UNIVERSITY OF WATERLOO

Faculty of Mathematics

Investigating the Correlation Between Properties of Botanical Specimens

Agriculture and Agri-Foods Canada

Ottawa, Ontario

Prepared by:

David Khazzam

20706704

2A Computer Science

July 16, 2019

Letter of Submittal

David Khazzam
July 16, 2019

Evaluators
WatPD
University of Waterloo
Waterloo, Ontario
N2L 3G1

Dear Evaluators,

This report, “Investigating the Correlation Between Properties of Botanical Specimens”, was created during my 1B work term, while working as a Knowledge Graph Developer in the Bioinformatics Group at Agriculture and Agri-Food Canada.

The purpose is to highlight and go through different strategies for manipulating and visualizing data, as well as mentioning ways to analyze this data. The information used comes from the Flora of North America, a multivolume work describing the flora of Canada and the United States.

Throughout the writing and especially revision process of this report, I have learned of the importance of ensuring that it must be easy to follow; there should be a flow throughout the sections and paragraphs. In addition, I became particularly aware of the importance of an aesthetic paper to aid comprehension.

I created this report, and it has not received any prior academic credit from the University of Waterloo, or any other academic institution. I have proof-read this report, and ensured that it complies with the Math Faculty work report guidelines. I would like to thank Mr. Joel Sachs for guidance with the work that is highlighted in this report.

Sincerely,
David Khazzam, ID# 20706704

Table of Contents

Table of Figures

Executive Summary

This report investigates the process of obtaining botanical data and then analysing it so as to discover trends or other conclusions that can be drawn from the data. It goes through techniques used to parse and organise plant information, and then possible ways to visualise this information so as to facilitate the discovery of trends related to plants.

Challenges faced in this process are explored, such as reducing the occurrence of errors and maintaining the validity of data with ever-changing ontologies.

1 Introduction

There exist over 390,900 plants known to science on earth, with over 13,000 of those occurring in North America (Morelle, 2016). Naturally, each species has dozens of different attributes; root length, corolla colour, anther size, etc. Over the course of decades of field work, much of this data has been written first into volumes of the Flora of North America, of which 20 are currently published, and then digitized to be more accessible to the general public onto the Flora of North America ("Main Page", 2019). This online repository, containing tens of thousands of different species as well as their many thousands of different attributes, stores a large amount of data. Though this information is available to the public, beyond researching each individual plant, it is tedious to analyze. With so much information at one's fingertips, it is enticing to try and use this data to answer questions about plant properties and how certain attributes affect others. However, due to the way the data is stored, this is quite difficult and almost impossible. For this reason, it is essential to try and reorganise all of this information in both an easily accessible and query-able format.

For this reason, the task of reorganising and parsing the original data to transform it into a more manageable format is currently being undertaken. This process will hopefully lead to the opportunity to develop more insights into the evolution and characteristics of plant species, and how to better understand the influence of climate change upon them.

This report will investigate processes being employed to analyze this data, as well as investigating the strengths and weaknesses of these processes. In addition, it will seek to illuminate some of the challenges facing those pursuing this undertaking.

2 Analysis

2.1 Parsing the data

The information about the properties and characteristics of the plants is stored on the Flora on North America in prose form, so it is necessary to parse this text to organise it in a more accessible format. This is done by putting all the data into RDF triples, using scripts written in PERL, Python, and Ruby. This is a format where all the data is in the form subject-predicate-object. For example, to say that the species Cirsium arvense had elliptical leaves, the subject would be Cirsium arvense, the predicate would be blade shape and the object would be elliptical ("Cirsium arvense", 2019). With the data stored in this way, it is much more manageable to query. This is done using SPARQL, a query language specifically designed for data in the form of triples (Hartig, 2017).

If the data is parsed incorrectly or differently than expected, this could have many negative repercussions. Those querying it in the future would perhaps make incorrect assumptions due to them using incomplete or incorrect data. It is therefore essential to attempt to clean the data as well as possible to mitigate potential future problems (Phillips, 2018). For this reason, scripts are being constantly updated and improved to maximize their effectiveness.

2.2 Collecting the desired data

When using SPARQL to query data, it is also imperative to determine precisely what one is looking for, and to use the correct query to obtain it. If sufficient care is not used, it is possible to procure the wrong numbers, which in turn may lead to incorrect assumptions, as mentioned above.

One example being researched is the correlation between the armature of a plant (these are its defense mechanisms) and the colouration of the plant. To ensure that the correct numbers are being collected, one must ensure that all possible armature and coloration properties are being queried, and that there are both no repetitions and no occurrences left out.

In situations like this, it is very important to predetermine exactly the conditions being searched for. For this reason, botanists who are experts in the field should be first consulted so as the correct assumptions are being followed. For example, it should be determined exactly which parts of the flower are deemed relevant to determining the coloration, or what exactly constitutes an armature. (J. Sachs, personal communication, June 27, 2019)

Often it is necessary to store the same data in different formats depending on the question being asked. In this case, instead of modifying the original data, it is advantageous to create new graphs in which to store the data in different triples. This enables the user to query the specific graphs needed, so as to specify exactly what they want. This both leads to more accurate results, as well as a much faster and efficient query as the server is not forced to look through all the graphs. This can be seen in **Figure 1**, where the SPARQL query specifies which graphs to use. The results from this query can be found in **Figure 2**.

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX fna-ontology: <http://www.agr.gc.ca/ontology/>
PREFIX : <http://www.agr.gc.ca/resource/>
PREFIX treat: <http://www.agr.gc.ca/resource/treatment/>

SELECT ?Main_Colour (COUNT (DISTINCT ?s) as ?Species_occ)
WHERE {
  GRAPH <file:///rdf-data/Treatments/V19-20-21_treatments.rdf-30-04-2019-12-46-06> {?s fna-ontology:family treat:Asteraceae}
  MINUS {?child fna-ontology:parent_taxon ?s}

  GRAPH <file:///rdf-data/Treatments/V19-20-21_treatments.rdf-30-04-2019-12-46-06> {?s ?color_prop ?detailed_col.}
  {GRAPH <file://created_subprop2.ttl-28-05-2019-04-57-34> {?color_prop rdfs:subPropertyOf fna-ontology:flower_property;
                                                                       rdfs:subPropertyOf fna-ontology:coloration_property.}}
  UNION
  {GRAPH <file://created_subprop2.ttl-28-05-2019-04-57-34> {?color_prop rdfs:subPropertyOf ?p1;
                                                                       rdfs:subPropertyOf fna-ontology:coloration_property.}
    GRAPH <file://created_subprop2.ttl-28-05-2019-04-57-34> {?p1 rdfs:subPropertyOf fna-ontology:flower_property}}
  UNION
  {GRAPH <file://created_subprop2.ttl-28-05-2019-04-57-34> {?color_prop rdfs:subPropertyOf ?p1;
                                                                       rdfs:subPropertyOf fna-ontology:coloration_property.}
    GRAPH <file://created_subprop2.ttl-28-05-2019-04-57-34> {?p1 rdfs:subPropertyOf ?p2}
    GRAPH <file://created_subprop2.ttl-28-05-2019-04-57-34> {?p2 rdfs:subPropertyOf fna-ontology:flower_property}}
  UNION
  {GRAPH <file://created_subprop2.ttl-28-05-2019-04-57-34> {?color_prop rdfs:subPropertyOf ?p1;
                                                                       rdfs:subPropertyOf fna-ontology:coloration_property.}
    GRAPH <file://created_subprop2.ttl-28-05-2019-04-57-34> {?p1 rdfs:subPropertyOf ?p2}
    GRAPH <file://created_subprop2.ttl-28-05-2019-04-57-34> {?p2 rdfs:subPropertyOf ?p3}
    GRAPH <file://created_subprop2.ttl-28-05-2019-04-57-34> {?p3 rdfs:subPropertyOf fna-ontology:flower_property}}

  GRAPH <file://color_subprop.ttl-05-06-2019-03-59-20> {?detailed_col rdfs:subPropertyOf ?Main_Colour}

  {GRAPH <file:///rdf-data/Treatments/V19-20-21_treatments.rdf-30-04-2019-12-46-06> {?s ?p ?o}
  GRAPH <file://created_subprop3.ttl-30-05-2019-10-55-15> {?p rdfs:subPropertyOf ?p2.}
  GRAPH <file://created_subprop3.ttl-30-05-2019-10-55-15> {?p2 rdfs:subPropertyOf fna-ontology:armature_property.}}
  UNION
  {GRAPH <file:///rdf-data/Treatments/V19-20-21_treatments.rdf-30-04-2019-12-46-06> {?s ?p ?o.}
  GRAPH <file://created_subprop3.ttl-30-05-2019-10-55-15> {?p rdfs:subPropertyOf ?p1.}
  GRAPH <file://created_subprop3.ttl-30-05-2019-10-55-15> {?p1 rdfs:subPropertyOf ?p2.}
  GRAPH <file://created_subprop3.ttl-30-05-2019-10-55-15> {?p2 rdfs:subPropertyOf fna-ontology:armature_property.}}
}
GROUP BY ?Main_Colour
ORDER BY DESC(?Main_Colour)
```

Figure 1 A Sample SPARQL query, querying for the number of occurrences of different flower colours in species in the Asteraceae family that possess armature

Main_Colour	Species_occ
http://www.agr.gc.ca/resource/yellow	476 *(xsd:integer)*
http://www.agr.gc.ca/resource/white	436 *(xsd:integer)*
http://www.agr.gc.ca/resource/red	133 *(xsd:integer)*
http://www.agr.gc.ca/resource/purple	317 *(xsd:integer)*
http://www.agr.gc.ca/resource/pink	163 *(xsd:integer)*
http://www.agr.gc.ca/resource/orange	38 *(xsd:integer)*
http://www.agr.gc.ca/resource/green	56 *(xsd:integer)*
http://www.agr.gc.ca/resource/gray	63 *(xsd:integer)*
http://www.agr.gc.ca/resource/brown	324 *(xsd:integer)*
http://www.agr.gc.ca/resource/blue	62 *(xsd:integer)*
http://www.agr.gc.ca/resource/black	47 *(xsd:integer)*
http://www.agr.gc.ca/resource/beige	312 *(xsd:integer)*

Figure 2 Results of the query from Figure 1

2.3 Visualising the data

Once the data has been queried and collected, only part of the job is done. It is then essential to use data visualization tools to properly illustrate the findings. This serves the dual purpose of making the results much more understandable, as well as making it much easier to draw conclusions from. By using tools such as D3 or Tableau, conclusions are drawn much more clearly. Though D3 is more difficult to learn than Tableau, it allows the developer much more freedom to create a visualization exactly according to their preference. For example, **Figure 3** illustrates occurrences of different types of armatures in species with respect to the flower colour of those species. This information may be difficult to interpret. By using D3 to visualize it as in **Figure 4**, it is much easier to notice which colours are more prominent, as well as determining which armatures are more likely to exhibit certain colours.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Armature	no_info	coloured	not_coloured	yellow	white	red	purple	pink	orange	green	gray	brown	blue	black	beige
2	no_info	100	52.73	47.27	26.901	23.652	8.04	19.196	9.682	2.178	4.154	2.714	15.712	4.891	3.685	12.161
3	armed	100	74.163	25.837	40.858	37.682	12.017	27.983	13.991	3.262	6.352	5.751	28.755	5.322	4.635	27.124
4	unarmed	100	39.011	60.989	17.967	14.67	5.495	13.571	6.923	1.484	2.747	0.769	7.363	4.615	3.077	2.582
5	barb	100	50	50	0	0	0	0	0	0	0	0	0	0	0	50
6	bristle	100	77.391	22.609	46.667	44.348	8.116	26.667	12.464	4.638	4.348	5.217	24.638	3.478	4.058	36.522
7	bristle-spine	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	ciliate	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	cilium	100	23.077	76.923	0	23.077	0	0	0	0	0	0	0	0	0	0
10	coma	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	comal	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	fibril	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	glochid	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	hair	100	74.324	25.676	39.865	35.135	16.216	26.351	15.541	3.378	4.054	2.027	29.054	4.73	4.73	9.459
15	hair-tuft	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	indument	100	30.435	69.565	13.043	21.739	0	0	21.739	0	0	0	0	0	0	8.696
17	indumentum	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	papilla	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	papillae	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	pappus	100	77.989	22.011	46.349	42.011	12.063	26.349	12.804	3.598	6.984	6.772	29.947	5.608	4.974	31.958
21	paraphyses	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	prickle	100	0	100	0	0	0	0	0	0	0	0	0	0	0	0
23	projection	100	100	0	100	0	0	100	0	0	100	0	0	0	0	0
24	seta	100	100	0	33.333	61.905	19.048	42.857	28.571	4.762	19.048	0	23.81	28.571	19.048	23.81
25	spine	100	75.214	24.786	13.675	35.043	11.111	54.701	30.769	1.709	4.274	3.419	45.299	2.564	3.419	17.949
26	spinule	100	100	0	50	0	50	100	50	0	0	0	0	0	0	0
27	thorn	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
28	tomentum	100	20	80	0	20	0	0	0	0	0	0	20	0	0	20
29	trichome	100	91.071	8.929	5.357	32.143	19.643	69.643	37.5	0	1.786	0	46.429	0	1.786	12.5
30	vestiture	100	0	100	0	0	0	0	0	0	0	0	0	0	0	0
31	vesture	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
32	wool	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 3 Percentage occurrences of species in the Asteraceae family with different flower colorations and armature types

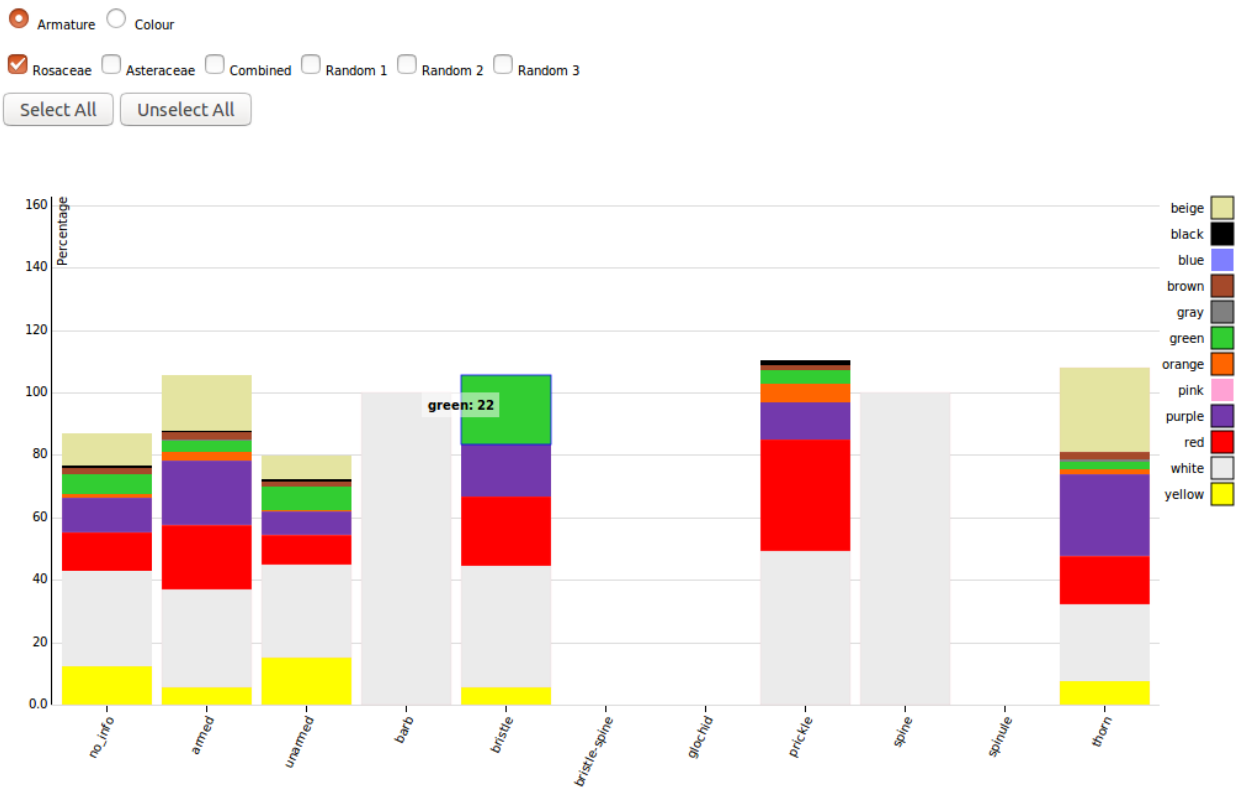

Figure 4 Interactive graphic created using D3 of the data from Figure 3

2.4 Maintaining Accuracy

A very significant issue when dealing with large amounts of data is ensuring both the validity and accuracy of the information. In the case of botanical data, both of these are important issues that must be thought about and dealt with.

First of all, ensuring the correctness of the data must be done as effectively as possible. There are many places where errors can be introduced, from field-work collection, transcribing, digitizing and parsing, among others. By taking care to execute these steps carefully and properly, problems that may occur in the future can be avoided. Even so, it should be assumed that there will always be some margin of error with the results that should be taken into consideration.

Additionally, the validity of the data must be taken into account. Botany is a unique field, as its terms and definitions are constantly changing. Especially with respect to botanical ontology, there is a constant need for change (Flouris, Manakanatas, Kondylakis, Plexousakis, & Antoniou, 2008). It takes years to document previously and newly discovered plants, and as the ontology of species is in constant flux, often plants are defined differently at different times. In addition, changes in ontology are not always sufficiently broadcasted, causing a mismatch between definitions used by different collectors. Finally, as individual occurrences of species may vary, specific attributes associated with different species may also vary depending on the collector. All these factors serve to further clutter the data, causing variance between different data sets.

It is thus imperative to maintain flexibility and be able to modify data with respect to current standards. This is also a reason as to why there are currently world-wide efforts to combine the silos of botanical data that belong to different organisations, so that it can all be located in one repository. One such project is called DINA, and developers around the world are in the process of assembling, managing and sharing data associated with natural history collections and their curation ("Welcome to the DINA project", 2019). Though a daunting task, ultimately it will save immeasurable time and effort, as well as provide researchers an unparalleled repository of knowledge.

3 Conclusions

Based on the analysis, the following conclusions can be inferred.

Firstly, when working with large amounts of data, there are always going to be errors. All steps should be taken to attempt to reduce the impact they may have upon the final results, but they will always occur.

Secondly, different forms of visualisation are necessary to properly understand the implications of the data. These different visualisations will help highlight different relevant messages and suggest possible interpretations. They may also suggest future questions to ask and routes to investigate.

Finally, continual efforts must be made to ensure data is kept up to date and accurate. Due to the constant evolution in the botanical field, it must be ensured that repositories containing information about different species must be continually revised.

4 Recommendations

Further investigations may look into other possible correlations between different plant properties. For example, the association between plant sexuality and whether fruits are fleshy or dry may be investigated, or looking to see if there is a correlation between chromosome count and whether a plant is monoecious versus dioecious.

In addition, further exploration may be done with association rules and investigating the conditional probability of different events occurring. This more mathematical-based investigation may yield more interesting quantifiable results.

Finally, the research could be done of the uses and potential strengths of employing knowledge graphs when working with biological data. As their queryable properties enables developers more freedom when manipulating the data, it might be interesting to compare the pros and cons of either cleaning the data before putting it into a graph, or querying afterwards (Page, 2019).

References

Cirsium arvense. (2019, January 2). Retrieved July 2, 2019, from http://beta.floranorthamerica.org/wiki/Cirsium_arvense

Flouris, G., Manakanatas, D., Kondylakis, H., Plexousakis, D., & Antoniou, G. (2008). Ontology change: Classification and survey. *The Knowledge Engineering Review*,*23*(2), 117-152. doi:10.1017/s0269888908001367

Hartig, O. (2017, June). RDF* and SPARQL*: An Alternative Approach to Annotate Statements in RDF. Retrieved July 2, 2019, from http://olafhartig.de/files/Hartig_ISWC2017_RDFStarPosterPaper.pdf

Main Page. (2019, February 15). Retrieved June 18, 2019, from http://beta.floranorthamerica.org/wiki/Main_Page

Morelle, R. (2016, May 10). Kew report makes new tally for number of world's plants. Retrieved June 18, 2019, from https://www.bbc.com/news/science-environment-36230858

Page, R. D. (2019). Ozymandias: A biodiversity knowledge graph. *Peerj*. Retrieved June 18, 2019, from https://peerj.com/articles/6739/.

Phillips, J. Z. (2018, May 16). Trusting Your Data: Clean, Verified, and Maintained. Retrieved June 18, 2019, from http://www.themeasurementstandard.com/2018/05/data-clean-verified-maintained/

Sachs, J. (2019, June 27). Talking with Joel Sachs [Personal interview].

Welcome to the DINA project. (2019, June 26). Retrieved July 16, 2019, from https://www.dina-project.net/wiki/Welcome_to_the_DINA_project

Acknowledgments

I would like to thank Joel Sachs for his explanations and thoughts on different botanical questions to ask and investigate.