Country Studies on the Scientific Landscape for Access and Benefit-Sharing

Notes on Methodology

Paul Oldham (Ph.D)

Introduction:

This document briefly describes the methodology used to develop the eleven African country studies on the Scientific Landscape for Access and Benefit Sharing.

The aim of the research is to:

1. Provide an overview of research activity in a country;

2. Identify research publications involving genetic resources and traditional knowledge from the country;

3. Identify the researchers, organisations and networks involved in research on genetic resources and traditional knowledge from the country as a basis for engagement activities.

Data Source:

The research involves searching the Thomson Reuters *Web of Science* database of scientific literature for references to a country name (e.g. South Africa) in two areas of the *Web of Science* database.

1. The address field (for authors)

2. The topic field (Title, Abstract, Author Keywords and Keywords Plus (citations))

3. The year period is set from 1945-2015 (mid-2015) and all available subsidiary databases are searched.

The aim of this approach is to identify those scientific publications that originate from researchers in a country and those publications that make reference to a country in the available data fields.

Data Preparation:

The results of the searches are downloaded and compiled (using VantagePoint analytics and text mining software) in a series of steps:

1. General Overview. This consists of all available publications on a country across all subject areas.

2. Species Data. This involves searching the available data fields for references to species names appearing in the Global Names Index (GNI) of over 6 million species

names including synonyms and variant spellings. The raw results are then used to query the Global Biodiversity Information Facility (GBIF) database using the *rgbif* package in RStudio to retrieve available kingdom data. In a second step, we use genus names. The data is then cleaned and cross-tested to ensure the correct allocation of kingdoms to species names. We are increasingly using the *taxise* package in R to improve name resolution and reduce the possibility of errors on genus matches.

The outcomes of this exercise is a dashboard presenting species information by kingdom and an overview of species related activity.

3. Organisation network. Organisation names are cleaned in VantagePoint and then visualized as networks in Gephi network visualization software. Network visualization displays the relationships between organisations with researchers who are publishing articles involving species. This is particularly useful for identifying research networks and clusters around species inside and outside the country of interest.

4. Author network. Author names are cleaned in VantagePoint using match criteria (organisation names) to combine variants of names of the same person. Networks of coauthors are visualized using Gephi. Clusters or communities of authors who carry out research together are coloured on these networks using the modularity class algorithm. Colours are allocated based on the strength of the co-authorship links between the researchers.

Author network visualization is very useful for identifying prominent researchers working on species and networks of researchers who work together.

5. Plants. The data is divided into a subdataset containing plant data to build a dashboard presenting an overview of activity relating to plants. Network visualization is repeated for organisation and author networks to focus on activity involving plants.

6. Traditional knowledge. In VantagePoint the words and phrases from titles, abstracts, author keywords and keywords plus are reviewed to identify traditional knowledge related terms such as traditional knowledge, medicinal plants, indigenous knowledge, local communities, pastoralists, ethnobotany etc. Because a wide range of terms may be used, as the research has developed a thesaurus of terms has been created to capture results likely to involve traditional knowledge or indigenous and local communities. In addition for each country anthropological data sources are consulted for names of indigenous peoples (e.g. Khoisan or Turkana or Luo) and any known variants (Khoe San etc.). These names are added to the thesaurus.

The raw traditional knowledge data for each country is manually reviewed because of the likelihood of noisy results. For example the term indigenous or tribe is most commonly used to describe the status of a biological species, not members of a human society. On this basis the raw TK data is refined to a final list.

An overview dashboard is developed for TK and the organizational networks and author networks are mapped in Gephi as above.

Further information on TK is provided below with respect to the limitations with the existing approach.

7. Marine. In some case marine dashboards and networks are provided. This data is based on the use of the World Register of Marine Species (WORMS) list of marine species. We use a refined list of 382,680 species that exclude known model organisms. However, the WORMS list suffers from the problem that it describes marine and aquatic species (aquatic terrestrial environments) and also organisms that live in association with those environments (e.g. mosquitos or shoreline plants). While technically correct this can distort the presentation of results because it will tend to privilege high intensity research areas (e.g. relating to malaria) at the expense of what might reasonably be considered to be marine or aquatic species.

In response to this problem efforts have been made to limit the data to marine species. However, significant additional work would be required to fully clean and classify the marine and aquatic data and is beyond the scope of the present research. Marine data, where available, should therefore be regarded as indicative rather than definitive.

The Tableau Workbook:

In the final steps the results are compiled into a Tableau Workbook for each country. Hyperlinks are added to all major data fields (organisations, authors, species etc.) to allow the reader to interact with and explore the data for themselves.

The data is prepared in Tableau Desktop Professional software and made publicly available using the free Tableau Public service. This allows a reader to:

1. Interact with the data in a web browser. Workbooks are presently located here: https://public.tableau.com/profile/poldham - !/

2. Download the workbook for use on their computer by signing up for Tableau public and downloading the software from here: https://public.tableau.com/s/

Limitations:

The approach described above provides a means to map the scientific landscape for access and benefit-sharing in a specific country and to identify relevant species, organisations and researchers. However, there are some limitations to this approach and it is important to understand them.

1. The data is limited to the available contents of the Web of Science database. Web of Science contains articles from nearly 12,000 journals as well as conference proceedings and a limited selection of books. The data and analysis will therefore privilege publications in journals listed in Web of Science. It will not cover local journals that are not covered in Web of Science, or books and book chapters that may be important forms of dissemination of biodiversity information either historically or into the present day.

2. The data will only capture records from researchers where the country name is mentioned in the address field or the topic field. This means that the publication portfolio for a specific author will not be complete in many cases (for example where a country is mentioned in the main text we do not have the ability to capture such results). When discussing the data with authors themselves note that the data is limited by the source (e.g. the journals in Web of Science) and by the search criteria. This will overcome potential misunderstandings where researchers believe their scores are too low.

3. At present the search results include Keywords Plus. These are terms taken from the titles of cited literature in the main results. It appears likely that this will introduce some false positives into the datasets but is presently difficult to control. This may be addressed in future work to adjust the landscapes. For the present keywords plus data is retained.

4. Species names. Because of the large scale of the words and phrases in the title, abstract and keyword fields, species name matching from the Global Names Index is presently carried out as an exact match. This will not capture abbreviations or non-binomial species names that are not in the Global Names Index. Future work could address this issue where it could be demonstrated that there are significant gains. For the present the most timely approach is to use exact matches.

5. Species names and sources. We presently seek to identify any species name that appears in the data and we do not seek to limit the results to species known to occur in the country of interest. The advantage of this approach is that it provides a fuller picture of research involving species (regardless of its origin).

5. Kingdom matching. We primarily use the Global Biodiversity Information Facility (GBIF service) to match species names with kingdom information using the rgbif package in RStudio. All taxonomic data sources are incomplete and this results in a list of 'missing' species names. We address this using genus name matching. However, the same genus name may appear under more than one kingdom. The main problem arises with plants and animals. In response to this the kingdom allocations in conflicting cases are manually checked against the full species name and corrected where relevant. We are experimenting with the use of the taxise package in R (part of the ropensci suite see https://ropensci.org/) to improve name resolution.

6. Department names. Extensive name cleaning (typically to 5 records) is carried out on organisation names by matching variants of organisation names using the author name field. However, department names are considerably noisier and are presently presented in raw form or partially cleaned. This means that data on department names will typically undercount the records for a particular department. Resolving this issue would involve a major investment of research time relative to the likely return. For that reason organisation names are the main focus of analysis with department names provided to facilitate engagement with relevant departments in organisations.

7. Traditional Knowledge and Indigenous Peoples and Local Communities. There is no definitive list of traditional knowledge related terms or of the names if indigenous peoples and local communities in particular countries (one option includes the use of the *Ethnologue* catalogue of language names). We are therefore developing a thesaurus of terms as individual country data is developed. However, this presents a problem relating to completeness (has everything been captured that needs to be captured) relative to accuracy (the introduction of irrelevant results). It is likely that additional investment of time could improve the balance between completeness and accuracy but is not foreseen in the present project. In addition, note that the focus of Web of Science on data in journals is likely to miss publications focusing on TK in books, edited volumes, theses and local journals not listed in Web of Science.