Ollscoil na hÉireann, Corcaigh

National University of Ireland, Cork

"Creating Humanistic interfaces for complex scientific networks"

Athene Storey Cosgrave - 119406136

University College Cork

Digital Humanities and Information Technology

Supervisor(s): Dr. Sabin Tabirca

Second Reader: Dr. Pedro Nilsson-Fernández

I declare that, unless otherwise stated, this report and the project described is entirely my own work.

Abstract

Developing from previous research in The Royal Botanical Gardens Kew, Science directorate, this project looks at creating a web based user interface for complex scientific networks. The aim of this project is to use open access methodologies to increase accessibility of a vast historical and scientific knowledge bank with data collected from community led initiatives such as IPNI(International Plant Names Index) and GBIF(Global Biodiversity Information Facility).

The data used for the end network of this project is an output from a combination of the abovementioned platforms and repositories that is cleaned and verified using a tool developed by both myself and Dr. Nicky Nicolson over the last two years which will also be explained at length through the course of this report as that is the core of this project as a whole and the development of the artefact interface for this project would not have happened without the tool. Later in the report to show exact uses for this data are small case study analysis for country to country data as well as looking at specimen movement through old imperial routes

The hope is that the methods and practice behind this work will help bridge the gap between experts and researchers in the realm of botany and data mobilisation by highlighting the advantages and learnings of interdisciplinary work. The drive behind this project stems from the aspiration to connect post colonial and indigenous communities with their botanical history, looking specifically at scientifically important specimens from their native biomes

Glossary of terms:..... 4
Introduction..... 5
Literature review and Environment Scan:..... 8
Implementation and technical breakdown:..... 15
Modules used:..... 20
Fixing coordinates in networkx:..... 22
FOR IPNI NETWORK IN GEPHI:..... 27
Case study analysis:..... 29
Results:..... 35
Reflection:..... 36
Bibliography..... 38

Glossary of terms:

Holotype:

- a single type specimen upon which the description and name and characteristics of a new species is based

Isotype:

- A plant specimen that is a duplicate of or very similar to the type specimen and can be used as a reference specimen if the (holo)type specimen is lost

Herbarium

- A herbarium is a collection of preserved plant specimens that have been stored appropriately, databased and arranged systematically to ensure quick access to students, researchers and the general public for scientific research and education (https://museum.wales/blog/1910/What-is-a-herbarium/

Collecting Event:

- A botanical expedition in which specimens are collected, often multiple type specimens at once

IPNI:

- International Plant names index A online database continually updated/maintained by teams in Kew's science division that contains all properly published new plant descriptions. Holds all published plants, not just digitised collections.

GBIF:

- Global Biodiversity Information facility: Similar to IPNI but only contains information from digitised collections

Introduction

The project was started within the department of Digital revolution in Kew's Science directorate [1]by myself and my supervisor, Dr. Nicky Nicolson. It began as a year-long proposal and has had enough support and interaction from external researchers that we have decided to continue on it voluntarily so as to ensure the ideas we have formed and developed came to fruition. This Digital revolution Department within Kew works closely with the digital humanities department in its partner colleges to ensure it can create new and accessible ways to mobilise our data while being in touch with new technologies.

The main research question being" Developing human interfaces for complex scientific networks" which will be the continuation of the work done in Kew by way of making a humanistic interface for the data developed so that those outside of the main research audience use this data and interact with it in a meaningful way. To understand why we need to get to the point of making this data accessible, beyond the fact that we should try to make all data in some way accessible, we need to look at the development and creation of the initial data tool that started in Kew and has continued into this time. *"A network model approach to understanding the global institutional distribution of type specimen data*".

The project developed as we pursued it and incorporated the input of other scientists and institutions. For context, About 2000 new plant species are published annually via a code governed process. Each description cites a set of type specimens, these are scientific specimens used to describe a new plant species, linking the abstract name to a concrete object. The members of type specimen sets are usually held in different institutional collections, often in different countries. When institutions share material from a specimen set, they share an interest and can be modelled as linked nodes in a network. Specimens are being digitised and aggregated to maximise access and reuse, but there is a mismatch between areas of high biodiversity (global south) and locations of science infrastructure (collections, labs) due to colonial history. Progress varies throughout institutions and digital representation is incomplete. This project applied

dynamic network modelling to analyse specimen data from Open Access platforms such as:

• IPNI (International Plant Names Index) [2]

• GBIF (Global Biodiversity Information Facility)[3]

• Index Herbariorum(New York Herbarium)[4]

These 3 data sources were chosen due to them being open source as well as wealthily populated with data. IPNI is based on Index Kewensis (originally funded by Darwin), the IPNI project editorial team record data from published protologues, standardise authorship and (since c 1997) record type specimens. This platform acts as a record base for validly published specimens around the world.

GBIF contains information on all digitised records, which although is a majority of them it does not cover all bases as many herbaria haven't the resources to digitise their entire collections. GBIF is maintained by a number of different institutions and governments. The data is collected via crawling and is first moved from multiple different sources to a Darwin Archive Core database then onto the GBIF servers. The coding tool after it was made does output the findings of the analyses into DWCA files which are compatible with the GBIF databases and get regularly rerun and updated onto their servers.

Index Herbariorum is a project developed by the Steere Herbarium in the New York Botanical Garden. This acts as a record of the world's active herbaria, containing information regarding location, collections, staff, etc. These resources are continually updated and improved upon by their respective teams which helps contribute to the sustainability of this project. Every Herbarium has what we call a herbarium code(type_holder_code in the scripts), this is a code assigned by the world biodiversity forum and council after their registration as a Herbarium, these codes are assigned to the specimens held within these institutions as a way to track them, over time these code get duplicated or transferred as institutions shut down or merge together and often the specimen information will have notes about this in the exploded specimen information however in some cases, because the community is so small and connected,

it requires someone to pick up a phone and call another researcher to ask where that specimen they collected twenty years ago has ended up and update the records by hand.

This interactivity by the community and open communication allowed for this project to flourish when we ensured that it was regularly shared in our internal forums and at conferences with other researchers to ensure we could continually receive feedback and wants from others in the community for the tool. We are all aware that the distribution of these important specimens was going to be inherently skewed towards past colonial and imperial powers but no one had yet to bring all that data together, clean and verify it against other sources and make it easily accessible for the research of others. The tool that was created was never meant to end up as something with a User Interface but as something that could assist other researchers in our field to continue their work and analysis in new and interesting ways.

Now being in Digital humanities does mean that everyone wants a UI for all data that someone might find useful, which is where the continuation of the above project comes in. The data output from the tool after it cleans and strips everything, then renders the network, is not in any way clear to those who aren't in this very niche field of research. This is something that never bothered us in Kew as we knew what we were doing and quite honestly didn't think anyone outside our field would be interested in using our data or findings and those within our field would have no issue understanding and navigating these complex files with tens of thousands of data points. The main point of this project was not to actually make an incredibly pretty and perfectly functioning web interface that would show exactly what we had achieved in our original tool but to look at how we might translate such a complex network into a functioning tool and how to incorporate the complex levels of analysis that were able to be computed within the command line interface to an interactive online network.

Literature review and Environment Scan:

The reason the project from Kew needed to be done was because there was so little done similar to this before and because of that there is no way to create a well developed literature review but listed below are pieces that were used in the research and development and then expanded upon here to give context for their relevance and how they shaped the flow of work.

Colonialism has shaped scientific plant collections around the world – here's why that matters [5]

(Park, Daniel. "Colonialism Has Shaped Scientific Plant Collections around the World – Here's Why That Matters." The Conversation, June 12, 2023. http://theconversation.com/colonialism-has-shaped-scientific-plant-collections-around-the-world-heres-why-that-matters-207375.)

Although this piece doesn't actually hold a great amount of information or be well sourced it is still a well written piece explaining the basis for the work all the herbaria are hoping to start or have started in the last few years. There is also a book that is developed and redone every few years by the team in the Kew Herbarium and the Natural History Museum but it caters to the specific scientific work within a Herbaria instead of looking at the ethical work of a conservator that many institutions are starting to focus on.

The end of the article outlines three different directives the team would like their project to take which you understand the author wrote as standards they had developed but these are actually mostly all baselines within the Nagoya Protocols that were developed back in 1997(most recent update being 2015 for the specific directive that author talks about) so it makes one less inclined to use such an article as a well researched piece but it does still allow one to understand a basis into this information.

Nagoya Protocols: Centre for Biodiversity (Unit, Biosafety. “About the Nagoya Protocol,” June 9, 2015. https://www.cbd.int/abs/about/.)[6]

This is a legal framework agreed at the Convention on Biological Diversity that allows for:

“

- Establishing more predictable conditions for access to genetic resources.
- Helping to ensure benefit-sharing when genetic resources leave the country providing

the genetic resources.”

These protocols have been developed to ensure fair sharing of genetic material and resources so that the practice of institutions from countries who are less biodiverse but hold high numbers of specimens do not ‘hoard’ the information they collect and extract. This framework ties into the work many institutions are doing to focus on the ethical work of a conservator and field collector to ensure that practices that had been around for hundreds of years and not questioned are now adjusted to reflect the thinkings and developments of today's society in regards to knowledge repatriation and fairness among all levels of researchers and institutions.

The data for this project once its finished will be used for at least two proposals for the amendments to the protocols by a botanist and taxonomist in the Natural History Museum London. The amendments to the protocols are only allowed to use proper verified data and ‘properly published’ information, for something to be properly published is a botany specific term and requires it to go through multiple peer reviews and councils to gain that title. Thankfully that is only when referring to a plant specimen so the data outcomes from this have been peer reviewed and are allowed to be used as a verified source for the proposal for amendment in the World Biodiversity conference.

Equity and Access(Decolonising collections):[7]
(“Equity and Access.” Accessed December 7, 2023. https://www.rbge.org.uk/about-us/equity-and-access/.)

This project has been an ongoing initiative by the Botanical Gardens in Edinburgh to decolonise their collections from its colonial roots, either by specimen repatriation, knowledge repatriation or by developing institutional networks to benefit herbaria negatively affected by the regime of the British empire.

They have since published a book regarding their work and what other herbaria can do to similarly affect change and Kew has begun to do so with their Economic Botany collections. An Economic Botany collection does not include plant specimens but it is just as important for the history of a botanical garden. It is a collection of collected artefacts that botanists have gathered during their travels to collect plant specimens. Often when collecting plant specimens the botanists will live in the local community for a long period of time, especially when it would be a medicinal specimen so as to gain the trust of the local community and gather the relative information associated with the plant to ensure that it will be useful to them on their return, due to their integration into these local communities they would often be gifted items and upon their return these items would be added to the Economic Botany collection of their institution. The people place plants project by Kiri Paramore along with the work my the archivers in Kews collections looks to illustrate the communities surrounding the artefacts as well as hold them in place for those that may not have the means to store them safely but also to ensure that researchers have all information currently studied from them as well as the ability to gain access to these collections at any time. A prime example of this is the Maori ceremonial feather cloak in Kew that has now been re blessed by elders in the Maori Community from Aoteroa and has been turned into multiple 3d models by the team to ensure that the indigenous community have access to this item from their past while still having it stored safely in the collections in Kew.

Personal Private Herbaria: (Roma-Marzio, Francesco, Lorenzo Peruzzi, and Gianni Bedini. “Personal Private Herbaria: A Valuable but Neglected Source of Floristic Data. The Case of Italian Collections Today.” *Italian Botanist* 3 (March 17, 2017): 7–15. https://doi.org/10.3897/italianbotanist.3.12097.) [8]

This was a piece of literature we came across after looking through the dataset we had already roughly cleaned and developed when working on the research pipeline. In our search we found that a number of the type specimens has 0.00 points for their latitude and longitude, this made no sense as all holotype and isotypes are accounted for where they are stored as we have ensured to only include ones that had herbarium codes attributed to them.

Going back to look at what had been going on only to realise all of these were being stored in private herbaria of botanists that apparently other botanists knew about but it was never made a point of information in official data so there was no way to double check things unless you knew the botanist personally, that investigation led me to the above piece of literature to see how we could find out similar information on a global scale.it was decided to then remove any specimens stored within private herbaria from the data, this removed nearly 3000 specimens from our analysis which absolutely has affected the outcomes of the research but there is no way to avoid this at this time. A future project proposed with the New york Steere Herbarium is that there be a record developed of all the type specimens within these private herbaria but that would lead to an individual having to actively go to theses herbaria and collect this information by hand from the botanists themselves.

The herbarium as personal: (Flannery, Maura. “The Herbarium as Personal.” *Herbarium World* (blog), November 18, 2019. https://herbariumworld.wordpress.com/2019/11/18/the-herbarium-as-personal/.)[9]

This work is not academic literature but it did make me look at the practice of holding a herbaria in a different way, of course I will be focusing on these large institutions but were there any way to get enough information to map historical personal herbaria of prominent individuals? It would be a fascinating project.

The idea for all our work to have started with a small passion from an individual to now have on average 6-8 million specimens. Kew began as a gift to the wife of King George and was then developed from her personal herbarium into the institution we know today, how many others have similar histories?

Similar Projects:

The projects outlined below cover those that have code and data very similar to what was being attempted with this project in regards to data collection, cleaning, analysis and overall visualiations, they are all open access and the repositories are linked where possible.

Brazilian Herbaria: An Overview

(Gasper, André Luís de, João Renato Stehmann, Nádia Roque, Narcísio C. Bigio, Ângela Lúcia Bagnatori Sartori, and Guilherme Salgado Grittz. “Brazilian Herbaria: An Overview.” *Acta Botanica Brasilica* 34 (August 3, 2020): 352–59. https://doi.org/10.1590/0102-33062019abb0390.)[10]

This Project looked at a specific set of data collected from questionnaires filled with targeted questions that was sent to all known active Brazilian herbaria. The questions were targeted to the best and most accurate data possible so that the later analysis of the information would yield the best and most useful results. Had they just used information from existing platforms, such as index herbariorum, they would not have

been able to do as comprehensive an analysis and in their analysis, they unintentionally created the perfect dataset for a case study of social impact mapping in regards to institutional relationships with local governments.

The project uses Sci-py as one of its modules as it does large scale language analysis and modelling, now it had been adapted for image models but at the time this project was being developed it was the best python module for the case. They also use gephi as a tool but quickly move away from it due to its inability to allow for accurate geoparsing. Continuing their work in Geopandas and Matplotlib. I went back to this project multiple times to look at their code when having issues in mapping my data through networkx and could not for the life of me figure out how they got it to work.

This would be incredibly interesting to see it done in more countries also to see it done on a decade long cycle to see how the movement of funding and specimens rotates around countries and institutions.

Index Herbariorum:
(“Index Herbariorum - The William & Lynda Steere Herbarium.” Accessed January 9, 2023. https://sweetgum.nybg.org/science/ih/?_ga=2.74776530.1754210885.1673300121-545275426.1663748227.)

This is a project created and sustained by the New York Botanical Gardens past director; Barbara Theere. We worked with Barbara on what she would like to see done with the data we had been developing back in Kew and how we could integrate it into the herbariorum website to look at how well herbaria are supported over the years.

This platform acts as a repository of information about all current active herbaria around the world, it is constantly being added to and developed as institutions shut down or simply disappear, as happened to herbaria in war torn countries, the Main herbaria in Al

Quds in Palestine has been since occupied and the current national Herbaria of Palestine sits in the West Bank, as of January 2023 many of the collections here have been moved to private collections in Jordan and Beirut for safety from the Occupational forces.

The recording of these institutions is important and the herbariorum will keep this information as up to date as it possibly can and it asks that the community who use the platform interact as much as possible with what they would like to see developed in the platform or if there are issues with their data as a whole.

The Coffee Network:

("Coffee_network/Coffee.Py at Main · Mmcordova/Coffee_network." Accessed March 24, 2023. https://github.com/mmcordova/coffee_network/blob/main/coffee.py.)
(Coleoni, Cláudia. "Beyond the Watershed: Mapping Global Coffee Trade Flows and Water-Related Teleconnections in Colombia," 2022. https://doi.org/10.13140/RG.2.2.12432.28167.)[11]

This project heavily affected coding the initial mapping aspect of the project, this is a network map of coffee trade routes in relation to how the historical geographical and topographical attributes of an area would have made them destined to be fertile soils for coffee plantations. The team working on this were incredibly talented and their well documented repository allowed for us in Kew to move forward with our mapping of data as it seems that every mapping module in python had begun to break or stopped being supported in the last 5 years.

The data behind this project is immensely impressive with an incredibly well coded network and analysis as well. The actual content we did not need as much, although it was incredibly interesting the main point of including this project was to show where some of the code methods came from and how we used projects tot develop our own.

Implementation and technical breakdown:

Although the data sources chosen for the project were all well maintained and monitored, there is still plenty of space for error and discrepancies. From the initial experimentation with these datasets we realised that the tool couldn't just act as a way to pull data, strip it and run a quick analysis for what we needed but that it would need to be verified against other sources to ensure that the data was cleaned of duplicates and other issues.

The number of issues with the initial exploration of the data showed a number of addition we would need to make to our data cleaning pipeline:

- The location of Type Specimens in Privates herbaria
    - Botanists have a habit of collecting specimens at all times no matter if they are associated with an institutions and often will have their own personal herbaria in their home, We do not have location data for their homes nor are the important specimens they have in their private collections associated with a herbarium code to assign to the collection event. In this case we had to add an additional script that went through the IPNI data and ensured that all the type specimen data had an associated herbarium code with an anctiver herbaria as often it will have an institutional code that might be out of date. In the case of trying to model this network we wanted it to have the endpoint of a visual tool so to have nearly 3000 data points that contained specimens with no location data at all, ended up with a heavily skewed network visually and when trying to analyse movement within biomes.
- Duplicates in specimens and institutions
    - Thankfully the number of duplicates within the network of Herbaria was

few but it did affect the analysis so much that until we resolved it there was no way to continue with our work as when we ran the full tool it would pull down fresh data files from the platform with the duplicates. This was resolved using a list of duplicates for specimens and Herbaria that was sent to the IPNI and IH team to adjust on their end in the platforms.

- Eg:

Data duplicates in the institutions DB (NYGB index herbariorum)

- node_id == 6 duplicates

    - 2 entries in the csv - 1 had code as NA, other has code as BARC

    - BARC contains more descriptive data than than NA but it missing lots of crucial information

    - in merge database only one node contains the location data and the rest are all null values, no clear indication as to why there are so many duplicates

    - assume NA was taken as a null value instead of as an institution code

    - same lat long for both inst but BARC does not seem to show up at all in exploded ipni types

- node_id == 219 duplicates

    - 2 different institutions with the same inst code

    - nong lam university

- University of Louisiana at monroe

- Status is permanently closed and does not contain location information

-code shows up as isotype holders in ipni exploded types but does not distinguish which institution it is

- node_id 5088 duplicates

- 2 entries into the csv, HIFPA, same institution for both

- location information same but reversed for both entries with slightly different input methods

- node_id 5498 duplicates

- UCBD for both nodes

- under 2 entries in the csv

- one for plant systemics institution

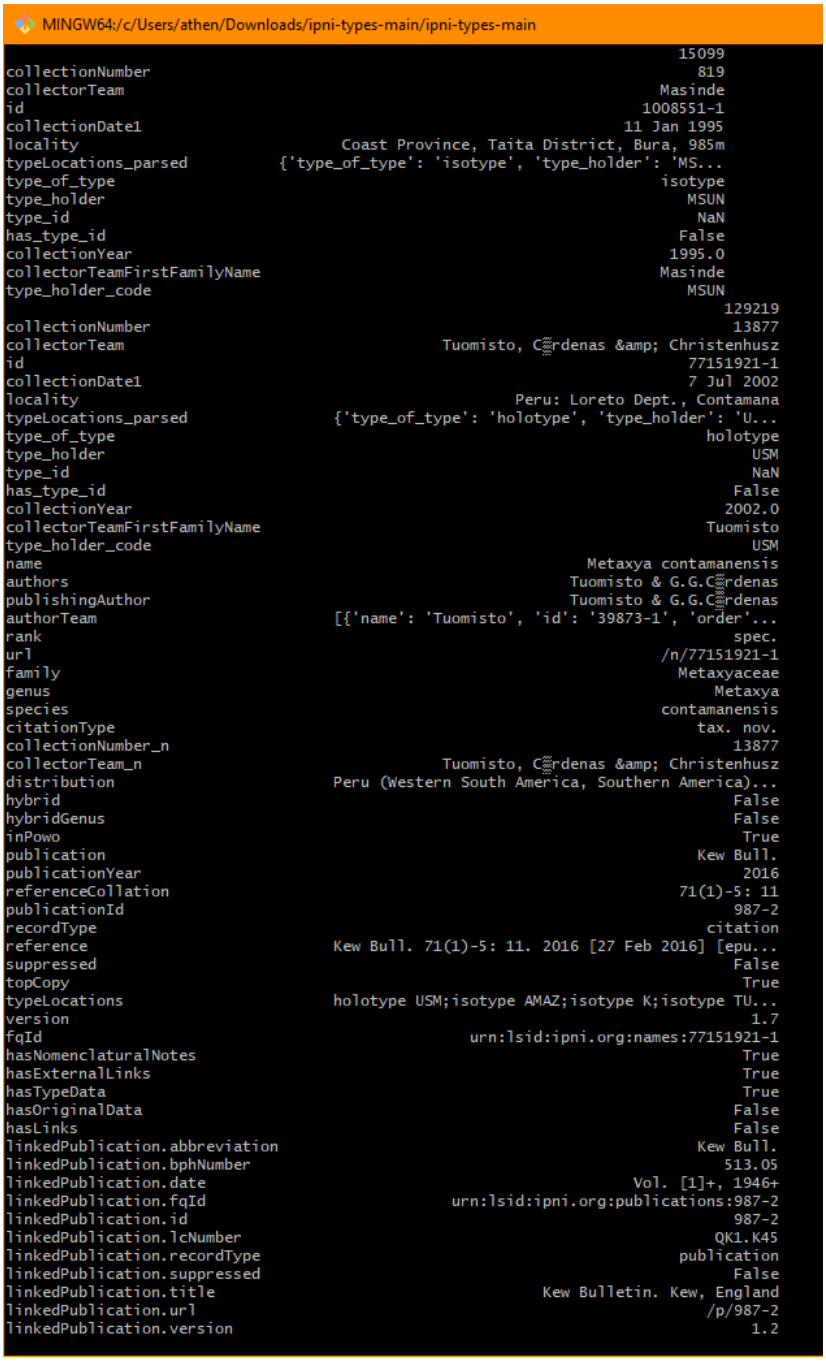
- one for botany

- both for taxonomic coverage of angiosperms

- looks as though one of the location datas has been rounded up which is the cause for the 2 entries.
- one entry contains slightly more detail info but is still the same institution

- Incorrect type allocation for specimens

    - This issue had to be resolved on a higher level than what we could interact with and needed review during the taxonomy nomenclature conference and we ended up just removing these plants from the data during the cleaning process to avoid having to try and navigate this as the project went on and we needed to do more complex analysis.



These sources contain a huge amount of data with each entry having ninety nine points of data and there being hundreds of thousands of entries in the databases. In the Image you can see the amount of data that is assigned to a single edge within the network, the final network has over 13,000 edges.

We ended up narrowing it down to 5 main attributes to assign to each node and edge.

- Herbarium Code (standardised code all herbaria are assigned)
- Herbarium Location (in the form of latitude and longitude)
- Number of type specimens shared between institutions
- ISO3 - This is a country code that correlates to a world code standard
- Node id (this was assigned within the data cleaning pipeline)

This project could have been coded in a few different languages, the main competitor to our final decision was R, as it was used by most other researchers in Kew for data analysis and visualisations but our decision to move away from it had a few factors

- Neither my supervisor nor I were comfortable coding in R
- We had example code from a few similar projects all done in python we could repurpose to make it easier on our end
- Python as a language has more resources for learning and is all in all a more mainstream language that meant that should the tool have issues when implementing it there’s more on the internet for them to use as a help or fix

The way the project was coded also follows Open Science methodologies, utilising code languages and modules that are well resourced and are community sustained so as to ensure the longevity of the project and pipeline. During the development of the tool we experimented with a number of different modules that promised some useful functionality but in the end were not well sustained or supported to include them in the end product without a large level of upkeep on our end that just wasn't feasible:

Modules used:

Dwca-writer

- Library that allows for the easy creation of Darwin archive files from rendered data

Geopandas:

Used in line with Networkx to plot the network using the node attributes of

Ipniutil:

- API used to access the IPNI libraries and data for initial retrieval

Ipysigma

- Allows for the exporting of the final data into a gefx files for use in gephi

matplotlib==3.3.1

- Working in conjunction with Networkx to plot the network and render the basemap

Networkx

- Python library used to render the network

pandas<2.0

- Used for data cleaning and analysis in the pipeline tool

Pygbif

- Used to access the data on the GBIF platform as well as assist in the data verification step of the tool

Pykew

- Used in conjunction with IPNIutil to act as a language standardiser

requests

Scipy

- Used to assist in analyses with ipni and gbif

Tqdm

- Used as a library to assist geopandas in the analysis of biome movement and Nagoya protocols

Xmltodict

- Common python library used in conjunction with scipy to create archive files and dictionaries needed for rendering the network.

Modules proposed but not used:

Mapping modules

- Cartopy
- Leaflet
- Basemap
- MPLLEAFLET
- Plotly
- Shapely

Network modules:

- contextlily

The pipeline tool is also made to need very little user input as to allow those who are not highly skilled in coding to use it, by using a makefile we were able to simplify the process for analysing the data down to the need for the user to only input a command line prompt for get back out cleaned and analysed data files and networks that they can then use for their own work.

Due to the fact that we decided to write this entire tool in python as it is a great language for scripting to create a self contained tool the best option for running everything was creating a makefile that would allow for easy changes to the code analysis tool by way of changing dates or data files but also to allow for localised analysis within the data by using specific make commands to target the building of specific files.

The only issue in this case was trying to get location information and basemaps to work alongside Networkx and python, it took about a month of work to figure out how to get everything to translate across to each other in regards to keeping the location data accurate and ensure it would actually show up in its correct locations.

Fixing coordinates in networkx:

- All attributes were contained within a dictionary as keys and values, with the index being the node_id, when trying to pass position as longitude and latitude it would give a key error or would say the objects were not iterable

    - the issue here was that network x was passing the work 'latitude' and 'longitude' as the objects for position for the values assigned to them. It was not a wild assumption to think that by giving it the position it would intuitively know to access the number value as some other modules would also do so.

    - one resolve was to pass long and lat into a single tuple called coordinates and pass that as the node position

    - the other resolution was to have it so label and pos were written to pass the value of the keys whenever it was accessed for each node which seemed to have worked well and showed a graph that vaguely looked like a world map, however when it then was added to put a basemap from geopandas underneath it it placed all the nodes on top of each other bar 1, I think the issue is the need for the module context lily which helps add basemaps to plots to give context to what needs to be done

    - FIXED: fixed by using geopandas to locate the nodes before plotting through a gdf instead of pushing everything as a dictionary which would have been fine was there not there requirement of a basemap and axis to match together,

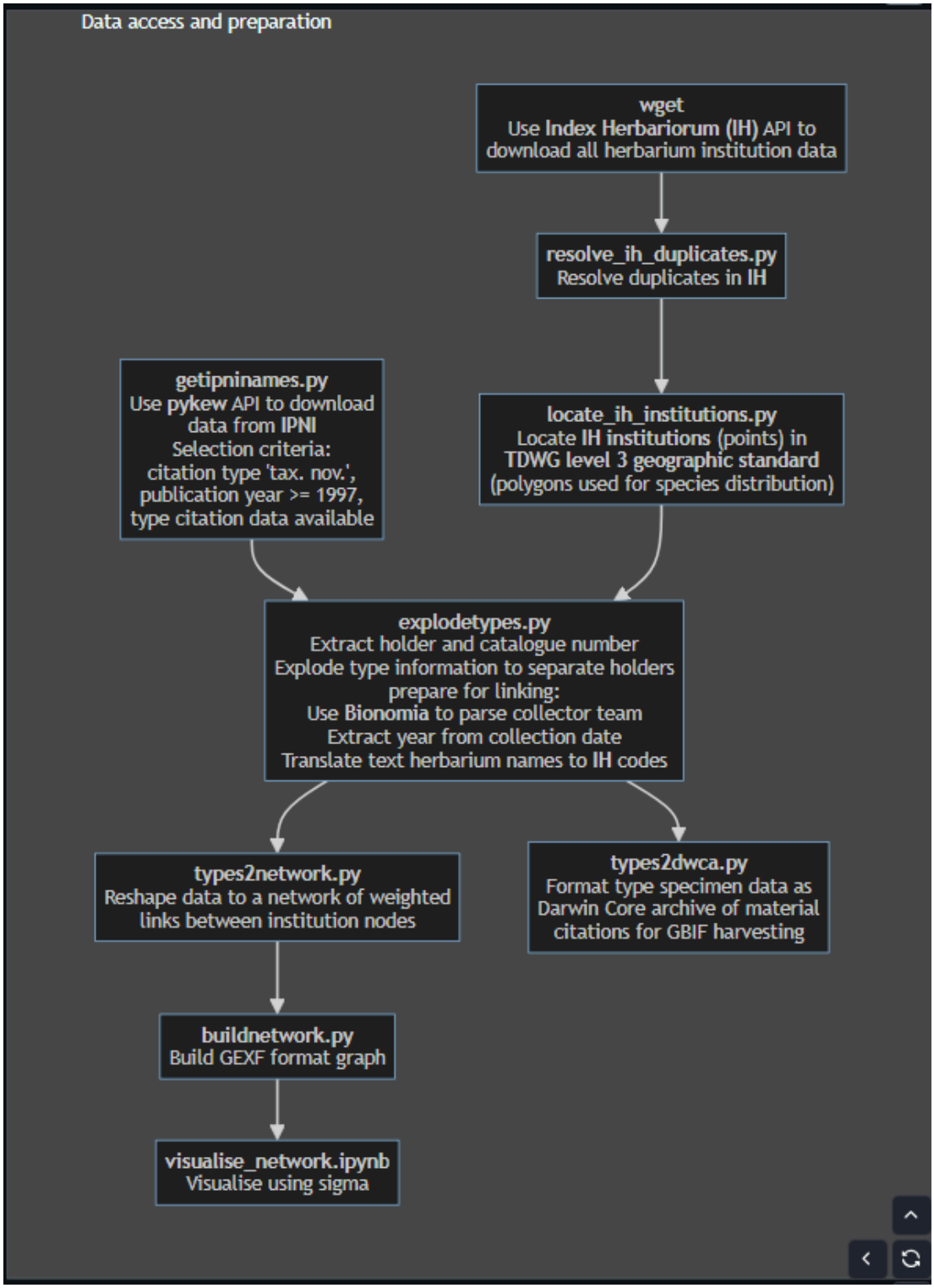


Fig 1

This fig shows the flow of the pipeline generated automatically by a git tool and adapts when there are major changes to the makefile.

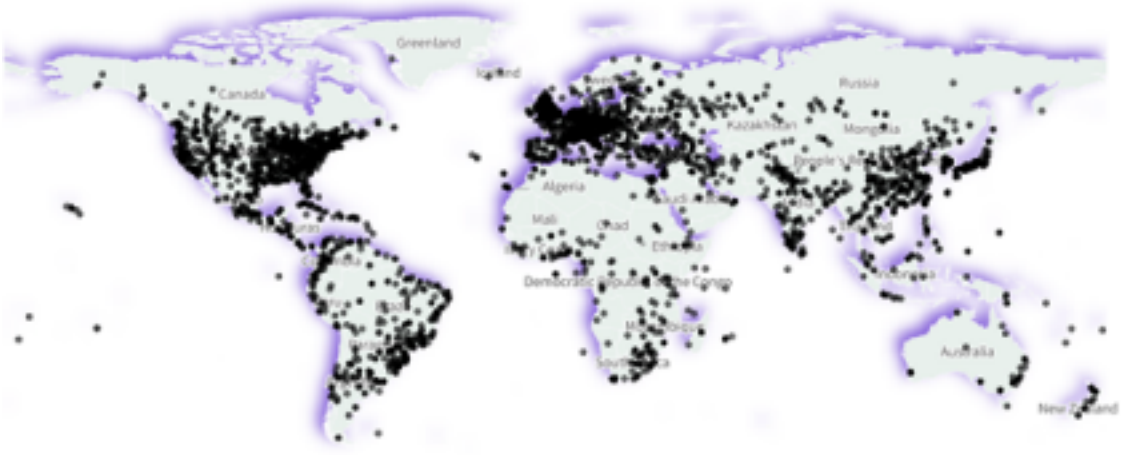


Fig2: Shows basic point map of all active verified herbaria in the globe

Fig3: jpg file output of project analysis showing herbaria with shared type specimen sets

Fig3 shows the result of the initial analysis done by the pipeline which is a network showing herbaria that share type specimens from the same collecting event. Going forward there are more analyses being added which have been influenced by individuals from other institutions who have shown interest in the end result of this project. The additional analysis will look at protocols that fall under the Nagoya Protocol.

One of the protocols dictates that specimens collected and taken from their country or polygon of origin to an institution in an alternate polygon must be published and named within 5 years of the collecting event to ensure that there is not a stall in the advancement of knowledge on the specimen. So to integrate this into the analysis the individual would input the year that the specimen was collected in and then the analysis would create a network showing herbaria that had collected in that year, where they collected from and those that had published the information about the specimen within the 5 year timeline, this proved to a lot more complicated to add into the pipeline than expected and is taking me more time than I expected to add in to the process.
Other layers of analysis are Social impact layers to help bring context to why institutions hold the relationships they do. For example how well a herbaria is resourced versus how well their collections are digitised or how often their online databases are updated. There is also the hope to, using an interface like Gephi, create an animated view of how relationships between herbaria developed over time and also use the context of time to give another view on how a country's socio-economic status at the time might have impacted the activity of a herbaria and its staff.

These analyses and outcomes were not part of the initial idea for the project and developed and we realised what could be done once we were able to properly mobilise the data, because of this the tools used had to change as well as our tactics as to how we approached the data and how much information we could handle. Initially there had been the idea to use interfaces or software to do the analysis for us as one of the first steps to help get an idea for how we would go about coding the pipeline and making it more accessible but we soon realised that it didn't actually give us a good enough idea of the process we would need to go through to properly clean and verify the data by hand and create a custom process.

The process within the coding pipeline is technically custom to the files we are using but with little altering can be used for other file formats and data orientations. Inside the makefile are a number of different commands that deploy smaller aspects of the whole analysis so that smaller changes can be made and looked at. These commands came from what other researchers wanted, such as above with the Nagoya Protocol, looking a

biome/polygon relocation and other functions that researchers deemed interesting enough to ask us to create.

The plan was to then make all of these available as interactive functions on the web app so that people can do quick analysis of the network as well as interact with it to better understand the data we created.

https://athenesc.github.io/Gephi-Viz/network/

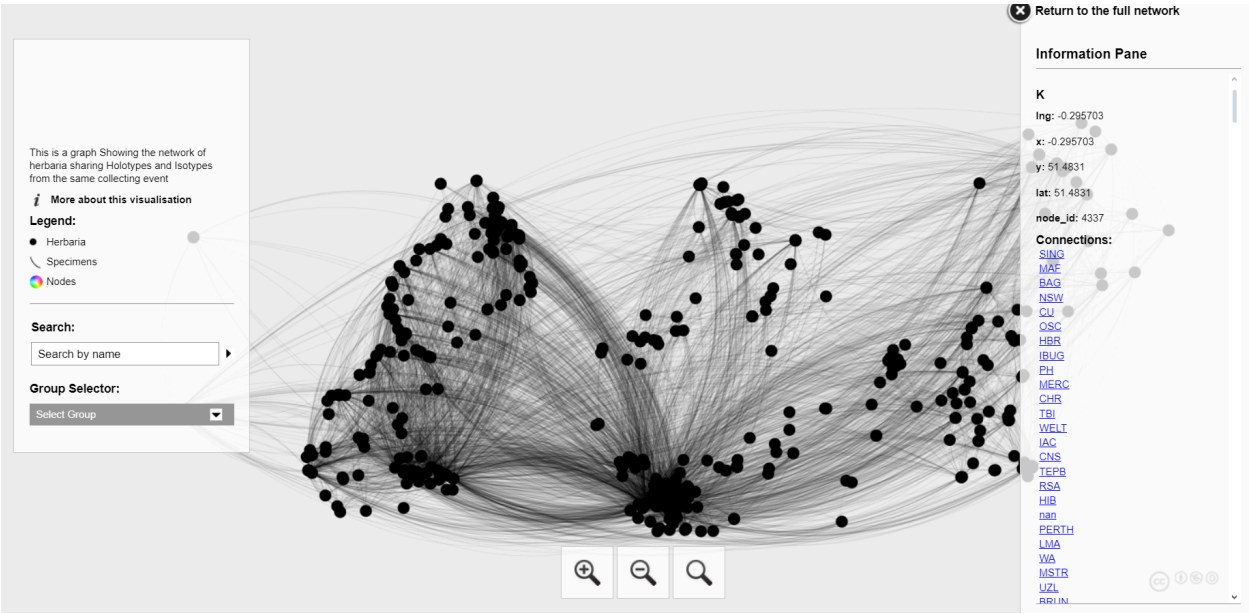


Fig4

The above image uses SigmaJS and github pages, The below image uses the gephi publish to web page functionality that uses the retina platform to execute the network. Visually the below image is nicer and more desirable but it is hindered by Retina's inability to handle custom functions and the fact that it has only been published in the last 9 months and needs regular work and updates done on it to keep it semi usable. I appreciate the Ouest was going for in creating this and it is useful for basic networks

https://ouestware.gitlab.io/retina/beta/#/graph/?url=https://gist.githubusercontent.com/ATHENESC/300c11f78d8987411bb3fddb416fe6ae/raw/2057a172f93f2bfeae743addfa7acfb29c6ac57c/network-2ba3c5ec-a19.gexf

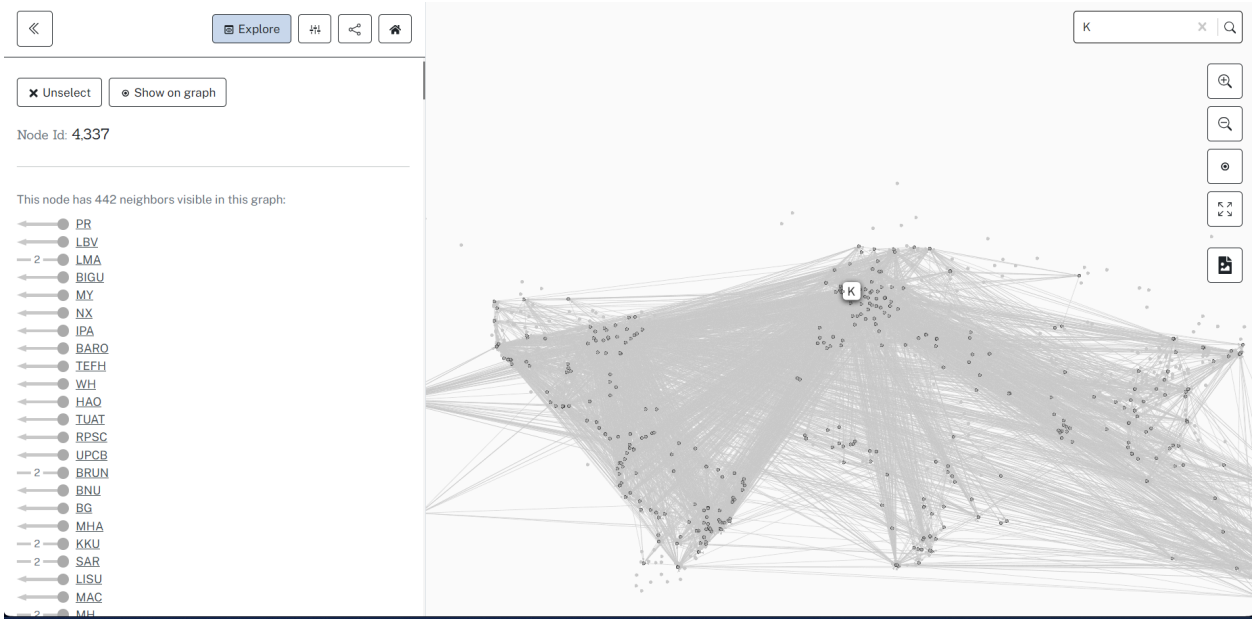


Fig5 retina visualisation

To get the network from the output file to gephi then to an online format, took an amount of testing. The output gexf file does work to transfer the network over to gephi as a finished network, but it does not work when trying to add the additional aspect of a basemap and search functionality just due the general limitations within gephi

FOR IPNI NETWORK IN GEPHI:

- load ipni node file first before adding any layouts into the node tab in gephi
- add ipni edge file next so that the ID on the node and edge files map
- if you have multiple workspaces going this wont work unless it is in the first workspace as gephi does continuous ID numbers, first node in the gephi node column must show as 0 or edges won't be able to locate the nodes
- graph will show up as square of dots in the overview, this is fine for now
- chose layout 'Map of Countries' and have projection as 'Mercator' (default) with the centre option unticked

- run layout
- go back to data laboratory and copy the data from the latitude and longitude columns to the lat/lng columns (lat/lng columns only appear after the map of countries layout has been run
- Go back to overview and now change the layout to 'Geo Layout' and change the projection to 'winkel triple'
- change the latitude and longitude option to select the lat/lng columns and keep the centre box checked
- run layout and the map of countries should line up with the node coordinates
- change node colour to make network more clear
- Export through Sigma Exporter
- Upload network file to public repository and deploy through github pages

Currently this is the easiest path to get the visualisation online, however , github has begun automatically using jekyll in all graphic deployments meaning that the stylised version of the network that make it easier to read an understand will not build through the pages and thus the artefact is stuck on a previous version until this is patched.

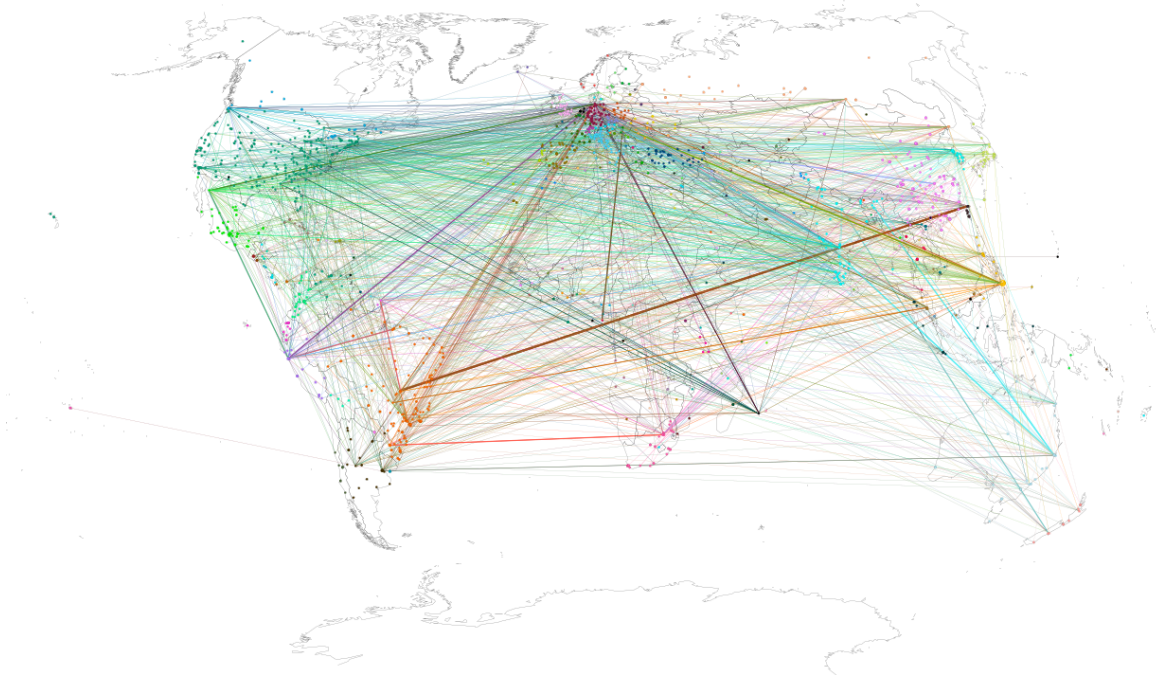


Fig shows all networks and source edges coloured to unique country ISO codes

Alternative web app options were:

Plotly with dash:

- This is a data visualisation library and hosting space that allows for quick integration and would have worked well in the case of visualising this project but does not currently have enough documentation and resourcing compared to gephi and although good for making a pretty visualisation, it was better to go with the more sustainable option

Storybook and Sigma

- This option only became apparent about 2 weeks before the dealine for this project but it uses a library similar to the exporter used to generate the webpage used in the end artefact and would have worked quite seamlessly had it been found about 2 months earlier to work with. Storybook is the online component to this and can be deployed through github pages as the gephi exporter had been for the final artefact.

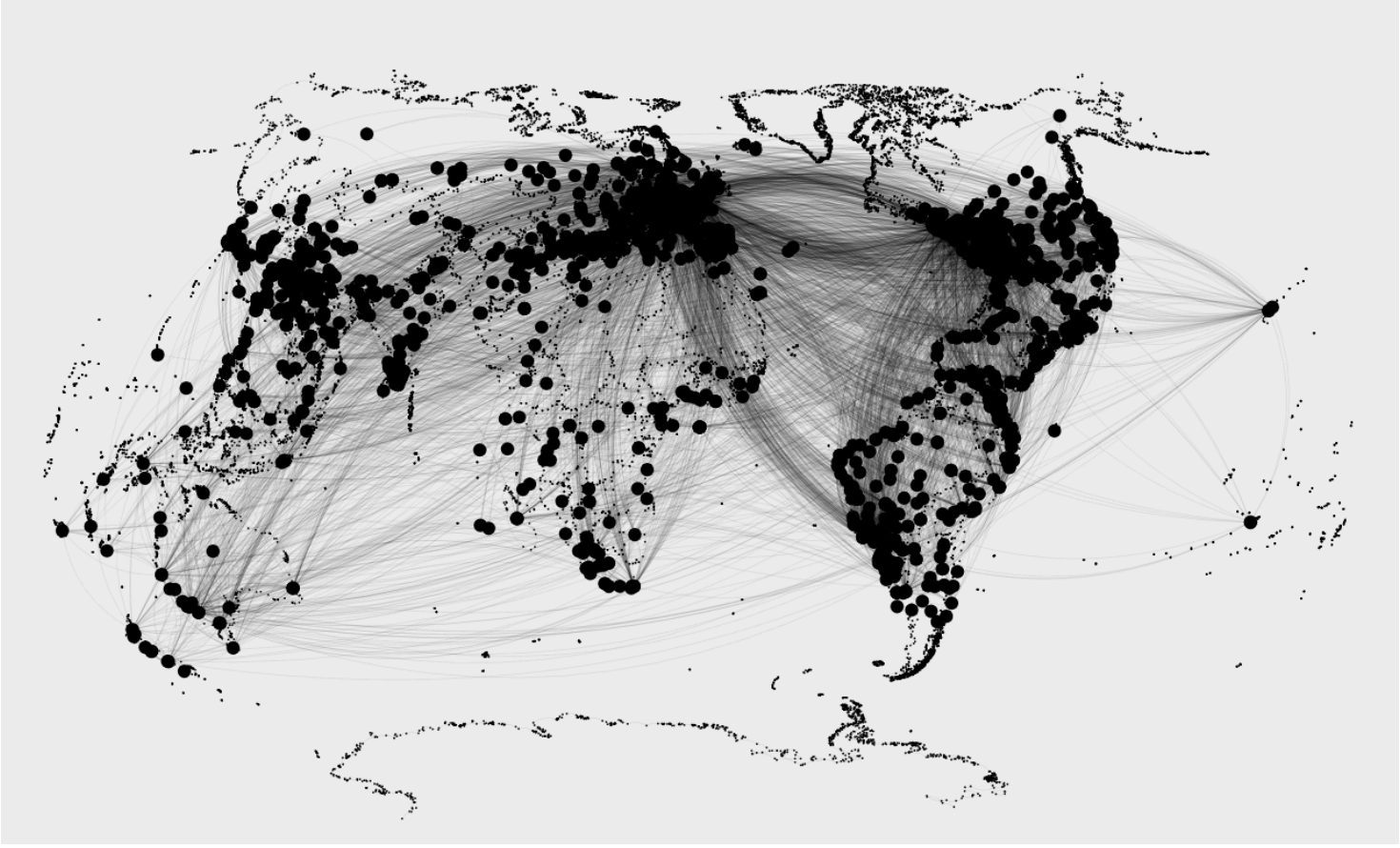
Case study analysis:

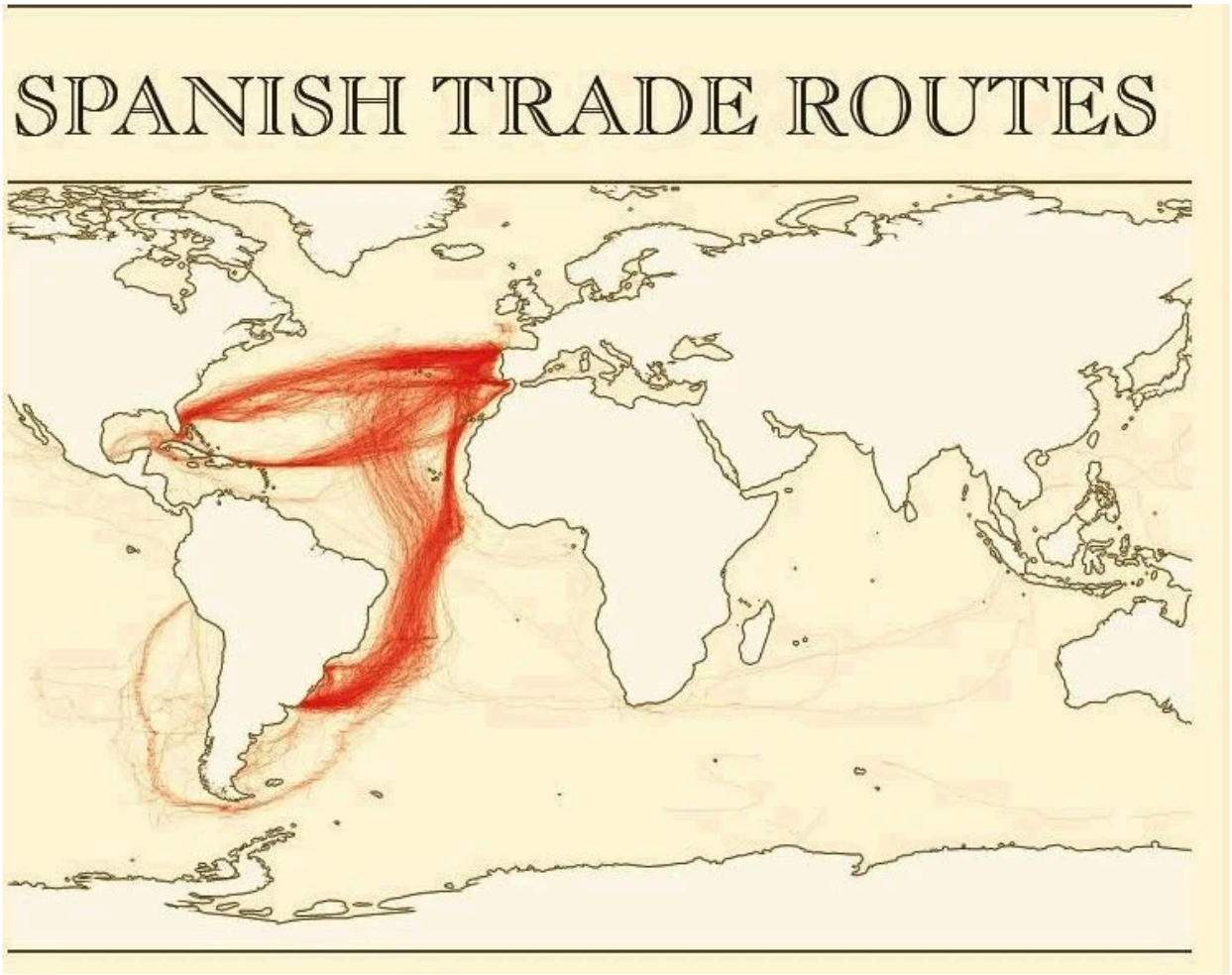
The key issues is that although we know how colonialism and other socio-economics elements have affected type specimen deposition and herbaria relationships, there is currently not much actual analysis to use to back up these claims, the reason for this

is to create a data pipeline that people can use as is or manipulate to help their own research, either to back up proposals or use a stepping stone to deeper regional analysis.

To give an example of how this data is used I'll be doing a small case study looking at the movement of important specimens from areas of high biodiversity and if past colonial relationships affect such movement.

The actual interactive network before any analysis gives us an incredibly clear image of how colonial relationships have been sustained between institutions over however many hundred years just at a glance.





Here we can see the shapes of these networks line up almost perfectly with each other but our common knowledge and understanding just isn't enough to base ideas on. We need to look at this also from the aspect of data analysis. So in this case we will be looking at an area of high biodiversity, Brazil and its relationship to its past colonial oppressors being mainly Spain and Portugal. We'll look at this using an analysis of their nodal betweenness centrality, the movement of the specimens as well as how well supported the actual affected herbaria are.

Brazil(BRA):

- Holds between 15-20% of the entire world's biodiversity and in a year up to 700 new species of angiosperm and animal are discovered. [12]
- 3rd most amount of active herbaria in the world making up 8.68% out of all countries(148)
- Holds largest indigenous run collections in the western hemisphere
- Highest betweenness centrality being 5071 with an edge weight of 180 specimens

Spain (ESP)

- Spain's Biodiversity index as of 2023 being 0.014% in regard to the rest of the world
- 8th most active Herbaria (47) at 2.75%
- Highest betweenness centrality being 227 (path average being 101) with an edge weight at highest of 11 specimens

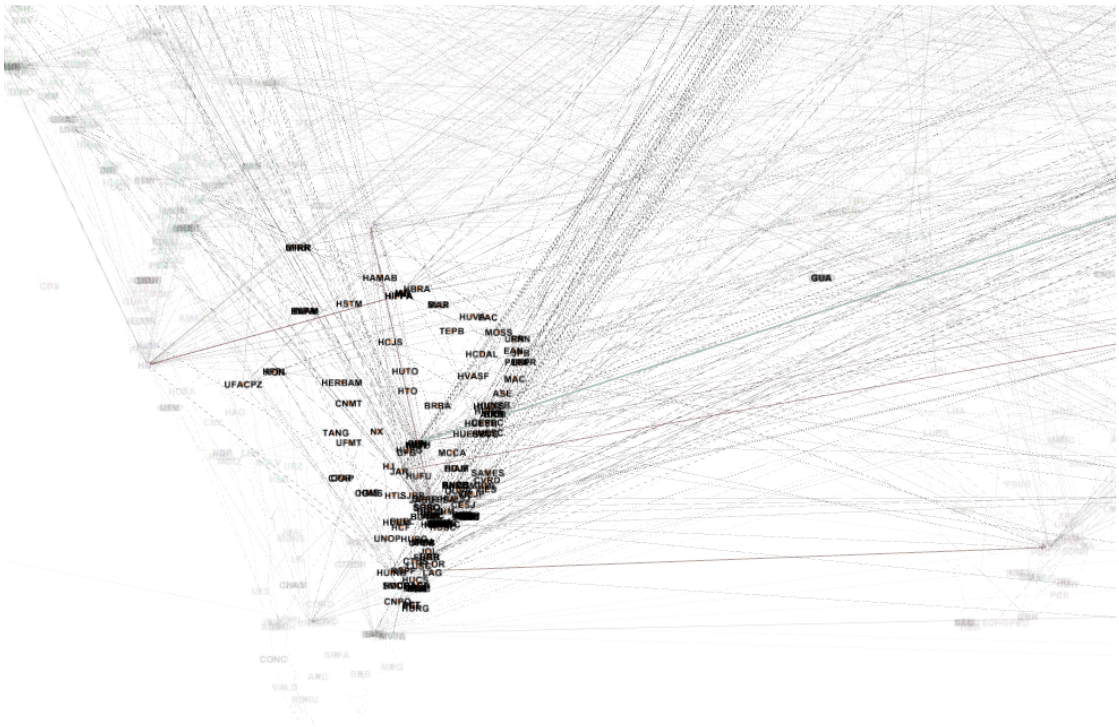
Portugal(PRT)[13]

- Highest biodiversity in Europe at 22%
- 0.89% percent of Global Herbaria (14)
- Highest centrality for Lisbon Herbaria 14,693 at an edge cap of 27 specimens

What does this data tell us?

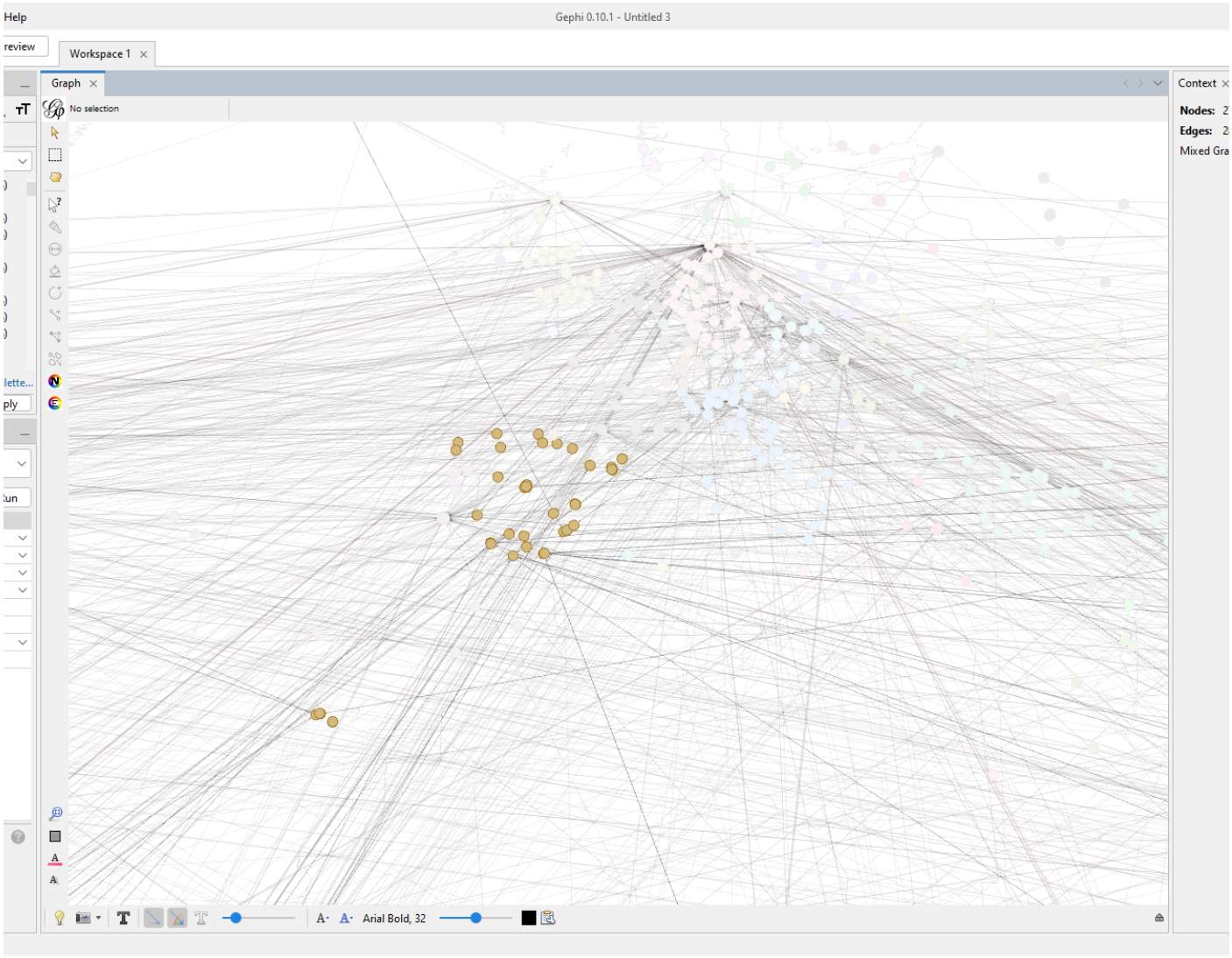
Well from looking at the networks, Brazil has a high number of shared specimens but only within its native polygon. Almost all of Spain and Portugals shared specimens came from non-native polygons and they hold the highest value specimen(holotype) and the native institution holds a much lower value specimen. We can see that

through Brazil's efforts they have been able to ensure that their plant data since 1997 has been kept in its native areas and although yes they share a huge amount with the outside communities and polygons, they have been able to keep much of their own rich data. Below we seen the extent of Brazils network



The data from Portugal illustrates just how incredibly skewed their resources and plant sharing is, to be a small country and have only 14 institutions but to hold almost 3 times a higher centrality score than the most biodiverse country in the world is not a testament to their ability to collect but shows an innate greediness and hoarding of resources compared to other countries.

Spain here lies in the middle. Its institutions are being affected in the last 20 years heavily by the fact it has had multiple conservative governments pulling funding from the science and research sector and although there is a large number of institutions, there is on average an incredibly low centrality score between them due to the shut down or lack of resourcing. Below we see the lack of network in Spain


Help
Gephi 0.10.1 - Untitled 3
Workspace 1
Graph
No selection
Context
Nodes:
Edges:
Mixed Gra
Arial Bold, 32

Results:

Implementation of the Nagoya Protocols:
The implementation of the Nagoya protocol to the command line interface tool (or pipeline tool) did not end up coming to fruition. It was the plan to ensure that it was a major part of this project but time and resource constraints meant that there was no feasible way to get it implemented before the due date of this project. It is in the process of being coded with the original team for this project in the Natural History Museum and Kew alongside my own volunteer work but was unable to be included here in any significant way

Implementation of Command line analysis into the web interface:
The Sigma.js exporter for gephi has a much more limited functionality that previously anticipated, yes there are other online visualisation tools but the point of this project was to ensure that all tools involved were open source and well sustained and after significant market research gephi was the only tool that met such criteria. Having previously met the creators of OUEST studio and working on a few updates with them for their online retina visualiser they have made it clear that there will be expanded functionality added to the Sigma exporter and to OUEST online visualisers in the future but at this stage the updates are coming at only a limited capacity meaning that adding the additional analysis from the command line would not have been doable unless a full dashboard was coded from scratch which did not fit within the timeline once the realisation that sigma could not support the functionality happened.

Creation of usable interface for complex network:
This did end up working out and although there were more things to be added to the interface it is still functioning and interactive to the extent that it can be used by other researchers to understand the relationships between biomes and institutions. The interface will continue to be worked on so that it can get to the stage where it acts as both a visually interesting and interactive tool but also eventually a high level analysis tool looking at the concepts and analyses I had previously mentioned.

Reflection:

The development of this project will continue by myself alongside the teams in the Natural History museum and Kew. This was never going to be a self contained project nor would it make sense for it to be as it began with collaboration and has developed in its parameters due to said collaboration between so many different institutions. The developments that can be made here include the addition of the command line analysis, the transfer of the interactive network to more suitable platforms, this being sigma and storybook deployed on github pages, and the gathering of feedback from other researchers on the wants they have for the development of this tool set.

The coding of this project shows how we can mobilise important data like this in multiple different forms. The reason it needs to be done should be obvious in the need for transparency and access to all forms of knowledge data to work against the barriers created by institutions from old practices of hoarding knowledge and resources. The ideas of this project works alongside the plans many institutions and botanical communities have for developing networks and systems of both knowledge repatriation but also resource sharing between institutions that have been taken advantage of by either old imperial powers or by poor practice.

My hope is that the development of the methodologies for these practices will be helped from what we have been attempting to do over the last two years by way of creating accessible tools and interfaces for large complex data that was previously only able to be analysed on a specimen to specimen basis. What would be an incredibly interesting, yet laborious task, would be to add temporality to this, to look at how the movement of these specimens has changed and developed over time is something that seems on the top set to be an easy thing to do by just adding an additional parameter to the analysis and adding that as an extra attribute to the edges within the network. This would be easy, yet in practice it wouldn't work at all and the best course of action would be to create a network that it separated year by year to look at these

changes starting with a smaller dataset that spans only a decade and then expanding this all into the 256 years of data we have within Kew.

This is a project for another day but the basis of it and the coding tool can be taken from what has been achieved here and with a team working alongside other communities and collaborators it would be a hugely useful creation to analyse our own past and adjust how we go about in our future endeavours as botanists.

Bibliography

Cited :

1. Knowledge for Health | Kew. Accessed January 21, 2024. https://www.kew.org/read-and-watch/medicinal-plants-amazon-yekwana-people
2. About | International Plant Names Index. Accessed September 13, 2022. https://ipni.org/about
3. Global Registry of Scientific Collections. Accessed January 24, 2023. https://www.gbif.org/grscicoll
4. Index Herbariorum - The William & Lynda Steere Herbarium. Accessed June 30, 2023. https://sweetgum.nybg.org/science/ih/
5. Park D. Colonialism has shaped scientific plant collections around the world – here's why that matters. The Conversation. Published June 12, 2023. Accessed December 8, 2023. http://theconversation.com/colonialism-has-shaped-scientific-plant-collections-around-the-world-heres-why-that-matters-207375
6. Regulations: The Nagoya Protocol on access and benefit sharing (ABS). GOV.UK. Published July 1, 2022. Accessed January 30, 2023. https://www.gov.uk/guidance/abs
7. Equity and Access. Accessed December 7, 2023. https://www.rbge.org.uk/about-us/equity-and-access/
8. Roma-Marzio F, Peruzzi L, Bedini G. Personal private herbaria: A valuable but neglected source of floristic data. The case of Italian collections today. *Ital Bot*. 2017;3:7-15. doi:10.3897/italianbotanist.3.12097
9. Flannery M. The Herbarium as Personal. Herbarium World. Published November 18, 2019. Accessed March 10, 2023. https://herbariumworld.wordpress.com/2019/11/18/the-herbarium-as-personal/
10. Gasper AL de, Stehmann JR, Roque N, Bigio NC, Sartori ÂLB, Grittz GS. Brazilian herbaria: an overview. *Acta Bot Bras*. 2020;34:352-359. doi:10.1590/0102-33062019abb0390
11. coffee_network/coffee.py at main · mmcordova/coffee_network. Accessed March 24, 2023. https://github.com/mmcordova/coffee_network/blob/main/coffee.py
12. Unit B. Main Details. Accessed April 26, 2024. https://www.cbd.int/countries/profile?country=br
13. Main Details. Accessed April 26, 2024. https://www.cbd.int/countries/profile?country=pt

Not Cited:

1. Holloway M. Colonial Era Shipping Routes. MoveHub. Published September 24, 2014. Accessed April 26, 2024. https://www.movehub.com/blog/colonial-trade-routes/
2. CITES Wildlife TradeView. CITES Wildlife TradeView. Accessed April 17, 2024. https://tradeview.cites.org/en/country
3. Gephi tutorial. Publishing interactive graphs online. Accessed April 16, 2024. https://blog.miz.space/tutorial/2020/01/05/gephi-tutorial-sigma-js-plugin-publishing-interactive-graph-online/
4. Knowledge for Health | Kew. Accessed January 21, 2024. https://www.kew.org/read-and-watch/medicinal-plants-amazon-yekwana-people
5. Plants for Health | Kew. Accessed January 21, 2024. https://www.kew.org/science/our-science/projects/plants-for-health
6. Liu J, Xiong Q, Shi W, Shi X, Wang K. Evaluating the importance of nodes in complex networks. Physica A: Statistical Mechanics and its Applications. 2016;452:209-219. doi:10.1016/j.physa.2016.02.049
7. Figueira R, Lages F. Museum and Herbarium Collections for Biodiversity Research in Angola.

In: Huntley BJ, Russo V, Lages F, Ferrand N, eds. Biodiversity of Angola: Science & Conservation: A Modern Synthesis. Springer International Publishing; 2019:513-542. doi:10.1007/978-3-030-03083-4_19
8. Gli erbari privati in Italia. Google Docs. Accessed March 7, 2023. https://docs.google.com/spreadsheets/d/1_6KgNo80Kad6Ek0cCxIIABou4zasDM6p_rIj6Rccy_A/edit?usp=embed_facebook
9. Cosmograph: Visualize big networks within seconds. Accessed February 27, 2023. https://cosmograph.app/
10. Iguana S. Large Graph Visualization Tools and Approaches. Medium. Published July 14, 2022. Accessed February 27, 2023. https://towardsdatascience.com/large-graph-visualization-tools-and-approaches-2b8758a1cd59
11. Sigma.js. Accessed November 30, 2022. https://www.sigmajs.org/
12. Nordling L. Seeding an anti-racist culture at Scotland's botanical gardens. Nature. 2022;611(7937):835-838. doi:10.1038/d41586-022-03797-z
13. Flourish | Data Visualisation & Storytelling. Flourish. Accessed September 29, 2022. https://app.flourish.studio/templates
14. International Code of Botanical Nomenclature. Accessed September 28, 2022. https://archive.bgbm.org/iapt/nomenclature/code/saintlouis/0013Ch2Sec2a009.htm
15. NetworkX — NetworkX documentation. Accessed September 28, 2022. https://networkx.org/