs they can be done in. Most of the techniques and content is able to be done through jupyter notebook or within R. I think its always useful to have one tool that can do multiple things rather than several tools. And the instructors were all informed on their topics as well as the other topics presented and thus could make these connections or inform us of the possible ways we can use everything in tandem.*
- *Data visualization. However, I would have preferred to also learn about other data visualization tools, not just maps. Also learning about EMEditor and learning new tricks in MS Access.*
- *The amount of content that was presented but to be specific to QGIS, GBIF, DATA SOURCES and data visualisation.*
- *The sessions were well structured and developed in a way that they are not content heavy and rather more practical.*
- *I definitely took something away from all the sessions. The section on using AI tools in research was especially informative & fun and I also enjoyed the data visualisation session*
- *Using maxent and QGIS to link species data from GBIF platform and Sentinel 2 images from EO browser*
- *Machine learning techniques and data visualisation techniques*

Appendix 1. Detailed agenda

Programme	
Day 1, 26th June 2023 **The Iziko South African Museums, Biodiversity Lab**	
9:00- 9:30	**Registration opens** Please proceed to the registration table to pick up your name badges
9:30-10:00	**Welcome and opening** Welcome by Institutional Host, Acting CEO of Iziko Museums of South Africa Dr. Bongani Ndhlovu Opening Remarks by SANBI's Head of Branch: Biodiversity Science and Policy Advice and Deputy Director General Ms. Carmel Mbizvo
10:00-10:30	**Introduction to SANBI-GBIF** Fatima Parker-Allie - SANBI-GBIF An outlining the SANBI-GBIF Initiative and efforts in supporting Biodiversity Informatics in South Africa will be presented
10:30-11:00	Tea and coffee Break
11:00-12:00	**Concepts** Francisco Pando - GBIF-Spain This session will focus on outlining data concepts, providing theoretical background on a range of topics, which will be explored during the week, including coding, algorithms, big data, Artificial Neural Networks, etc.
12:00-12:30	Software installations Time allocated to ensure all software is in working order
12:30-14:00	Lunch break
14:00-15:00	Software installations Time allocated to ensure all software is in working order
15:00-15:30	**Data Sources** Katia Cezon - GBIF-Spain In this presentation, we will explore a diverse range of data sources that are crucial for conducting analyses in the field of biodiversity. Our focus will be on identifying various biological and environmental data sources to gain valuable insights and make a significant impact in biodiversity research.
15:30-15:45	Tea and coffee Break
15:45-16:30	**GBIF API** Francisco Pando GBIF-Spain This session will cover the basics of the GBIF API, how to download and retrieve data from the GBIF website and will include a hands-on exercise. The exercise will include building API requests and see if they work and how they work
16:30-17:00	Wrap Up

Day 2, 27th June 2023 The Iziko South African Museums, Biodiversity Lab	
9:00- 10:30	**Jupyter Notebooks** Vernon Visser SEEC-UCT This session will include a presentation with an introduction to Jupyter Notebooks, including a practical hands-on example. You will learn about - why we use them, how to use them for Python and R, how to do some basic Markdown formatting, etc. We will also run through a basic introduction to Python notebook.
10:30-11:00	Tea and coffee Break
11:00-12:30	**Data Visualisation** Katia Cezon GBIF-Spain Here, we will explore common tools used to visualize primary biodiversity data and environmental data. Our main focus will be on keppler.gl, an open-source web tool developed by Mapbox and Uber. This powerful tool is specifically designed to visualize & analyze spatial data on interactive maps. This session will include an exercise.
12:30-14:00	Lunch break
14:00-15:30	**Finding Outliers / Data Reduction Techniques – Time, Space, Name** Francisco Pando GBIF-Spain We will conduct this session on a particular case but keeping in mind that how to approach this problem is more about thinking what we need and how to get it than about mastering any particular technique or software.
15:30-15:45	Tea and coffee Break
15:45-16:30	**Finding Outliers / Data Reduction Techniques – Time, Space, Name** Francisco Pando GBIF-Spain Continued from above
16:30-17:00	Wrap Up
Day 3, 28th June 2023 The Iziko South African Museums, Biodiversity Lab	
9:00-10:30	**Molecular data mobilization – eDNA** Morne Du Plessis UFS A presentation which will explore the background, applications, and technologies in eDNA based research. Incorporated in the presentation will be group interaction around matching eDNA strategies to sequencing approaches. This is followed by a data quality interpretation exercise.
10:30-11:00	Tea and coffee Break
11:00-12:30	**Molecular data mobilization- eDNA** Morne Du Plessis UFS A presentation focused on eDNA analyses workflows followed by the approaches used to organise and evaluate the data in order to publish it. Incorporated in this section will be group discussions on limitations of eDNA research, exploring the Darwin Core Categories and then a data exploration exercise on GBIF examples of the various categories for eDNA publication.
12:30-14:00	Lunch break
14:00-15:30	**Image Segmentation – Machine Learning using EO browser/ Sentinel playground** Fernando Aguilar GBIF-Spain Image segmentation is a computer vision technique to divide a picture in different regions depending on its visual characteristics. It can be applied to map layers or remote sensing images, which could be useful for biodiversity management or species distribution. In this session, we will find out how supervised and

	unsupervised learning algorithms can be used for this purpose. Using Python and different libraries we will divide a Sentinel-2 satellite data in different regions like water, forest or urban surfaces.
15:30-15:45	Tea and coffee Break
15:45-16:30	**Image Segmentation – Machine Learning using EO browser/ Sentinel playground** Fernando Aguilar GBIF-Spain Continued from above
16:30-17:00	Wrap Up

Day 4, 29th June 2023 **The Iziko South African Museums, Biodiversity Lab**	
9:00- 10:30	**Distribution models using Maxent and Machine Learning Technique**s Francisco Pando GBIF-Spain In this session we will use a practical approach to introduce the basic steps, tools, and data to create an SDM. Out of the many scenarios in which SDM can be applied, we will work with two: potential distribution of invasive species in a new territory, and changes in species distribution in future climate change scenarios.
10:30 -11:00	Tea and coffee Break
11:00-12:30	**Distribution models using Maxent and Machine Learning Techniques** Francisco Pando GBIF-Spain Continued from above
12:30-14:00	Lunch break
14:00-15:30	**National Biodiversity Assessment: Species & Ecosystem Assessment** Andrew Skowno, Dewidine Van Der Colff & Maphale Monyeki SANBI This session will have three talks: an overview of NBA (Andrew Skowno), followed by an in depth look at species red listing workflow (Dewidine Van Der Colff) and ecosystem red listing workflow (Maphale Monyeki). There will be a participatory task on the Red List of Ecosystems using Rstudio (Maphale Monyeki).
15:30-15:45	Tea and coffee Break
15:45-16:30	**Conservation planning** Douglas Harebottle – SPU In this session we will briefly introduce systematic conservation planning, its foundations and concepts and introduce Marxan as a conservation software tool. We will then assess the use of a waterbird index (and briefly highlight bird atlas data) as a theoretical application and unpack its relevance and limitations within the conservation planning paradigm.
16:30-17:00	Wrap Up

Day 5, 30th June 2023 **The Iziko South African Museums, Biodiversity Lab**	
9:00- 10:30	**Conversation on search engines- ChatGPT** Francisco Pando GBIF-Spain In this session we aim to explore collectively the impact of this AI (a Generative Pre-trained Transformer Chatbot) in our work. We will explore a few scenarios in which they can be used (e.g., research, report writing, translation, code writing). We will pay some attention to some of its

	limitations; we may linger a bit on what it is called "AI hallucinations". Besides ChatGPT, we will explore and compare some other chatbots such as perplexity.ai or elicit.org.
10:30 -11:00	Tea and coffee Break
11:00-11:30	**Course Evaluation & Closing remarks** (Fatima Parker-Allie SANBI-GBIF/IZIKO)
	Lunch

Appendix 2. Workshop trainers, participants and support staff

Trainers		
Fatima Parker-Allie	Node Manager, South Africa	SANBI-GBIF
Francisco Pando	Node Manager, Spain	GBIF-Spain
Katia Cezon	Data Science expert	GBIF-Spain
Fernando Aguilar	Data Science expert	GBIF-Spain
Douglas Harebottle	Director: Risk and Vulnerability Science Centre	Sol Plaatje University
Morne Du Plessis	Lecturer- Genetics	University of Free State
Vernon Visser	Research fellow/Senior lecturer	University of Western Cape
Andrew Skowno	Lead: National Biodiversity Assessment	SANBI (NBA)
Maphale Monyeki	Impact Assessment- Ecosystems	SANBI (NBA)
Dewidine Van Der Colff	Impact Assessment- Red List Scientist	SANBI (NBA)
Participants		
Bianca Greyvenstein	Researcher	North-West University
Christie Anne Craig	Conservation Biologist	Endangered Wildlife Trust
Claire Julia Paranzee	Intern	SANPARKS
Mafanela Clearance Msini	Scientist	SANBI
Curtley Wayne Tonkin	Intern	SANBI (NBA)
Dian Spear	Data Management Programme	SANPARKS
Gabriella Leighton	Post-Doctoral Researcher	Rhodes University
Janine Rose Baxter	Scientist	SANBI
Morena John Mapuru	Scientist	SANPARKS
Judith Botha	Senior Manager	SANPARKS
Aliza le Roux	Associate Professor	University of Free State
Precious Letebele	nGAP Lecturer	University of Free State
Martin M. Chari	Post-Doctoral Researcher	University Fort Hare
Nonkululeko Ntshangase	Principal Herbarium Technician	SANBI
Oliver Cowan	Conservation Scientist	Endangered Wildlife Trust
Daksha Naran	Volunteer	Albany Museum
Piet Monegi	Post-Doctoral Researcher	ARC
Robyn Manuel	Intern	SANBI
Helen Snethemba Ndlovu	PhD Candidate	UKZN
Siyabonga Zamisa	Database Technician	KZN Museum
Tiwonge Mzumara-Gawa	Lecturer	Malawi University of Science and Technology
Hannelie Snyman	BRAHMS Database Content Manager	SANBI
Support Staff		
Tshepiso Mafole	Intern	SANBI
Minoli Appalasamy	Intern	SANBI

Acknowledgements

Ms. Parker-Allie and Dr. Francisco Pando on behalf of SANBI-GBIF and GBIF Spain would like to express a warm note of appreciation to the GBIF secretariat and the CESP review panel for selecting this project, with this training event being the culmination of this project. A joint project between the Nodes has been on the wish list for many years, and the implementation of this initiative has therefore enabled this cross-continental collaboration, with the aim to support biodiversity informatics science at the national, regional, and global levels.