Enhance the GBIF clustering algorithms (Topic 3)

Aim/problem/goal

GBIF aggregates biodiversity data from many different sources such as citizen science platforms, specimen data from collections, literature and sequences. In 2020, GBIF developed a clustering algorithm to cluster these different types of data based on scientific name, location, date, etc. The aims of this topic were to explore enhancements to this previous work:

1. To build a common understanding of the process and software currently run by GBIF for detecting links across records and make improvements to this.
2. To explore the DataBricks platform as a tool for analysing and scripting processes that run across large data exports from GBIF.
3. To trial the use of a cloud environment to assess its suitability for collaboration across institutions.
4. To accommodate other research ideas from members in the hackathon.

Method

A databricks cluster was established on Microsoft Azure, using credits kindly donated by the Microsoft Planetary Computer programme to support our effort. An induction programme was run, presenting the DataBricks environment to the members of the group.

The team split into individual and group tracks and explored the following:

1. A data analysis of the EMBL's European Bioinformatics Institute (EMBL-EBI) datasets published in GBIF identified catalog number formatting that was not handled in the clustering algorithm. This was presented to the group as a requirement to address to the group. A fix to the issue identified was coded.
2. The clustered result occurrences were used to compare with the Plazi TreatmentBank datasets (publishingOrgKey = "7ce8aef0-9e92-11dc-8738-b8a03c50a862", as all datasets published by Plazi). Questions were first drafted on the relationship between digitised material citations from literature and their corresponding physical curationship. SQL queries were then issued to the databricks cluster in order to retrieve the answers in tabular form.
3. The clusters formed by the Meise Herbarium records published to GBIF were taken as a test case and further explored. The records published by Meise that clustered were also investigated on a taxonomic and spatial level.
4. Modifying the code to consider the possibility of multiple values being stored in the otherCatalogNumbers field. If multiple values were stored in this field, they should be split, then the individual component parts could be compared across other fields, potentially finding more clustering matches between records. A branch was made to explore this which functioned well on a smaller subsection of records, but encountered performance issues when run on larger quantities of data.

Results

The changes made during the week increased the count of records that link in GBIF from 43.7M to 50.5M with the ENA dataset increasing from 720k to 1.1M. Using the clustered result, as a practical application, we were able to acknowledge that Landcare Research, California Academy of Science and Museum national d'Histoire naturelle are the top three organisations that each holds more than 5000 specimens cited by literature digitised by Plazi. We were also able to further explore the comparison by grouping with the type status and georeferencing

status which enables data quality check between collection metadata and material citations (https://bit.ly/gbif-clustering-plazi). We did not have time to fully analyse the impact and results of this beyond these simple metrics.

Conclusion

The team achieved the goal of a shared understanding of the current implementation operating at GBIF. The algorithm was improved with minor improvements but a mechanism to monitor improvements to the algorithm - or to assess new approaches - is still needed and alternative algorithms should be explored. A key outcome was the confirmation that using a shared cloud environment enables collaboration that otherwise would be difficult to achieve. This environment allows easy exploration of a large dataset, and having access to common, shared cloud computing capabilities with up-to-date exports of GBIF has great potential to enable easy exploration of GBIF data.