Position Paper for ALS 13

Towards/Developing a legume systematics portal

Outline

Introduction

Discussion

- Legume Online Resources
- Online resources available for plant systematics and evolution
- Taxon-based portals, examples from other taxa
- Lessons learned from past efforts
- Resources and information that should be integrated in a legume portal
- Effective pipeline for developing a legume portal (Ways forward?)
- Uses of Web Portals and Future Applications

Concluding Remarks

Table 1 - List of selected currently available resources and portals (name, url, brief description)

Fig. 1 - Legume Systematics Portal: Front end showing visual aspects of portal and search entry point

Fig. 2 - Legume Systematics Portal: Links to external data sources and portals. Green boxes are dynamically updated (with the latest updated version visible), blue boxes are updated periodically (how often?). Arrows represent APIs and Webservices

Fig. 3 - Legume Systematics Portal: Components, tools and functions (Back end showing informatics tools and structure required/suggested)

Abstract

Objectives/Aims of paper

State of the situation

What is needed and how to move forward

Introduction

- What will be talked about in this paper - past, existing and what needs to happen
- Why a taxon-focused portal is still relevant and necessary in the age of massive global databases (ex, all plants of the world, all images, all specimen data, all genetic sequence data…).

- Why a legume portal is important. Whom is it for?
- Legumes as an exemplar group of plants.
- How information can be used to ameliorate legume research.

The need for scientists to exchange and share data has resulted in a marked increase in the number of research data portals in the past decades (ref). These data portals have varying objectives and functions but generally are considered informatic “systems that provide remote access to data repositories for discovery and distribution of reference data, the upload of new data for analysis and/or integration, and data sharing for collaborative analysis” (Chard et al. 2018). This proliferation of data portals is no less true in the field of systematics and taxonomy, where researchers have adopted informatics to make data publicly available (e.g. O'Leary and Kaufman 2011; Benson et al. 2012; Goswami 2015). Wheeler et al. (2004) have suggested that cyber infrastructure could revitalise taxonomy, by allowing quick access to bibliographic information, biological data, and specimen information.

This is illustrated clearly by the continued development and maintenance of portals dedicated to taxonomic literature (BHL), nomenclature and synonymy (Catalogue of Life, IPNI, Tropicos, ITIS), georeferenced specimen data (GBIF, Atlas of Living Australia, BIEN, iDigBio, VertNet), taxon lists and hierarchies, phylogenies, etc. (see Table 1).

From the late 1990s onwards, numerous taxonomic communities developed taxon-based portals that grouped the information about a particular taxon into readily accessible web pages and databases (ants, spiders, plants, etc.) (Neotropikey, Scratchpads e.g. Solanum Source, ref?). There has also been a trend, particularly evident in the past decade, to group species data into massive international projects that focus on a particular data type and which in many ways can present through taxon-specific searches information specific to a taxon. The question then becomes, whether a taxon-based portal remains a pertinent (and viable) option for data sharing, storage, aggregation and analysis. Here we argue that a community of taxonomists and systematists focused on a particular taxonomic group have a vested interest, a strong allegiance and the knowledge necessary to ameliorate data quality and actively contribute to broader more general data providers, and thus that taxon-based portals continue to serve a useful function that has not been made obsolete by large international data-focused portals.

Additionally now that many family and order level taxonomies have been reworked to be equivalent to monophyletic lineages, these portals can represent snapshots of evolutionary time and allied data such as distributions, morphological traits and phenology are robust to quantitative statistical analyses.

Legume systematists have been at the forefront of using informatic interfaces for sharing of data, and particularly of taxonomic information, with the development in 1985 of the International Legume Data Information System (ILDIS). However, due to lack of continuity of resources, institutional support and community involvement, as well as evolving software protocols, the need for distributed data curation and increased data complexity, ILDIS has ceased to be curated by the community for more than ten years. Nevertheless ILDIS's taxonomic data are now deployed through the Catalogue of Life. The entire ILDIS database (with use, common names, etc) can be searched via http://www.legumes-online.net/ildis/aweb/database.htm as well as at the original location https://www.ildis.org. These other classes of ILDIS data have been incorporated into Plants of the World Online having been assigned to a more modern consistent taxonomy. This example points to the synergy of in taxon specific community projects that inform higher level international initiatives and also benefit from the data aggregation and standardization provided by them.

After several years of discussion, the legume systematics community has expressed interest to develop a new portal dedicated specifically to deployment of resources pertinent to researchers in this field. In 2010, the Legume Phylogenetics Working Group was founded, and it has since published three papers under this name, including a new subfamily classification for the family. This research community has been collaborating and sharing data for several decades, since the first International Legume Conference in 1978 held at The Royal Botanical Gardens, Kew (Kew) and organised by Roger Polhill and colleagues, but the exchange of information often occurs casually...xxxxx. Bean Bag, LoWo, ALS volumes (This contrasts remarkably with some other research communities where sharing of data is considered detrimental to a researcher or at least more problematic, e.g., Tenopir et al. 2011). See Heaton et al. (2018) about new ways that scientists work and share data.)

In order to facilitate data sharing, knowledge of what is available, what is missing and what is erroneous, to facilitate cross analyses and increase collaboration within the legume systematics community and with other research communities, we explore here the best practice for developing a legume focused portal. Our objective is to take full advantage of existing data sources, informatic tools and protocols to develop an easily manageable, scalable and interactive portal that will be used and contributed to by the legume systematics community. Below we discuss existing resources pertinent to our needs, explore taxon-based portals for other taxa to determine which features are most useful and of interest to our community and which appear problematic to their long term sustainability and usefulness, and discuss lessons learned from past projects. We then propose a vision and a pragmatic way forward for the development of a Leguminosae systematics portal.

2. History of Leguminosae Portals / Past initiatives managing Legume data

Legumes are special

The size, economic importance, diversity and biological importance of the Fabaceae has meant that there have always been comparably intense studies of the family by both systematists and other biologists. The Advances in Legume Systematics Series contributed to the creation of an active global community of legume systematists. These factors meant that legumes were a focus for early initiatives to develop and deploy information technologies for capturing, sharing and disseminating information about plants. Such efforts often led the way in developing methods and thinking with regard to taxonomic data management more generally.

Early days

In the late 1980's the Vicieae Database Project, led by Frank Bisby and Richard White in Southampton, UK built a monographic information resource for the c. 600 species of the tribe Vicieae [Adey et al. 1984]. The project was established to prototype taxonomic information management explicitly seeking to test use of the software and technologies available at that time: to demonstrate what was possible and where development was necessary. Software limitations required creation of separate, albeit interlinked, database modules. These included a "taxonomic nomenclatural backbone", "geographical distribution", "morphological descriptions", "chemical constituents" and "bibliography". The choice of software for each module essentially

depended upon the nature of the data being stored. Software tools were developed within the project to link the modules and fill in obvious gaps. A herbarium, seed and living plant collection were created alongside the digital resources to assist data collection. All data entry and management were undertaken by a small team in one institution and, in a pre-Internet era, project outputs were numerous printed catalogues published by the University, identification keys generated using DELTA software [Dallwitz, 1993] and a series of research papers describing and discussing the methodologies [Allkin 1984; White 1984].

Success of the Vicieae project led subsequently to two parallel developments. The first involved attempts to use data modelling for management of taxonomic monographic data for the first time [Allkin and White 1988; White et al. 1993] and to build software capable of managing species checklists and descriptive data for the first time that could interact with specimen collection management and DELTA identification software. These initiatives led to creation of several of the first taxonomic data standards adopted by the then newly formed Taxonomic Databases Working group (e.g. Allkin and White 1993).

ILDIS

At the same time the Vicieae project led to a more ambitious initiative to build a comprehensive checklist of the entire Legume family involving systematist from many different countries with taxonomic expertise in the family. The International Legume Database and Information Service (ILDIS) was to be built in phases [Zarucchi et al. 1993]. Phase 1 involved construction of a taxonomic checklist including full synonymy, geographic distribution and a very modest set of habit features. Phase 2 sought to add depth to that taxonomic framework in several dimensions: morphological, chemical etc. Lack of funding mostly prevented progress with Phase 2, other than definition of the data standards to be used. There were however notable exceptions such as inclusion of two different sets of root nodulation data [Refs]. Independent, albeit collaborative, efforts had led to significant publication of accounts of the seeds and fruit morphology of Legumes [Gunn ref] and these were envisaged to form part of a final information product.

Implementation of the ILDIS database required software to manage the core checklist and capable of extension to handle sophisticated data types in Phase 2. It required well-defined data standards and terminological controls and software capable of enforcing these. The widely

spread network of collaborators and contributors also necessitated data exchange mechanisms (export, import and integration) to enable data sets to be shared between collaborating centres. These in turn led to development of further data standards and exchange protocols [refs] of wider utility to botanical data management. A strength of ILDIS was that each and every observation relating to a plant or plant name was linked to the data source(s) from which that data was obtained facilitating acknowledgement and analysis.

ILDIS needed to give particular effort and attention to integrating datasets with different scopes (including those from partners with either a geographic or taxonomic focus) and to establishing effective mechanisms for merging data sets with overlapping scope (and thus potentially conflicting opinions). It is no coincidence that the vast bulk of the original ILDIS species checklist was built through a series of non-overlapping regional checklists compiled from the literature by Mike Lock working at and supported by a large institution, namely the Royal Botanic Gardens, Kew [e.g. Lock 1989) Each publication was compiled using Alice software and ILDIS data standards with the expectation that the data sets would be merged into the central ILDIS database which subsequently, they were.

Defining and agreeing data standards and protocols amongst partners proved to be an administrative overhead for ILDIS but did ensure the involvement and engagement of all parties. This did lead to development of further standards under the Taxonomic Databases Working Group Umbrella (e.g. Holllis and Brummitt XX?) The frequent exchange and integration of data sets between the coordinating centre and partners required significant data administration and support effort centrally.

Sustaining ILDIS within a university department required a continual search for grant funding which ultimately proved impossible as research funding agencies prioritised innovation and ever more challenging targets rather than completion of the task in hand. Accordingly, the ILDIS coordinators became focussed on wide challenges of creating a checklist of all plants and ultimately all biological species. Species 2000 (ref) and its product the Catalogue of Life (ref) were the result. The focus shifted to the discovery and stimulation of further “global species databases” analogous to ILDIS, with the consequence that the ILDIS database received reduced attention, curation or input from partners. Without further funding centrally efforts to complete the species checklist (phase 1) or particularly to develop more ambitious Phase 2

modules containing richer data became more disparate despite some efforts from staff at Kew, Missouri Botanical Gardens and Reading University.

LOWO

The "Advances in Legumes Systematics" series presented the output from major conferences arranged periodically (at 10 year intervals????) to review the state of legume systematics. Ultimately these led to a different kind of publication which presented in a single volume an illustrated account of each Genus of Legume summarising our knowledge of the Genus and its position within the family classification. Legumes of the World (LoW Lewis et al 2005) was edited at Kew though the content for each Chapter was provided by one or more experts in that Tribe. 24 legume systematists contributed to this landmark publication.

Since LoW was published using traditional publishing techniques, any future update would effectively require a new book written from scratch. Nevertheless, the structure of the book and the consistent format of each Genus account lent itself to digitisation to permit subsequent reuse and manipulation. A studentship (initially) and modest subsequent investment at Kew saw the content of LoW imported into Kew's Content Management System and made available via Kew's website as "Legumes of the World Online" (LOWO). This digital version had a number of significant benefits over the book. In addition to permitting consultation by many more people via the internet, LOWO also offered the ability for readers to browse the data in ways not possible with the original book. The book included indexes of names and was broken into chapters, each dealing with a single Tribe, much as in previous publications. For LOWO, however, the phylogenetic trees included as diagrams in the original book were also digitised and users might browse genera using either the conventional taxonomic structure of subfamily and tribe or using the phylogenetic tree represented in those diagrams. A further advantage was that online publication removed constraints on the numbers of images that might be included for any Genus. The most significant advance however was that in addition to "publishing" the data online, LOWO also provided a means for the Kew legume team to curate and extend the content: adding new images (including photographer acknowledgements) and withholding online view of photographs for which permission for use had yet to be given. New Genera were added, species lists refined, corrections made to data values in many Tribes and in some cases completely new revisions were undertaken. This work proceeded slowly envisaging ultimately formal publication as a second Edition with input from the rest of the community.

LOWO adopted and adapted existing software: the Content Management System employed by Kew. This reduced development costs considerably but subsequently required several migrations as Kew opted to change the CMS used for its institutional website. Each release brought enhancements and had the significant benefit that in principle, at least, legume systematists from around the world could be given access to LOWO so as to edit remotely those "chapters" or taxa for which they were responsible. There was no need for a constant data exchange. Editing remotely was trialled by staff working at RBGE who were able to make edits to LOWO but the feature had not been fully taken advantage of when Kew replaced their CMS for the fourth time with a more recent version of Drupal. An outstanding task, therefore, is to upgrade the Drupal code used in LOWO to match the Version of Drupal now used by Kew.

Mapping the ILDIS species checklist to LOWO had been problematic given that the former failed to reflect current taxonomy. Inclusion of Kew's current legume species checklist had been envisaged, however, and now that is complete could be linked directly to LOWO.

Current Leguminosae Portals

Kew's World Checklist - since this covers all plant species I guess that it doesnt fit here - though it is referred to in proceeding

ILDON – Nodulation data on legumes

ILDON, The International Legume Database of Nodulation, is a current initiative which builds upon one of the original ILDIS Phase 2 Modules for root nodulation data, ILDON intends to create and maintain an authoritative online repository of peer group reviewed records of legume nodulation and to derive and make available digital publications from this repository.

The nomenclatural and taxonomic backbone for the ILDON repository will be that already maintained by RBG Kew. The nodulation records will come from the published literature and from authenticated records collected by scientists in the field. ILDON looks to improve its value and utility in two ways: by progressively developing the detail of the nodulation records in

collaboration with scientists working on nodulation and by working with other legume scientists to ensure the ILDON data can be linked to and kept in sync with other online digital legume data resources.

LIS - Legume Information System

The Legume Information System (LIS), developed by the National Center for Genome Resources (NCGR) in cooperation with the USDA Agricultural Research Service (ARS), integrate genetic, genomic, and trait data across legume species enabling cross-species genomic and transcript comparisons and facilitating crop improvement (Gonzales et al 2005, Dash et al 2016). At the time of this writing, users have access to browse 15 genomes of legumes (https://legumeinfo.org).

Other Legume Portals / Resources that maybe could be mentioned somewhere

- Leucaena Specimen Database (https://herbaria.plants.ox.ac.uk/bol/leucaena) , BRAHMS Database
- Acacia database (worldwidewattle.com curated by Bruce Maslin) The WorldWideWattle (WorldWideWattle ver. 2. Published on the Internet at: www.worldwidewattle.com [Accessed*]) is a collaborative project with a goal of providing authoritative information on the species of *Acacia* sens. lat. It is the brainchild and lifework of Bruce Maslin of the Western Australian Herbarium within the Department of Biodiversity, Conservation and Attractions with collaboration from the Western Australian Shire of Dalwallinu. The WorldWideWattle delivers authoritative information relating to taxonomy, distributions, traits and includes images, line drawings, a detailed information gallery, and links to the Flora of Australia. The portal also includes access to Wattle, the Lucid based interactive key for the Australian *Acacia* taxa. The portal is a static information source that is not integrated into international databases, but it contains the type of data desired in all biodiversity portals.
-
- Legume Federation (https://www.legumefederation.org/), PeanutBase (https://peanutbase.org/), KnowPulse (http://knowpulse2.usask.ca/portal/), Cool Season Food Legume Crop Database Resources (https://www.coolseasonfoodlegume.org/)

- LegumeIP (http://plantgrn.noble.org/LegumeIP/),an integrative database for comparative genomics and transcriptomics of model legumes, for studying gene function and genome evolution in legumes (Li et al 2012)
- Chemical data - proposal by Geoff Kite (Kew) to develop a chemical database for legumes
- Cultivated and economically important legumes
- Phytozome (https://phytozome.jgi.doe.gov), the plant comparative genomics portal for accessing, visualizing and analyzing plant genomes and datasets (Goodstein et al 2011)
- Phylogenomic Database for Plant Comparative Genomics (GreenPhylDB): A comparative and functional genomics in plants including gene families based on gene predictions of genomes, covering a broad taxonomy of green plants, including six legume species (Conte et al 2008; http://www.greenphyl.org/cgi-bin/index.cgi).

3. Web resources presently available for (plant?) systematics and evolution

Over the past three decades numerous online resources have been developed that facilitate the work of taxonomists and systematists. These include taxonomic and nomenclatural resources and databases, geospatial portal based on data from collections and observations, databases focused on traits, morphology, chemistry?, DNA sequences, full or partial genomes, molecular resources, data repositories and phylogenetic assemblers (term) (Table 1). In practice systematists often will consult multiple resources to access the desired information.

A. Taxonomic Resources

Taxonomic and nomenclatural resources are of two main types: curated original datasets and aggregators of multiple sources. Both types tend to be under the hospices of large natural history museums or of international consortia of museums, and respond to particular and sometimes specific needs (lists of plant names, synonyms, reference literature, basic descriptive information). Some of these important resources have not been updated in terms of recent nomenclatural changes and decisions and some are informatically out of date. For this reason, researchers and curators will tend to consult several different taxonomic and nomenclatural resources to arrive at nomenclatural decisions and gather information. Here we provide a short

description of the main resources that are presently available. The first four are curated datasets, where as the The Plant List, the Catalogue of Life and the GBIF taxonomic backbone are taxonomic aggregators.

The International Plant Names Index (IPNI: http://www.ipni.org/) is a catalogue of, and default reference, for scientific names of ferns and angiosperms. It provides information on authorship of names (including standard abbreviations) and their place of publication, but does not give information on what are the currently accepted names or synonyms. Increasingly journals (e.g. Phytokeys and Kew Bulletin) are automating addition of new names to IPNI. IPNI has external links such as to the Biodiversity Heritage Library https://www.biodiversitylibrary.org/ and a new beta version of IPNI (https://beta.ipni.org/) makes these more explicitly to the protologues (available in online articles). IPNI is being actively curated by Kew Gardens, Harvard University and the Australian National Herbarium (In 2018, 7342 names were published, 18864 records were added and 33547 names were updated). It is the modern descendant of Index Kewensis that was active from 1893 to 2000 and of the Gray Card Index, which was computerised in 1992.

The World Checklist of Selected Plant Families (WCSP: https://wcsp.science.kew.org/home.do) is compiled by Rafaël Govaerts at Kew. It contains the accepted names, full synonymy and distribution derived taken from published literature. It also presents information on alternative taxonomic classifications where these are known. Importantly WCSP is actively curated with c. 10,000 changes to the status of plant names being published every year. New plant names are derived directly from IPNI and, in 2018 a far higher rate of 12,000 new names were published. Once a family account (synonymy and distribution data) has been compiled they enter a peer review process in which systematists with expertise in particular genera or groups of genera invited to review and correct before the family account is published online. The species checklist for Leguminosae is complete for 98% of all legume species (October 2018) and is now entering the review phase with various specialists contacted. The Legume checklist has not therefore been published on the WCSP website yet but is visible through Plants of the World Online .

Plants of the World Online (http://www.plantsoftheworldonline.org/) is a taxon-based portal that uses Kew's core taxonomy (pulling data from WCSP and IPNI) and publishes additional information provided by Kew and its partners, such as descriptions (from floras or monographs),

static maps and links to other databases. The legume checklist in POWO derives from WCSP. POWO includes most of the common names, use, habit and other descriptive information derived from ILDIS and from Legumes of the World Online (see below, Part 3).

Tropicos (http://www.tropicos.org/Home.aspx) from the Missouri Botanical Garden is a rich source of botanical information particularly for the New World. It encompasses information from MBG's collections and various published floras. Each flora project includes synonymy, distribution, specimens and images (when available) and thus Tropicos offers the user conflicting taxonomies for some groups of plants. Initially created for internal uses, Tropicos was made publically available online in year? and now contains over 1.3 million names and references 4.6 million specimens.

In addition to the curated databases listed here, many countries or geographic regions, as well as communities of taxonomic specialists (see below), have developed and maintain online species lists with synonyms, vernacular names, maps, images, etc. (e.g., Australian Plant Census (https://www.anbg.gov.au/chah/apc/), VasCan (http://data.canadensys.net/vascan/search?lang=en)).

The Plant List (TPL: http://www.theplantlist.org/) published in 2012 as a "working list of plant species" as a response to Target 1 of the Global Strategy for Plant Conservation provides a synonymised checklist of all plants, ferns and mosses using APGIII. The data used in TPL are aggregated from WCSP (published and unpublished), Tropicos and monographic data sets for Compositae, Rosaceae and Legumes. Automated mechanisms were employed to detect conflicting taxonomic views between and within overlapping data sets and conflicts were resolved using a logical rule set seeking to implement standard botanical decision making. Inevitably this led to some errors. TPL provided for search, browse and data download functions. Taxonomic decisions were labelled with one, two or three stars depending on the relative confidence or reliability of that taxonomic judgement. TPL has become the most popular plant checklist (1.7 million unique users / year) despite its underlying data. TPL always suffered from data errors and gaps inherited from its contributors and introduced further errors when merging multiple plant lists automatically. Sadlly TPL has not been updated since 2012. It is, for example, linked to the R packages such 'Taxonstand' (Cayuela et al. 2012), which allows

automated standardization of taxonomic names in bioinformatics pipelines using The Plant List nomenclature.

Catalogue of Life (http://www.catalogueoflife.org/) is an aggregator site bringing together (mostly non-overlapping) checklist data sets from diverse sources for all living organisms. The Catalogue of Life supports several biodiversity and conservation information services such as GBIF, the Encylopedia of Life and the IUCN Red List of threatened species. COL provides web services for querying their database. As for TPL, the legume data in COL derives from ILDIS and is not being actively curated.

GBIF also is now acting as a taxonomic aggregator. Although its taxonomic backbone is mostly based on COL, increasingly GBIF is using other sources of information

B. Biodiversity Portals, Specimen Information and Occurrence Data

Biological collections around the world are the host of impressive biological information, and importance of this type of collection is acknowledge in the scientific community (Meineke et al. 2018). Getting access to the rich historical data curated in herbaria, entomological collections, natural history collections, etc, traditionally meant visiting the collections of interest or asking for a loan of a specific set of specimens. With the development of databasing, imagery and Internet, it quickly became apparent that access to specimens could be virtual. Institutional portals arose in the early 2000's, giving access to some collections. However, the need to aggregate data, in order to efficiently discover and analyse them from a single entry point, led to the development of data aggregators and standards of publication. The Global Biodiversity Information Facility (GBIF) is nowadays the most important data aggregator, harvesting data from a network of national aggregators. This type of aggregation would not be efficient without standardized publication protocols. In 2009, the Biodiversity Information Standard, formerly the Taxonomic Database Working Group (TDWG), developed the Darwin Core vocabulary, a standard to publish biodiversity data (https://www.tdwg.org/). With the recent development of citizen science, several initiatives have contributed to the publication of millions of observation data points that are now available on GBIF (e.g., eBirds).

As the governance structure of GBIF developed in the late 1990s and 2000s, several countries developed biodiversity portals and platforms that aggregated data at a national level, acted as a

biodiversity information resource for the country and became the central point from which GBIF could harvest data (e.g., GBIF France (http://www.gbif.fr/), GBIF Spain (http://www.gbif.es/en), Brazil Species Link (http://www.splink.org.br/), Brazil Biodiversity Information System (http://www.sibbr.gov.br/), Colombia Biodiversity Information System (https://sibcolombia.net/), Canadensys (www.canadensys.net), NBN Atlas Scotland (https://scotland.nbnatlas.org/), Belgian Biodiversity Platform (http://www.biodiversity.be/), Biodiversity Atlas Sweden (https://bioatlas.se/), BISON - USGS (https://bison.usgs.gov/#home), iDigBio (https://www.idigbio.org/), Plants of Southern Africa (http://newposa.sanbi.org/), Japan Node of Global Biodiversity Information Facility (http://www.gbif.jp), and many more around the world; for a complete list of GBIF nodes and participants, see https://www.gbif.org/the-gbif-network). DiSSCo (Distributed Systems of Scientific Collections) is a new European initiative of 21 European to unify natural history collections (https://dissco.eu/). In particular, the Australian government funded the development of the Atlas of Living Australia (ALA) portal (https://www.ala.org.au/), which has become a model in the biodiversity informatics community. Computationally complex but with the source code fully available, ALA answers to the research community by providing numerous modules, entry search points, analytical tools and dense information about Australia's taxa. The biodiversity informatics community has been adopting and adapting the ALA model for developing national biodiversity portals (Living Atlases Community). It is clearly advantageous if a curated list of "accepted" taxonomic names exists side-by-side with these biodiversity portals (for example the Australian Plant Census and the ALA); although, there are possibilites for competing taxonomies to exist side-by-side.

Although not the focus here, numerous citizen science projects have been developed in the past years, which contribute importantly to our knowledge of biodiversity through BioBlitzes, publication of observations, and even digitisation of specimens (e.g., iNaturalist (https://www.inaturalist.org/), eBird (https://ebird.org/), eButterfly (http://www.e-butterfly.org/), Les Herbonautes (http://lesherbonautes.mnhn.fr/), to name just a few).

C. Morphological, Trait and Genetic Data

A number of different online tools have been developed to share basic data on the morphology and molecules of organisms. Molecular data are commonly shared through the National Center for Biotechnology Information (NCBI). It includes repositories on nucleotide sequences, as well as proteins, genomes, reference sequences and others. NCBI's GenBank (Benson et al. 2012)

is probably the most prominent sequence repository, even though there are other platforms for this type of data (e.g. European Nucleotide Archive; DNA Databank of Japan; Barcode of Life Data Systems). There are more limited resources linking biodiversity data portals to molecular data. One such initiative is the Barcode of Life, BoLD, http://www.boldsystems.org/, which links specimen records with DNA barcodes for various groups of organisms, on a user defined "project" basis.

Sharing of molecular data in dedicated online repositories is common practice, but the same cannot be said about morphology. Morphological data are still mostly restricted to descriptions in taxonomic works, and tabulated datasets are usually confined to tables or supplementary files accompanying publications. This is true for both categorical and continuous data, but the situation is worse for the later. Authors of morphometric studies usually present only tables with mean and variance values for each variable. Besides being available on journal websites, categorical data, commonly used in phylogenetic studies, are also shared as nexus files via TreeBASE (Piel et al. 2009). This overall absence of sharing of morphological data is in part related to the lack of dedicated open access portals. The TRY Plant Trait Database (Kattge et al. 2011) and MorphoBank (O'Leary and Kauffman 2012) are the two best known options available for publishing, searching, and downloading morphological data. MorphoBank is dedicated to morphology and allows inclusion of continuous and categorical data, plus supporting 2D and 3D images, including CT scans. TRY aggregates information on functional traits, including morphology, but consists only of tables of observations. The web includes other data resources on plant traits, such as chromosome numbers (IPCN), mass spectral data (MassBank; Horai et al. 2010) and metabolites (KNApSAcK; Afendi et al. 2012).

Although there are a number of resources for molecular and morphological data sharing, fragmentation of information in different databases is a problem. In addition, researchers must contend with problems related to the unbalanced proportion of data available for taxa and of taxonomic accuracy in the databases. DNA sequences are much more abundant than any type of morphological data, but molecular repositories show a significant amount of misidentification (Bridge et al. 2003, Vigalis 2003). There are no studies on the taxonomic quality of morphological databases, but we expect that anatomical data, commonly produced by taxonomists, to be more accurate, whereas the situation in chemical datasets may be similar to that of molecular databases.

Although not taxonomically centered, the National Ecological Observatory Network (https://www.neonscience.org) is an open access portal that provides comprehensive occurrence data and samples in the USA, specifically from 81 ecological field sites, all collected through standard protocols. NEON provides data from terrestrial, airborne, or aquatic environments, such as soil microbe occurrences, LiDAR imagery, or chemical properties of groundwater. Biodiversity samples consist of DNA extractions, frozen soil samples, bulk and pinned insect collections, herbarium vouchers, and partial or entire vertebrate specimens. This portal, and other such ecologically oriented portals (e.g. xxx), integrate biodiversity sampling and data with abiotic and biotic factors under an open access umbrella - all geared towards collaborative science intent on furthering our understanding of ecology through time and space.

D. Images and Photos

Digitization “electronic capture of data, including images” began decades ago in some herbaria, but a revolution is currently happening in the way herbarium specimens are being accessed and used (Soltis 2017). While many museums and herbaria worldwide have embraced online data opportunities, simultaneously the opportunities for computer-based image recognition are also rapidly improving (Nelson and Ellis 2018). New machine learning and artificial intelligence approaches of digitized specimens of herbaria will benefit taxonomic identification (Schuettpelz et al. 2017, Wäldchen and Mäder 2018). Botanical applications include taxon recognition from photos in the context of citizen science (Gardiner and Bachman 2016; Kress et al 2018; Yunis et al 2018). Here we give examples of citizen science networks and herbaria.

iNaturalist (http://www.inaturalist.org), founded in 2008, is an online citizen science network that collates biodiversity observations and identifications and presents them in a manner accessible and suitable for scientific and conservation research. With more than 1.9 million contributed observations, iNaturalist represents an enormous data where users post pictures of species and interact with expert taxonomists to resolve species identification (van Horn et al 2017; Bowser et al 2014).

Global Plants, a collaboration between JSTOR and some 300 herbaria, is now the world’s largest database of digitized plant specimens (http://about.jstor.org/content/ global-plants). It contains images of 2,482,901 million herbarium sheets (Unger et al 2018). The herbarium of the Museum National d'Histoire Naturelle (MNHN) in Paris (P) has recently digitized over 5 million

specimens, or about 90% of its entire collection of vascular plants, and these records are searchable on both the MNHN site (https://science. mnhn.fr/institution/mnhn/collection/p/item/search) (Bras et al 2017) and iDigBio. The Chinese Virtual Herbarium (CVH http://www.cvh.ac.cn/en), one of the National Science and Technology Infrastructures of China represents a collaboration of approximately 30 herbaria in China with collective holdings of 6 million specimens. It serves specimen records from the flora of China, with nearly 4 million digitised specimens currently online . The Australasian Virtual Herbarium (AVH http://avh.chah.org.au/), a consortium of Australian and New Zealand institutions, now serves over 7 million specimen records of plants, algae, and fungi and and can be accessed through the Atlas of Living Australia (ALA, http://ala.org.au). Collaborations in Brazil have produced the Institutos Nacionals de Ciencia e Tecnologia - Herbario Virtual da Flora e Dos Fungos (http://inct. florabrasil.net/), with over 5 million specimen records currently online (Soltis and Soltis 2016). Anthos (www.anthos.es) is a Spanish platform built upon the development of the *Flora Iberica* project. The database holds c. 1.8 million records specially with distribution points for plants in the Iberian Peninsula and Macaronesian islands. Information on the taxa treated generally include scientific drawings, synonyms, chromosome numbers, vernacular names, among other information, and interestingly photographs of fresh/living material for many of the species, for which the identification has been double checked by taxonomist working for *Flora Iberica*.

The Moscow University Herbarium (MW), the second- largest herbarium in Russia after the Komarov Institute (LE), digitised all Asian collections in 2016–2017 and published them online at https://plant.depo.msu.ru/, including 3283 specimens from Turkey. Currently, MW holds 331 type specimens of 285 taxa described from Anatolia—19 holotypes, 47 isotypes, 48 isolectotypes, and 203 syntypes (Seregin et al 2018). The William and Lynda Steere Herbarium of The New York Botanical Garden has been digitizing specimens since 1995. Approximately 2.5 million specimens have been databased and 1.4 million have been digitally photographed (http://sweetgum.nybg.org/science/vh/).These data are served through the Garden's C. V. Starr Virtual Herbarium and are shared through other data portals as well. Completion of the digitization of all American specimens (roughly five million) is projected by 2021 (Thiers et al 2016). Three major botanical institutions, The Morton Arboretum, the Field Museum of Natural History, and the Chicago Botanic Garden, have developed the online, searchable database vPlants ("virtual Plants") that provides free plant specimen data and digital images of specimens to anyone with internet access (The vPlants Project 2009). The JACQ Virtual Herbarium

(http://herbarium.univie.ac.at/ database/index.php), which has at its core 5.5 million specimens of the Naturhistorisches Museum Wien (W; http://www.nhm- wien.ac.at/en/research/botany) and 1.4 million specimens of the University of Vienna (WU; http://herbarium.univie.ac.at/), as well as more than 25 additional institutions, primarily from central and eastern Europe.

African Plants— A Photo Guide (www.africanplants.senckenberg.de) which covers at present ca. 3200 African plant species and contains ca. 25,000 photos. It combines the three regional portals launched earlier (i.e., West, Central, East African Plants) into one and enables the user at the same time to search in each region independently. Nomenclature generally follows the African Plant Database of the Conservatoire et Jardin Botanique in Geneva and SANBI Pretoria (African Plant Database). Free usage for non-commercial scientific and educational purposes is granted. The photo database is searchable according to systematic criteria (family, genus, species name, and—so far as entered— vernacular name in different) (Stefan et al 2014).

Live Plant Photos (https://plantidtools.fieldmuseum.org/en/nlp/5304). The Field Museum's Plant Identification Tools website - a curated botanical collection of plant specimens from Central and South America. The site is accessible in 3 languages, serving scientists around the world. The collection includes 100,000+ high-quality digital images of live plant photos and herbarium specimens, covering more than 35% of all species in the Neotropics. Images on this site are displayed with the consent of the photographers and are subject to the Field Museum's Conditions and Norms for the Use of Data and Images.

Leafsnap (Kumar et al 2012), a free mobile iPhone and iPad app, uses visual recognition of leaves to help identify tree species distributed in the Northeastern United States and Canada currently covering all 185 tree species (http://leafsnap.com). Leafsnap UK includes trees from across the United Kingdom with species information and imagery provided by the Natural History Museum in London. With over three million recorded leaf images, this citizen science app has potential on monitoring the effects of climate change and exotic species on tree distributions (Kress et al 2018).

E. Phylogenetic Information

Trying to understand biodiversity without considering evolutionary relationships is like viewing fine mosaic artistry as a pile of its individual tiles - the bigger picture is lost to view. An ideal legume portal would include a phylogenetic browser that would integrate individual phylogenies and their datasets or enable an overview of legume evolutionary relationships via tree grafting from numerous studies. Deposition of datasets and inferred phylogenies is good practice for biodiversity research (Penev *et al.* 2017) and many journals actually require it for publication. Current methods for linking phylogenetic trees to data include stand-alone databases such as TreeBASE (https://www.treebase.org) and DRYAD (https://datadryad.org), sometimes individual github accounts, act as repositories for molecular data and accompanying phylogenetic inferences. The next stage in usability would integrate subtrees across phylogenetic databases and The Open Tree of Life (OTOL; https://tree.opentreeoflife.org) is a good example (Hinchliff *et al.* 2015). Rather than simply compiling and re-analyzing datasets, OTOL integrates analysis pipelines with taxonomy (Rees & Cranston 2017), all to enable a supertree hypothesis of evolutionary relationships across life. Numerous wrappers and pipeline tools have been developed to enable interactive use and integration of OTOL into other databases and portals (e.g. Michonneau *et al.* 2016 and https://github.com/OpenTreeOfLife).

Use of phylogenetic knowledge can increase our ability to understand life around us and its dependence on and use of abiotic and biotic factors entwined across the globe. As an example, a recent ecological study reported that optimizing phylogenetic community trees with the OTOL backbone increased the reliability of phylogenetic diversity and dispersion indices (Gastauer *et al.* 2018). An excellent example of integrating biodiversity knowledge with phylogenetic information is Phylolink (https://www.biodiversityintelligence.com/phylolink), the successor to PhyloJIVE (Jolley-Rogers *et al.* 2014; Miller et al. 2018), which used the legume, *Acacia,* as the pilot taxon (see: https://www.ala.org.au/uncategorised/phylojive-integrating-biodiversity-data-with-phylogenies/). Integration of botanical specimen data with phylogenetic and spatial data have been noted as particularly powerful (Soltis et al 2018). Phylolink, in association with the Atlas of Living Australia, provides a phylogenetic view of biodiversity data. Viewing the distribution of a single species is a common and powerful integration of biodiversity data and most major biodiversity informatics portals have this capability. However, linking from a phylogenetic tree to be able to view the distribution of sister taxa in a single visualization is less common and more powerful form an evolutionary perspective. From any terminal or node of a tree Phylolink links out to

species pages. Phylolink is presently built into the Atlas of Living Australia and uses the ALA APIs to link to spatial distribution data and Phylolink visualizations can be transferred to the ALA spatial portal where climate and soil layers can be placed under the spatial distribution to give a complex view of the biodiversity. Further data integration will make these modules even more informative.

A large struggle in this area is the difficulty of easily viewing and navigating large phylogenies. The current tree viewer in Phylolink does not work well when the tree is larger than a few hundred species. Onezoom (http://www.onezoom.org/) is an elegant phylogenetic viewer for large phylogenies that works upon a fractal basis to easily zoom in and out of clades. The OneZoom web presence currently provides phylogenies, based on Open Tree of Life, including the Fabaceae (http://www.onezoom.org/life/@Fabales=956360#x-30,y906,w1.5118). Images and traits are easily incorporated into the visualization. Both OneZoom and traditional viewers like PhyloLink's visualization have advantages and disadvantages and they can be used to compliment each other.

4. Existing taxon-based portals: GOOD examples from other taxa

Numerous taxon centric portals have been developed over the last decades, each answering to specific needs within a community, and each using differents technologies. The landscape of biodiversity data sharing has tremendously changed in the last years, giving access to data in different ways. Designing a taxon centric portal today is not the same thing as designing one ten years ago. Still, looking at the strengths and weaknesses of existing portals is important to help design one that will be useful for the legume systematics community and beyond, that uses and takes full advantage of tools and services that are currently available, and that is designed to be capable of evolving as needed.

Aesthetic may not be a focal point in research projects, but it needs to be taken into consideration during the development of a portal. eMonocots (http://www.emonocot.org/), now included in Plants of the World Online (http://plantsoftheworldonline.org/), is a good example of attractiveness. The design is appealing, with high quality pictures and a simple research access to the database. Simplicity of navigation through the information is also a key component of a useful portal and may be one of the key features, for example, that has led to the popularity of The Plant List (http://www.theplantlist.org/). Citizen-science portals like eBird (https://ebird.org/home), eButterfly (http://www.e-butterfly.org/), leafsnap

(http://leafsnap.com/species) or the more generalistic iNaturalist (https://www.inaturalist.org/), are also good examples of appealing portals, where it is easy to navigate and search for specific information.
Numerous taxon centric portals visited are static, without dynamic links to other type of data, or direct update of information. However, several portal give access to different types of information, like litterature (), occurrences (specimens or observations), molecular sequences, traits, phylogenetic information,

- Encyclopedia of Life (http://eol.org/) … species pages, Wikispecies (https://species.wikimedia.org/wiki/Main_Page)
- eMonocots http://www.emonocot.org/) (now under Plants of the World Online): graphically beautiful but static. Taxonomy available but classification not easily visualized. No link to other types of data (sequences, traits), no links to specimens.
- Solanaceae (http://solanaceaesource.org/), BRAHMS database and information, descriptions, useful species, specimens collected, literature, links to other databases (project in progress?): visually less appealing, navigation more heavy, some tabs miss information, link to specimens through table visualization (but possibility to filter data)
- Global Compositae Checklist (https://compositae.landcareresearch.co.nz/) - moribund: not really pretty, links to maps not working
- Brassicaceae Database (https://brassibase.cos.uni-heidelberg.de/) - nice example: graphically more attracting, easy access and navigation in the classification, visualization with phylogenetic tree (but view is too general), possibility to export the tree (great for phylogenies). Possibility to visualize specimens in a table but not easy to navigate. A lot of tools, but everything is too disconnected from each other.
- Genera of Gesneriaceae (http://www.genera-gesneriaceae.at/) : not pretty, highly static, but history of classification is interesting to have. Basic species page, but could be prettier.
- Ants: Global Ants Database (http://globalants.org/): pretty with a lot of information but not a lot of links to navigate within all the information, maps are static, need a login to access database (a big no!), interesting documentation available; AntBase (http://antbase.org/ moribund?): more of a portal to access other databases than a real portal; AntWeb (https://www.antweb.org/): need to use their own filter, which is not practical, map seems clickable but not working, need to login to access some tools, so seems pretty but not as

easy to use; Ant Profiler (http://www.antprofiler.org/): database not responding ; Ant Genera of the World (http://robdunnlab.com/antgenera/): just a series of map without any more information, so not really useful.

- eBird (https://ebird.org/home): no specimens data. Pretty, easy to use but map not clickable, and no easy access to specific observations. Pros: published on GBIF eButterfly (http://www.e-butterfly.org/), iNaturalist (https://www.inaturalist.org/) - focused on observations and citizen science, data quality issue.
- Fishbase - one of the most widely used, successful models
- Vascan - no link to specimens, species pages static
- MyCoPortal (http://mycoportal.org/portal/index.php)
- BryoQuel (http://societequebecoisedebryologie.org/Bryoquel.html)
- VertNet

Species pages portals

Discussion

(Part I: Existing Resources -what is usable, how to not reinvent the wheel)

5. Lessons Learned (from past efforts in legumes and other taxa)

- This might lead into a section called “desired features”

Before planning a new online resource for legumes it is worthwhile to reflect on past efforts and consider both what worked well and what led to their demise.

Technological change and planning for future re-use of data

Continuing and rapid technological change is inevitably one factor. Any digital infrastructure need be designed, as far as possible, with future change in mind. Whilst no software nor data platform survive (although older software - such as that underlying ILDIS - still run on modern machines albeit with archaic interfaces) the data collated in older publications or databases are as relevant today as ever. It is thus beholden to those building data platforms to provide for future transfer and evolution of content.

Data standards are required

ILDIS succeeded in bringing together data contributions from multiple systematists and other legume scientists by communicating clearly and rigorously implementing the data standards to be followed. It remains the case that effective exchange of data records requires such standards. Ensuring that contributors use and interpret terms consistently makes data retrieval and analysis possible. A limitation of LOWO for example (where the content derives from the

text contained within a traditionally published book) is that a search for “medicinal”, for example, misses any genera with use as “herbals”.
Standards created through the taxonomic databases working group have, in more recent years, focused on data exchange protocols and formats such a Darwin Core. There remains, however, a critical need for “terminological controls” to constrain country boundaries, habitat or habit definitions, categories of use etc. Selection, and possible development of standards would need to be an early stage for a new initiative.
Application programming interfaces permitting data exchange are increasingly implemented for relevant information systems but the technology cannot resolve fundamental differences in how data values are used and interpreted within different systems.

A common taxonomic backbone

The obvious starting point for any form of standardisation would be the scientific names (and synonyms) and particularly the taxonomic arrangement adopted by each participant or databases to be linked. Many botanical databases exist in which the same plant is erroneously listed twice under alternative synonyms. Databases using alternative accepted names for a species can still share data provided a fully synonymised list is available which would equally be required to link the resource with databases using different taxonomies.
Data exchange becomes meaningless, however, where species concepts differ. LOWO was unable to employ the ILDIS legume checklist since it had been built from a much older taxonomic viewpoint. Discussion and adoption of a consistent taxonomic backbone is essential. It is equally clear that such a backbone will only be fit for purpose if it is actively curated since otherwise it will quickly cease to fulfil the requirement. Thus TPL, for all its merits, is unsuitable for such a purpose.

Central versus Distributed

The success of TROPICOS, the World Checklist, GBIF along with most other current digital initiatives demonstrate the merit in having a coordination centre based in a single institute. The latter assumes responsibility for building and sustaining the digital infrastructure whilst contributors may be based in many different institutions. ILDIS illustrated this. The drive to make things happen, funding applications and concept development all occurred in Southampton University where the ILDIS “coordinating centre” built, communicated with and promoted a global network of participating specialists. A number of committees were formed to develop thinking and practice in different areas including content, software and use.

Disadvantages of relying on a strong coordinating centre include a reliance on a single institution to provide core funding and infrastructure. The community of contributors are beholden to continued support from one institute. A centralised endeavour also, inevitably perhaps, reduces the sense of joint ownership. ILDIS worked hard through regular workshops and meetings to maintain the network of specialists each providing a geographic or monographic data set. Nevertheless the demise of a coordinating team communicating and driving led, eventually, to a cessation of curatorial or promotional activity. LoW and the website

built from it (LOWO) were equally organised centrally, this time by Kew. Chapter authors contributed text for inclusion in the book but were not subsequently engaged in the design of the online version. Nor, initially, were they able to edit the content for the chapter in which they held expertise. It was tragic that LOWO had begun piloting editing by contributors in different institution when Kew changed its Content Management System once more - demonstrating its dependence on Kew.

University vs Research Institute

Whilst the focus of all institutions change over time, the rate of change is that much greater within university departments which are reliant on research funding which is primarily short-term in nature. Established institutions, in the past at least, could count on core funding to curate their physical and digital assets. Provided an information resource continued to be seen to fulfil a useful role for the institute the initiative survives. Thus IPNI, which builds on Index Kewensis (a series of volumes published by Kew from 1893) has evolved over the decades, including its merger with Gray Cards and the Australian Plant Name Index to take its current form. The botanical community depend on this key resource and it continues to be sponsored almost entirely by Kew. Such longevity is however rare and arguably, far more difficult to achieve within a university department since research funding is rarely made available for practical tasks required to construct and curate data sets.

Linking systems

It is relatively straightforward to download data from one system for import and use elsewhere. This is increasingly commonplace but is problematic that the latest updates in one system may not be reflected in all systems. It often depends upon refreshing data uploads using an agreed frequency and protocol (or data dump). Linking information systems which share common data in such a way that edits and improvements in one system are automatically reflected in the other is becoming more common with the growing popularity of application programming interface (APIs). APIs are now commonly used to call information from one source to display or integrate it another such as displaying GBIF species distribution data on the webpage of EOL. As more data is available via APIs it will be easier to link data and generate modular, user-specific data integrations and visualizations.

The social dimension

There are undoubtedly technical challenges inherent in building a new information system, particularly one that will link to many existing platforms and further challenges as to harnessing and curating content in a sustainable manner which meets our goals. What cannot be overestimated, however, are the social challenges in organising the ideas, contributions, data and efforts of the wider community and, in particular, in sustaining the interest and shared commitment over time. This begins with agreeing a sufficiently clear and tightly defined purpose and target audience. Subsequently some form of coordination or management group will inevitably be required to oversee and support implementation. Sustaining such coordination

has, arguably, been the most challenging issue facing earlier efforts and the single biggest factor in their lack of continuity. Any new project will require a considered proposal covering the multiple social aspects including how credit and recognition will be achieved.

Part II: What we want and need (who is the target audience?) (Fig. x)

6. Uses of Web Portals and Future Applications

- Examples of what can be done with these portals by botanists and those in many other fields
    - Online floras
    - Id keys
    - Online digital Floras ?
        - Kew's digital Floras
        - Flora do Brasil
    - Identification aids
    - Neotropikey(http://static1.kew.org/science/tropamerica/neotropikey.htm) - Lucid-based identification resources for neotropical plantsneotropikeys
- Many botanists either can't remember a time when specimen data wasn't available online or if they do have found it so important in their work-flow that it has revolutionised our work and we wonder how we worked without it
- "Cleaned datasets" in real time
- For mapping
- For modelling climate change
- Species distribution modelling
- For automated IUCN extinction risk assessments
- Invasive species
- Cultivated and economically important legumes

In the first comprehensive continent-scale initiative to digitise herbarium specimens, in early 2000s the Australasian Virtual Herbarium (AVH) went online. In fact, databasing of the legume genus Acacia at the National Herbarium of Victoria, was used as a pilot for the AVH initiative (Tim Entwisle pers comm). As such the AVH can serve as a test-case for some observations on data usage by the botanical research and broader community. Usage records for the Australasian Virtual herbarium have been analysed (Cantrill 2018) to reveal the most common

uses of the data, with the largest category of usage being for scientific research, and namely ecological research. There was also a large component of non-scientific use for biosecurity, conservation management and education; although, somewhat surprisingly, the use of botanical specimen data for taxonomic and systematics research was low (Cantrill 2018). Other major uses of digitized botanical specimen data have been in the fields of global environmental change, phenology, climatology and various ecological modelling studies (Lang et al 2019, Soltis 2017, Soltis et al 2018).

While some contributors to this paper may favour a portal serving mainly legume systematists, there are strong arguments for building the portal around other user needs such as plant identification. Facilitating plant identification for every type of user from 1) farmers identifying pests, 2) conservationists identifying rare, threatened or endangered species, 3) ecologist identifying sterile plot vouchers, 4) teachers teaching differences between legume species, genera, etc., 5) foresters, agronomists and biotechnologists needing reliable names, and 6) food producing companies and chefs checking the correct names of their ingredients, will not only attract a wide range of users, but also reach out to and engage a larger community to provde financial support for a new revamped legume portal. Fishbase (www.fishbase.de), one of the most successful, globally most visited, taxon-based non-profit organisations, is an exellent example demonstrating how a wide user community has assisted the board of Fishbase secure financial support for development and maintenance over more than 20 years.

Examining the uses and searches in Fishbase: a) it had in excess of 300K unique visitors per month in the first six months of 2016 (www.Fishbase.de/Weduse.php); b) 24% of users searched the comprehensive species summaries (e.g. Rhincodon typus = whale shark), 5% common name, 7% looked up photos, 3% searched by country information, 5% on references, 7% on scientific name, 2% identification, and 1% entered via the extensive glossary, the remaining search terms received below 1% hits. While this provides some indication of user preferences, it also serves to say that the user interface of Fishbase could do with a revamp which would likely see an increase in users in general. In comparison, in 2016 Neotropikey (http://www.kew.org/science/tropamerica/neotropikey/families/glossary.htm) received on average 5,000 visitors per month with the family key receiving 6% of visits, the three legume subfamily pages each 3.5% of visits, the comprehensive glossary 3.3%, and the Myrtaceae page 2% of visits. We predict that with a revamp, translation into Spanish and Portuguese, and

relaunch the traffic to Neotropikey would increase significantly. Interestingly the Neotropical Plants image database hosted also on the on the Neotropikey landing page receives more visitors (+++check how many) than the family key and the family webpages, indicating that users prefer photo galleries over keys for identification.

Surveying the communities of plant taxonomists and portal builders (+++refs.), it becomes evident that 1) traditional dichotomous keys are too complicated for most non-specialist users, 2) Lucid keys are costly and time consuming to build well and to maintain, 3) image-based gallery filtering systems such as eMonocots, iNaturalist, Jstor global, and (+++refs.) are favoured by scientist as well as non-scientist user communities, and 4) not yet possible to use artificial intelligence for reliable and accurate species identification. The advantage of a filtering system is that it could rely on a limited number of easily recognisable diagnostic characteristics and still efficiently guide the user through the portal, if all species were provided with short, coded descriptions including these diagnostic characteristics (e.g. habit, leaf position and type, flower colour, geographical range, and uses). Sources of exhaustive glossaries of legume terms exist already. Providing a complete glossary with images and linking it to the short descriptions would facilitate identification for an even wider range of user communities.

Geographic occurrence data is needed by a range of user communities, e.g. ecologists (for modelling climatate change), taxonomists (for mapping taxa), conservationists (for IUCN extinction risk assessments). Currently cleaning and georeferencing specimen data is extremely time consuming; and this process is often repeated for the same collection (by different researchers in different herbaria) or even for the same specimen. This is mainly because big data aggregators such as GBIF do not provide a facility for feeding cleaned GBIF georeferenced data back into the portal for future use by other researchers. Suppling a future legume portal with a facility for hosting cleaned georeferenced legume data will undoubtedly make life easier for legume researchers while, at the same time, attract a range of additional users seeking reliable locality data, as has already been seen for the online Australasian Virtual Herbarium.

7. Resources and information that should be integrated in a legume portal

- General aspects: easy to use????!!!

- easy to maintain, easy to contribute to, robust, scalable, find and reuse data, functions and tools from other sites, efficient implementation of a range of functions
- Reporting? Can report errors to data providers
- Data Export functions?
- Real-time? [1] updating web portal with abundant resources
- taxon list, collecting information, distribution, morphological traits, habitat, some useful pipelines or hyperlinks to other tools
- Legume names (accepted and synonyms) being dynamically updated and names curated by specialists; and corrections returned to source DB
- Links to GBIF records and other specimen databases (cleaned up records???) to names
- Ideally cleaned specimen records will be georeferenced and imaged for easy identification
- Lucid keys to higher as well as lower taxa
- Bean Bag repository, or make the Bean Bag into a dynamic part of this portal
- Open Tree of the Fabaceae (https://www.biodiversityintelligence.com/fabaceae-opentree)
- Agreed Editorial control procedures and mechanisms to support
- Sustained with a business model to upkeep for 20 years?

>Advantages of a taxon portal: integration vs user; quick access to information pertaining to the taxon; curation; improve data quality;

A taxon-based portal is a powerful tool that allows users to have different types of datasets easily accessible and integrated. It is important to proper define who are the users and what kind of data to make available in order to increase the portal's lifespan and usability. Legume systematists and evolutionary biologists are likely the main users, but a legume portal would gain track if it could also be used by data scientists. Thus, it is important to develop a model that can be used by both communities. In this context, two factors play a major role: data quality and exportation capabilities. Download formats have to be included in the initial definition of the platform's architecture. The design should also take into account different options for data curation and cleaning, either

1

automatically or by experts. This second feature resonates with and justifies a taxon-based portal, but it is also more difficult to implement.

A solid database is the most prominent feature of a taxon-based portal, but it can also serve as a hub for other types of information and tools. Interactive digital keys for legumes would be interesting for a broader range of users, particularly citizens. Implementation of data visualization tools such as the Open Tree of Life (Hinchliff et al. 2015 - opentreeoflife.org) can help users to identify potential relationships between different data partitions, besides an aid for teaching. Sections of the portal could be a source of news on current research and a repository for past and current references on the increasing number of publications in Leguminosae. The possibilities are almost endless, but it paramount not to lose sight of the main objective of such a portal: integrate and curate information useful for the development of our knowledge on legume systematics and evolution.

(moved from above) These problems supports the need for taxon-based portals in which information would be aggregated and curated. A centralized view of data by taxon could also help to drive research aiming to fill knowledge gaps. It is possible to merge data from multiple sources in a single platform, but integration would still be superficial. Efficient unification could be achieved with the widespread use of ontologies and semantic annotations. These are available for genomes, but rarely used to annotate phenotypes (Deans et al. 2015). Besides being an useful tool to make good information on a particular taxon readily available, a taxon-based portal can be a way to promote the adoption of semantic tools. Such changes have the power to build powerful discovery instruments (Deans et al. 2015) out of our datasets.

Open Tree of Life has developed a Application Programming Interface (API) for the synthetic tree. This allows a user to send a list of taxa to the Open Tree API and receive back a subset of the synthetic tree that contains only the taxa of interest. A legume portal could take advantage of this technology by always accessing the latest edition of the synthetic tree at only the Fabaceae node. An advantage of this legume community can then take “ownership” of the curation of the Fabaceae on Open Tree. Areas of missing data or poor resolution can be identified by the legume community and published trees identified, uploaded and curated in Open Tree to fill the gaps or new studies initiated to generate the missing data.

Part III: How do we make this project happen: Effective pipeline for developing a legume portal

8. Appropriate technologies and tools (Fig. ?) - constraints and possible technologies...depends on timing and resources, etc.

- This is the more techy section of how we make this happen
- What is the central identifier that links the multiple sources? Taxonomy? from
- Potential existing informatic tools and software: TaxonWorks, ALA tools, ALA Hub, TDWG, DarwinCore
- Controlled vocabularies and ontologies
- APIs
- Direct points to other resources from each species page or Searches of external sources using the relevant taxon name
- Web Services to download data subsets and refresh
- Data cleaning tools and pipelines: OpenRefine, R
- An important technical issue, needed for a coherent infrastructure, regardless of where data are served, is the need for unique identifiers for specimens and for people

Need for researchers to publish data in standardised DWc format - via national nodes. Legume portal will be useful and pertinent if researchers contribute

Above we have highlighted how taxon based portals still have an important role in biological data sharing, but we need to devise a plan to build a system that would relevant for legume systematics. In this context, it is clear that information should be aggregated around the classification system. Names and associated vouchers would link the different types of data extracted from other databases. The modular nature of the portal would allow people to bypass taxonomy, but centralizing on names facilitates curation of the data, the most desired feature of such an information system. To develop a legume portal, it is first necessary to define a basic model on top of which future developments can be made. Construction of the model relies on three main steps: 1. Definition of type of information to aggregate; 2. Selection of online data sources; 3. Development of bots to centralize information; 4. Production of an online graphical user interface (GUI) to retrieve and visualize information.

There is a large number of organismic data available, but it is necessary to select a smaller number of information to begin with. Considering the nature of the portal, data on species names, herbarium records (vouchers), dna sequences, morphological data, and occurrence data are probably the most relevant. All this data can be automatically retrieved from existing databases, as done for *Arabidopsis* proteomics data (Joshi et al. 2011), using web scraping. This method facilitates data aggregation, but it is important to define priorities, particularly focusing on the portal identity as done in a previous section.

Having chosen the data we want in the portal, the second problem we face is related to decentralization of information in the web. For instance, there are multiple sources for sequence data, probably with unbalanced taxon sampling. Although we would like to sample all of these sources, each may have particular features that affect automatic data extraction (Stein 2003). Thus, until a resilient system is built, it is important to choose among alternative data sources on the first steps of development of the portal. Other sources can be included during late implementations in order to increase data coverage during scraping. The amount and quality of data on legumes are two basic principle that may guide selection of primary databases.

Data scraped from different sources should be stored in a central database in order to provide easier — and faster — access to the information. This highlights the need to have a base server, and to develop bots responsible for automatic data extraction data from other websites. Both these factors are the most expensive, but creation of the web scrapers is probably the most effort-consuming task of such a project. Particularly, long term maintenance of the tool is necessary in order avoid regular crashes and to keep up with updates in the source databases (Stein 2003). Nonetheless, there is a trend for public databases and services to build API's (e.g. Miller et al. 2015) that facilitate the use of their tools or gathering of their data. Collaboration between the legume systematics and data scientists communities is probably a good way to facilitate writing of bots, and support the cyber infrastructure needed create a dataset useful for answering different scientific questions, as well as other potential users.

Continued usage depends large part on the ability of users to query, find, and extract data in an easy and free way. Thus, even though creation of a GUI is the last step of this development model, it is central for increasing the lifespan of a data portal. For that end, it is imperative to define would be the user of a taxon portal. Although systematists — and data scientists — are likely the focal group, expansion of usage to citizens and other scientists may increase usage and justify acquisition of funds for development and maintenance.

Other functionalities can be included in subsequent versions of the portal. Although there is a large number of desired features, some should be prioritize. Implementation of data cleaning, transformation, and exportation processes would be central at this point. The two last steps of this chain have to be automatic, but data cleaning can also be done by users. Human curation highlights the need to have unique identifiers both for data and for people. That way it is possible to pinpoint changes and contribute to a coherent database. Additional steps during refinement of the portal would include increased interaction between datasets via API's, and implementation of controlled vocabularies and ontologies. Ideally, the portal would evolve together with technological advances, but is important to keep in mind that curation and easiness of use are central to connect taxonomists, citizens and data scientists. With that goal we may be able to have a life lasting legume portal.

9. Organisation, People, Resources and Sustainability

- Section on social aspects (community endorsed, community driven, etc.)
- People who work together !

Workshops to establish a) Audiences b) Need c) what we have and what we missing d) benefits to users e) benefits to community building f) risks and addressing them

As we have advocated above, taxon based portals are still relevant tools for the systematics community, and other people. We have outlined a development model, but we still have to discuss the means to achieve our goal. Considering the task at hand — to build an information portal for the third largest family of flowering plants — we propose to use the design sprint method.

In design sprint, a group of creators get together to develop a desirable product (Banfield et al. 2015, Knapp et al. 2016). The goal is to design the product as well as to build, and test a prototype with a small set of users, particularly looking for flaws that would lead to failure of the project (Knapp et al. 2016). The creators finally move to actual production, or to fixing design problems through an iteration of the design sprint. Design sprints are supposed to occur in person and last full five days (Banfield et al. 2015, Knapp et al. 2016), but they do not have to follow this formula. It is possible to use online tools and adapt the frequency of the activities according to availability and still so follow the guidelines of a design sprint.

We can adapt this methodology to build a legume data portal. Instead of trying to focus on the family, different research groups could try to build a cyber infrastructure to their legume taxa of interest. This would produce a number of parallel iterations of design sprints that could lead to a model — or models — with higher chances of being applied to the Leguminosae as a whole. An advantage of this approach is that it may increase chances of funding, as the need to aggregate information from multiple sources can be linked to objective scientific questions.

Research groups that already have a multidisciplinary approach to taxa probably would have a better chance of developing interesting prototypes. Nonetheless, legume systematists interested in the ideas presented here could try to organize science hackathons (Groen and Calderhead 2015). Besides forming such multidisciplinary groups, these meetings could lead to new scientific questions (Groen and Calderhead 2015), besides contributing to develop and test a portal model.

Having multiple small prototypes tested, we would then proceed to choose the solutions that would work better for a large portal. Ideally, people not associated with any of the design sprints should conduct the evaluation process to avoid biases. Also, even though we want the prototypes do differ from each other, too much divergence can be a problem. This can be minimized if we restrain particular variables, such as source databases, language for scrapers design (e.g. python; Mitchell 2018) and databasing (e.g. MongoDB; Chodorow 2013). That way, it would be easier to build a hybrid portal based on the best solutions found by each team.

Concluding Remarks

References

Adey, M.E., Allkin, R., Bisby, F.A., White, R.J. & Macfarlane, T.D. (1984) The Vicieae Database: an experimental taxonomic monograph. In: R. Allkin and F.A. Bisby, eds. Databases in Systematics, London, Academic Press, Systematics Association Special Volume 26, 175-188

Allkin, R. (1984) Handling Taxonomic Descriptions by Computer. In: R. Allkin and F.A. Bisby, eds. Databases in Systematics, London, Academic Press, Systematics Association Special Volume 26, 263-278.

Allkin, R. & White, R.J. (1988) Data management models for biological classification. In: Bock, H.H., ed., Classification and related methods of data analysis, Elsevier (North-Holland), Amsterdam. Pp. 653- 402

Allkin, R. & White, R.J. (1993). XDF Data exchange format. In “Advances in Computer Methods for Systematic Biology: Artificial Intelligence, Databases and Computer Vision”, pp. 474-5. R.Fortuner (ed.), Publ. John Hopkins University Press, Baltimore and London.

Afendi FM, Okada T, Yamazaki M, Hirai-Morita A, Nakamura Y, Nakamura K, Ikeda S, Takahashi H, Altaf-Ul-Amin M, Darusman LK, Saito K (2012) KNApSAcK family databases: integrated metabolite–plant species databases for multifaceted plant research. Plant and Cell Physiology 53(2), e1. DOI: 10.1093/pcp/pcr165.

Banfield R, Lombardo CT, Wax T (2015) Design Sprint: A Practical Guidebook for Building Great Digital Products. (O'Reilly Media, Inc.: Sebastopol)

Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW (2012) GenBank. Nucleic acids research 41(D1), D36-42.

Bisby, FA (2000) The quiet revolution: biodiversity informatics and the internet. Science 289: 2309-2312.

Bowser A, Wiggins A, Shanley L, Preece J, Henderson S (2014) Sharing Data While Protecting Privacy in Citizen Science. Interactions 21 (1), 70-73.

Bras G L, Pignal M, Jeanson ML, Muller S, Aupic C, Carré B, Flament G, Gaudeul M, Goncalves C, Invernón VR, Jabbour F, Lerat E, Lowry PP, Offroy B, Pimparé EP, Poncy O, Rouhan G, Haevermans T (2017) The French Muséum National d'histoire Naturelle Vascular Plant Herbarium Collection Dataset. Scientific Data. 4, 170016.

Bridge PD, Roberts PJ, Spooner BM, Panchal G. (2003) On the unreliability of published DNA sequences. New Phytologist 160, 43-48.

Cantrill, D. J. (2018). The Australasian Virtual Herbarium: Tracking data usage and benefits for biological collections. Applications in plant sciences, 6(2), e1026.

Cayuela L, Granzow-de la Cerda I, Albuquerque FS, Golicher JD (2012) Taxonstand: An R package for species names standardization in vegetation databases. Methods in Ecology and Evolution 3(6), 1078-1083.

Chard K, Dart E, Foster I, Shifflett D, Tuecke S, Williams J. 2018. The Modern Research Data Portal: a design pattern for networked, data-intensive science. *PeerJ Computer Science* 4:e144 https://doi.org/10.7717/peerj-cs.144

Conte, Gaillard, Lanau, Rouard, Périn (2008) GreenPhylDB: a database for plant comparative genomics. Nucleic Acids Research 36, D991–D998. doi:10.1093/nar/gkm934.

Dallwitz, M.J. (1993) DELTA and INTKEY. In 'Advances in computer methods for systematic biology: artificial intelligence, databases, computer vision', pp. 287–296. (Ed. R. Fortuner.) The Johns Hopkins University Press, Baltimore, Maryland. Also available at delta-intkey.com.

Dash S, Campbell J, Cannon E, Cleary A, Huang W, Kalberer S, Karingula V, Rice A, Singh J, Umale P, Weeks N, Wilkey A, Farmer A, Cannon S (2016) Legume information system (LegumeInfo.org): a key component of a set of federated data resources for the legume family. Nucleic Acids Research 44, D1181–D1188. doi:10.1093/nar/gkv1159.

Deans A, Lewis S, Huala E, Anzaldo S, Ashburner M, Balhoff J, Blackburn D, Blake J, Burleigh J, Chanet B, Cui H, Dahdul W, Das S, Cooper L, Dececchi T, Dettai A, Diogo R, Druzinsky R, Dumontier M, Franz N, Friedrich F, Gkoutos G, Haendel M, Harmon L, Hayamizu T, He Y, Hines H, Ibrahim N, Jackson L, Lecointre G, Lapp H, Jaiswal P, James-zorn C, Ko S, Lundberg J, Macklin J, Mast A, Lawrence C, Nove N, Mungall C, Oellrich A, Osumi- D, Midford P, Parkinson H, Ruttenberg A, Schulz K, Segerdell E, Seltmann K, Sharkey M, Smith A, Smith B, Specht C, Squires R, Thacker R, Thessen A, Fernandez-triana J, Vihinen M, Vize P, Vogt L, Wall C, Walls R, Westerfeld M, Wharton R, Wirkner C, Woolley J, Yoder M, Zorn A, Mabee P (2015) Finding our way through phenotypes. PLoS biology 13(1), e1002033. DOI: 10.1371/journal.pbio.1002033

Gardiner LM, Bachman SP (2016) The role of citizen science in a global assessment of extinction risk in palms (Arecaceae). Botanical Journal of the Linnean Society, 182 (2), 543–550.

Gastauer M, Caldeira CF, Trotter I, Ramos SJ, Neto JA. Optimizing community trees using the open tree of life increases the reliability of phylogenetic diversity and dispersion indices. Ecological Informatics. 2018 Jul 1;46:192-8.

Gewin, V. 2002. All living things, online. Nature 418: 362-363

Godfray, H.C.J. 2002. Challenges for taxonomy. Nature 417: 17-19.

Gonzales M, Archuleta E, Farmer A, Gajendran K, Grant D, Shoemaker R, Beavis W, Waugh M (2005) The Legume Information System (LIS): an integrated information resource for comparative legume biology. Nucleic Acids Research 33, D660–D665. doi:10.1093/nar/gki128.

Goodstein D, Shu S, Howson R, Neupane R, Hayes R, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N, Rokhsar D (2012) Phytozome: a comparative platform for green plant genomics. Nucleic Acids Research 40, D1178–D1186. doi:10.1093/nar/gkr944.

Goswami, A (2015) Phenome10K: a free online repository for 3-D scans of biological and palaeontological specimens. www.phenome10k.org.

Groen D, Calderhead B (2015) Cutting Edge: Science hackathons for developing interdisciplinary research and collaborations. Elife 4, e09944

Hinchliff CE, Smith SA, Allman JF, Burleigh JG, Chaudhary R, Coghill LM, Crandall KA, Deng J, Drew BT, Gazis R, Gude K (2015) Synthesis of phylogeny and taxonomy into a comprehensive tree of life. Proceedings of the National Academy of Sciences 112(41), 12764-12769.

Horai H, Arita M, Kanaya S, Nihei Y, Ikeda T, Suwa K, Ojima Y, Tanaka K, Tanaka S, Aoshima K, Oda Y. (2010) MassBank: a public repository for sharing mass spectral data for life sciences. Journal of mass spectrometry 45, 703-714.

James, S. A., Soltis, P. S., Belbin, L., Chapman, A. D., Nelson, G., Paul, D. L., & Collins, M. (2018). Herbarium data: Global biodiversity and societal botanical needs for novel research. Applications in plant sciences, 6(2), e1024.

Jolley-Rogers G, Varghese T, Harvey P, dos Remedios N, Miller JT. Phylojive: Integrating biodiversity data with the tree of life. Bioinformatics. 2014 Jan 17;30(9):1308-9.

Joshi HJ, Hirsch-Hoffmann M, Baerenfaller K, Gruissem W, Baginsky S, Schmidt R, Schulze WX, Sun Q, Van Wijk KJ, Egelhofer V, Wienkoop S (2011) MASCP Gator: an aggregation portal for the visualization of Arabidopsis proteomics data. Plant physiology 155, 259-270.

Kattge J, S. Diaz, S. Lavorel, I. C. Prentice, P. Leadley, G. Boenisch, E. Garnier, M. Westoby, P. B. Reich, I. J. Wright, J. H. C. Cornelissen, C. Violle, S. P. Harrison, P. M. v. Bodegom, M. Reichstein, B. J. Enquist, N. A. Soudzilovskaia, D. D. Ackerly, M. Anand, O. Atkin, M. Bahn, T. R. Baker, D. Baldocchi, R. Bekker, C. Blanco, B. Blonder, W. J. Bond, R. Bradstock, D. E. Bunker, F. Casanoves, J. Cavender-Bares, J. Q. Chambers, F. S. Chapin, J. Chave, D. Coomes, W. K.

Cornwell, J. M. Craine, B. H. Dobrin, L. Duarte, W. Durka, J. Elser, G. Esser, M. Estiarte, W. F. Fagan, J. Fang, F. Fernandez-Mendez, A. Fidelis, B. Finegan, O. Flores, H. Ford, D. Frank, G. T. Freschet, N. M. Fyllas, R. V. Gallagher, W. A. Green, A. G. Gutierrez, T. Hickler, S. Higgins, J. G. Hodgson, A. Jalili, S. Jansen, C. Joly, A. J. Kerkhoff, D. Kirkup, K. Kitajima, M. Kleyer, S. Klotz, J. M. H. Knops, K. Kramer, I. Kuehn, H. Kurokawa, D. Laughlin, T. D. Lee, M. Leishman, F. Lens, T. Lenz, S. L. Lewis, J. Lloyd, J. Llusia, F. Louault, S. Ma, M. D. Mahecha, P. Manning, T. Massad, B. Medlyn, J. Messier, A. T. Moles, S. C. Mueller, K. Nadrowski, S. Naeem, U. Niinemets, S. Noellert, A. Nueske, R. Ogaya, J. Oleksyn, V. G. Onipchenko, Y. Onoda, J. Ordonez, G. Overbeck, W. A. Ozinga, S. Patino, S. Paula, J. G. Pausas, J. Penuelas, O. L. Phillips, V. Pillar, H. Poorter, L. Poorter, P. Poschlod, A. Prinzing, R. Proulx, A. Rammig, S. Reinsch, B. Reu, L. Sack, B. Salgado-Negret, J. Sardans, S. Shiodera, B. Shipley, A. Siefert, E. Sosinski, J.-F. Soussana, E. Swaine, N. Swenson, K. Thompson, P. Thornton, M. Waldram, E. Weiher, M. White, S. White, S. J. Wright, B. Yguel, S. Zaehle, A. E. Zanne, C. Wirth. 2011. TRY - a global database of plant traits. Global Change Biology, 17:2905-2935.

Knapp J, Zeratsky J, Kowitz B (2016) Sprint: How to solve big problems and test new ideas in just five days. (Simon and Schuster: New York, London, Toronto, Sidney and New Delhi)

Kress, W. J., et al. (2018). "Citizen Science and Climate Change: Mapping the Range Expansions of Native and Exotic Plants with the Mobile App Leafsnap." BioScience 68(5): 348-358. 10.1093/biosci/biy019

Kumar N, Belhumeur P, Biswas A, Jacobs D, Kress J, Lopez I, Soares J (2012) ‘Leafsnap: A Computer Vision System for Automatic Plant Species Identification.’ (springer) doi:10.1007/978-3-642-33709-3_36.

Lang, P. L., Willems, F. M., Scheepens, J. F., Burbano, H. A., & Bossdorf, O. (2019). Using herbaria to study global environmental change. New Phytologist, 221(1), 110-122.

Lewis, G.P., Schrire, B., Mackinder, B. and Lock, M. eds. (2005) Legumes of the World, Royal Botanic Gardens, Kew.

Li J, Dai X, Liu T, Zhao PX (2002) LegumeIP: an integrative database for comparative genomics and transcriptomics of model legumes Nucleic Acids Research 40, D1221–D1229. doi:10.1093/nar/gkr939

Lock, J.M. (1989) Legumes of Africa: a checklist.Royal Botanic Gardens, Kew.

Meineke EK, Davies TJ, Daru BH, Davis CC. 2018 Biological collections for understanding biodiversity in the Anthropocene. Philosophical Transactions of the Royal Society B: Biological Sciences 374: 20170386.

Michonneau F, Brown JW, Winter DJ. rotl: an R package to interact with the Open Tree of Life data. Methods in Ecology and Evolution. 2016 Dec 1;7(12):1476-81.

Miller JT, Pirzl R, Rosauer D, Jolley-Rogers G, Varghese T. 2018. Phylolink: Phylogenetically-based profiling, visualisations and metrics for biodiversity. Bioinformatics. doi.org/10.1093/bioinformatics/bty792

Nelson G, Ellis S (2018) The history and impact of digitization and digital data mobilization on biodiversity research. Philosophical transactions of the Royal Society of London Series B, Biological sciences 374, 20170391. doi:10.1098/rstb.2017.0391.

O'Leary MA, Kaufman S (2011) MorphoBank: phylophenomics in the "cloud". Cladistics. 27(5), 529-37.

O'Leary, MA, Kaufman S (2012) MorphoBank 3.0: Web application for morphological phylogenetics and taxonomy. http://www.morphobank.org

Penev L, Mietchen D, Chavan V, Hagedorn G, Smith V, Shotton D, Ó Tuama É, Senderov V, Georgiev T, Stoev P, Groom Q, Remsen D, Edmunds S (2017) Strategies and guidelines for scholarly publishing of biodiversity data. Research Ideas and Outcomes 3: e12431. https://doi.org/10.3897/rio.3.e12431

Pennisi, E. 2000. Taxonomic revival. Science 289: 2306-2308.

Piel WH, Chan L, Dominus MJ, Ruan J, Vos RA, Tannen, V. 2009. TreeBASE v. 2: A Database of Phylogenetic Knowledge. In: e-BioSphere

Rees J, Cranston K (2017) "Automated assembly of a reference taxonomy for phylogenetic data synthesis." Biodiversity Data Journal 5: e12581. https://doi.org/10.3897/BDJ.5.e12581

Schuettpelz E, Frandsen P, Dikow R, Brown A, Orli S, Peters M, Metallo A, Funk V, Dorr L (2017) Applications of deep convolutional neural networks to digitized natural history collections. Biodiversity Data Journal 5, e21139.

Seregin AP (2016) Making the Russian flora visible: fast digitisation of the Moscow University Herbarium (MW) in 2015. Taxon 65(1), 205-207.

Soltis DE, Soltis PM (2016) Mobilizing and integrating big data in studies of spatial and phylogenetic patterns of biodiversity. Plant Diversity 38(6), 263-270.

Soltis PS (2017) Digitization of herbaria enables novel research. American Journal of Botany 104 (9), 1281-1284

Soltis, P. S., G. Nelson, and S. A. James. 2018. Green digitization: Online botanical collections data answering real-world questions. Applications in Plant Sciences 6(2): e1028

Stefan D, Schmidt M, Zizka G (2014) Introducing African Plants—A Photo Guide—An Interactive Photo Data-Base and Rapid Identification Tool for Continental Africa. Taxon 63(5), 1159-1161

Stein LD (2003) Integrating biological databases. Nature Reviews Genetics 4, 337-345.

Tenopir C, Allard S, Douglass K, Aydinoglu AU, Wu L, Read E, et al. (2011) Data Sharing by Scientists: Practices and Perceptions. PLoS ONE 6(6): e21101. https://doi.org/10.1371/journal.pone.0021101

The vPlants Project (2009). vPlants: A virtual herbarium of the Chicago region. Retrieved December 31, 2009, from www.vplants.org.

Thiers BM, Tulig MC, Watson KA (2016). Digitization of The New York Botanical Garden Herbarium. Brittonia 68(3), 324-333.

Unger J, Merhof D, Renner S (2016) Computer vision applied to herbarium specimens of German trees: Testing the future utility of the millions of herbarium specimen images for automated identification. BMC Evolutionary Biology 16, 248 . doi:10.1186/s12862-016-0827-5.

van Horn G, Mac Aodha O, Song Y, Shepard A, Adam H, Perona P, Belongie S (2017) The iNaturalist Challenge 2017 Dataset. Retrieved from http://arxiv.org/abs/1707.06642

Vilgalys, R (2003). Taxonomic misidentification in public DNA databases. New Phytologist 160, 4-5.

Wäldchen J, Mäder P (2018) Machine learning for image based species identification. Methods in Ecology and Evolution 9, 2216–2225.

Wheeler, T.A. 2004. Bioinformatics and misinformatics: the missing link between taxonomic data and taxonomic databases. Newletter Biol. Surv. Can. 23: 1-6.

Wheeler QD, Raven PH, Wilson EO (2004) Taxonomy: Impediment or expedient? Science 303, 285.

White, R,J. (1984) Implementing small database systems with specialised features. In R. Allkin and F.A. Bisby, eds. Databases in Systematics, London, Academic Press, Systematics Association Special Volume 26, 291-308.

White, R.J., Allkin, R. & Winfield, P.J. (1993) Systematic databases: the Baobab design and the Alice system. In: Fortuner, R., ed., *Advances in Computer Methods for Systematic Biology: Artificial Intelligence, Databases, Computer Vision*, Johns Hopkins University Press, Baltimore, pp. 297-311

Wilson, E.O. 2000. A global biodiversity map. Science 289: 2279.

Wilson, E.O. 2003. The encyclopedia of life. Trends in Ecology Evolution. 18: 77-80.

Younis S, Weiland C, Hoehndorf R, Dressler S, Hickler T, Seeger B, Schmidt M (2018)

Taxon and trait recognition from digitized herbarium specimens using deep convolutional neural networks Botany Letters 1-7, DOI 10.1080/23818107.2018.1446357

Zarucchi, J.L., Winfield, P.J., Polhill, R.M, Hollis, S., Bisby, F.A. and **Allkin, R.** (1993). The ILDIS project on the world's legume species diversity. In "Designs for a Global Plant Species Information system". pp. 131-144. Bisby, F.A., Pankhurst, R.J. & Russell, G.R. (eds). Oxford University Press, Oxford.

Table 1. Online resources pertinent for taxonomy, systematics and phylogenetics.

Resource	**URL**	**Primary Curator(s)**	**Brief Description**
Taxonomic & Nomenclatural Resources			
Plants of the World Online (POWO)	http://www.plantsoftheworldonline.org	Kew Gardens	Names, descriptions, uses, maps and images of plant species

International Plant Names Index (IPNI)	http://www.ipni.org/index.html	Kew Gardens, Harvard University Herbaria, Australia National Herbarium	Database of names and associated basic bibliographical details of seed plants, ferns and lycophytes
Tropicos	http://www.tropicos.org/Home.aspx	Missouri Botanical Garden	Nomenclatural, bibliographic, and specimen data accumulated in MBG's electronic databases during the past 30 years. Includes c. 1.3 million scientific names and c. 4.6 million specimen records
World Checklist of Selected Plant Families (WCSP)	https://wcsp.science.kew.org/home.do	Kew Gardens	Accepted scientific names and synonyms of selected plant families, searchable by name or geographic region
Germplasm Resources Information Network (GRIN)	https://www.ars-grin.gov/npgs/aboutgrin.html	U.S. Department of Agriculture's Agricultural Research Service	Germplasm information about plants, animals, microbes and invertebrates
The Plant List	http://www.theplantlist.org	International consortium of several institutions and projects	Working list of all known plant species. Last updated in 2013. It includes accepted names and synonyms
Catalogue of Life	http://www.catalogueoflife.org	Species 2000 Secretariat, Naturalis	Online database of known species of animals, plants, fungi and microorganisms
ITIS	https://www.itis.gov/	Integrated Taxonomic Information System	Database of taxonomic names of plants, animals, fungi, and microbes of North America and the world
Biodiversity Portals, Specimen Information & Occurrence Data			

Global Biodiversity Information Facility (GBIF)	https://www.gbif.org	International initiative governed by the GBIF Secretariat based in Copenhagen	Aggregator of more than a billion occurrences, including specimens, observations, event-based datasets and checklists
Atlas of Living Australia (ALA)	https://www.ala.org.au	CSIRO	Collaborative, open infrastructure national Australian biodiversity project that aggregates biodiversity data from multiple sources
Ocean Biogeographic Information System (OBIS)	http://www.iobis.org	International Oceanographic Data and Information Exchange (IODE) of UNESCO	Data and information on marine biodiversity
Distributed System of Scientific Collections (DiSSCo)	http://dissco.eu	European natural history collections	A new European programme that aims to to facilitate mobilisation and publication of biodiversity data housed in European natural history collections
Morphological, Trait and Genetic Data			
NCBI	https://www.ncbi.nlm.nih.gov/	National Center for Biotechnology Information	Nucleotide, protein, genome, reference sequences and others.

	https://www.ebi.ac.uk/ena	European Nucleotide Archive	Nucleotide sequence
	https://www.ddbj.nig.ac.jp/index-e.html	DNA Databank of Japan	Nucleotide sequence
	http://www.boldsystems.org	Barcode of Life Data Systems	Barcode sequence
MorphoBank	https://morphobank.org	Stony Brook University	"House, save and share information on anatomy, physiology, behavior and other features of species"
TRY Plant Trait Database	http://try-db.org	Future Earth; Max Plank Institute for Biogeochemestry	Curated repository for plant morphological, anatomical, biochemical, physiological or phenological traits
IPCN	http://www.tropicos.org/Project/IPCN	Index to Plant Chromosome Numbers	List of chromosome counts by taxon
KNApSAcK	http://kanaya.naist.jp/knapsack_jsp/top.html	KNApSAcK Core System	System integrating Mass Spectrometer peaks, molecular weight and molecular formula of metabolites by species
Images			
iNaturalist	http://www.inaturalist.org	California Academy of Sciences	Citizen science portal that collates and publishes biodiversity observations and identifications

Global Plants JSTOR	https://plants.jstor.org/	JSTOR	Images of type specimens from >300 herbaria globally
Leafsnap	http://leafsnap.com/species/	Columbia University, the University of Maryland, and the Smithsonian Institution	Citizen science-based mobile app that helps to identify trees found in the Northeastern United States and Canada
African Plants-A photo guide	www.africanplants.senckenberg.de		A photo guide of African plant species (except Madagascar)
Live Plant Photos	https://plantidtools.fieldmuseum.org/en/nlp/5304		The Field Museum's Plant Identification Tools website and it covers plants specimens from Central and South America
Phylogenetic Information			
TreeBASE	https://www.treebase.org		
DRYAD	https://datadryad.org		
Open Tree of Life (OTOL)	https://tree.opentreeoflife.org		
Phylolink	https://www.biodiversityintelligence.com/phylolink)		
Figshare	https://figshare.com		Users get 20 GB of free private storage with DOI for the projects

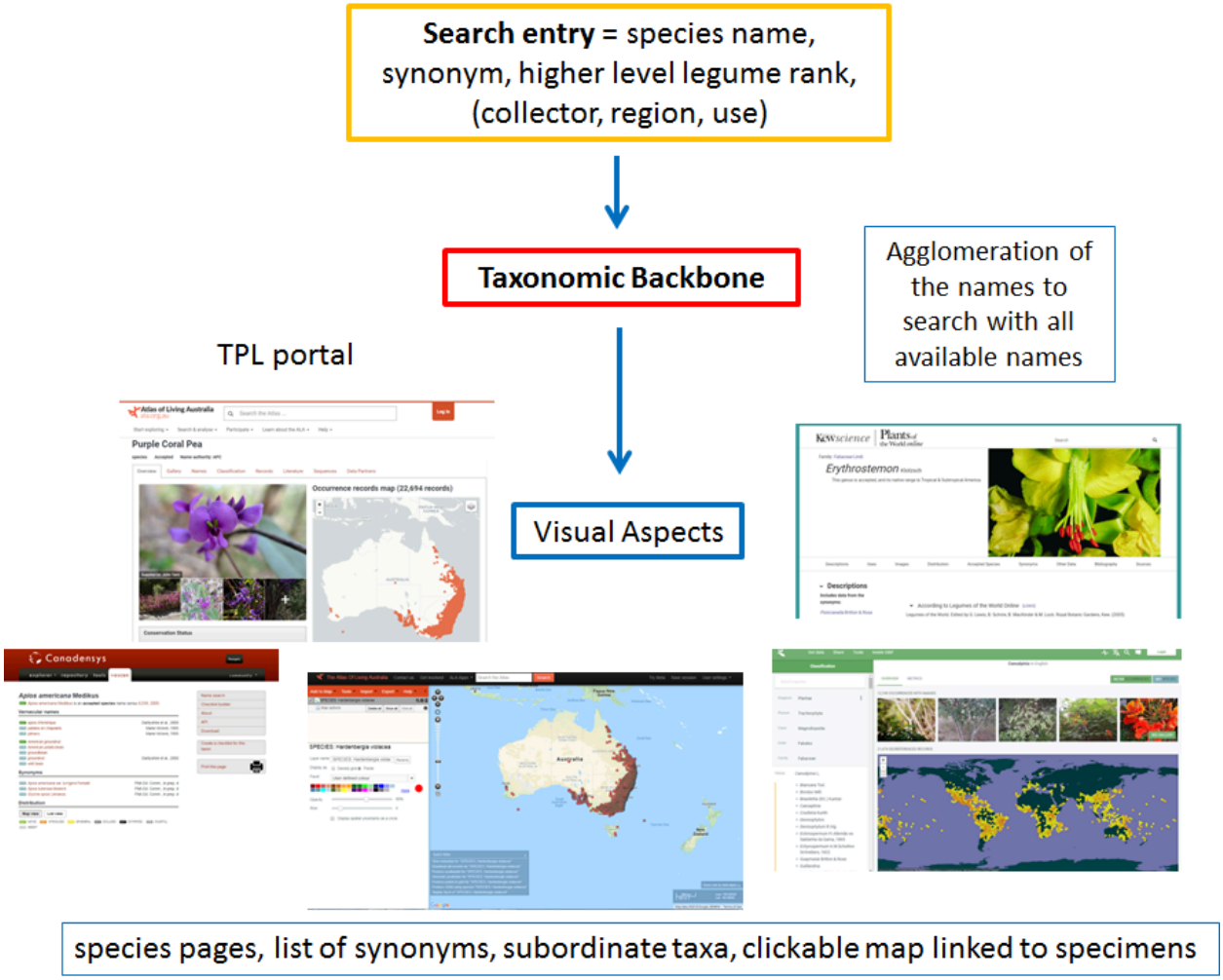

Fig. 1 - Legume Systematics Portal: Front end showing visual aspects of portal and search entry point

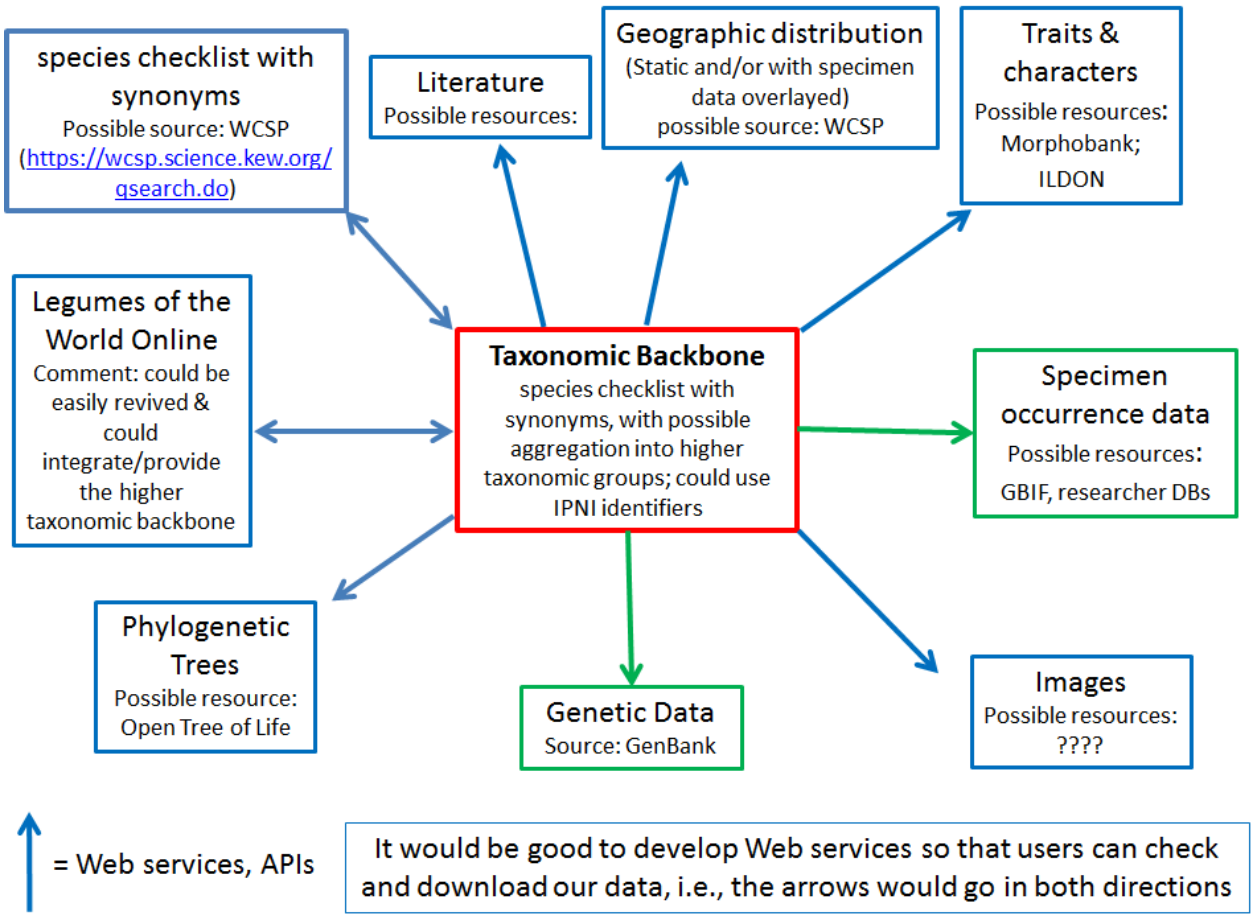

Fig. 2 - Legume Systematics Portal: Links to external data sources and portals. Green boxes are dynamically updated (with the latest updated version visible), blue boxes are updated periodically (how often?). Arrows represent APIs and Webservices

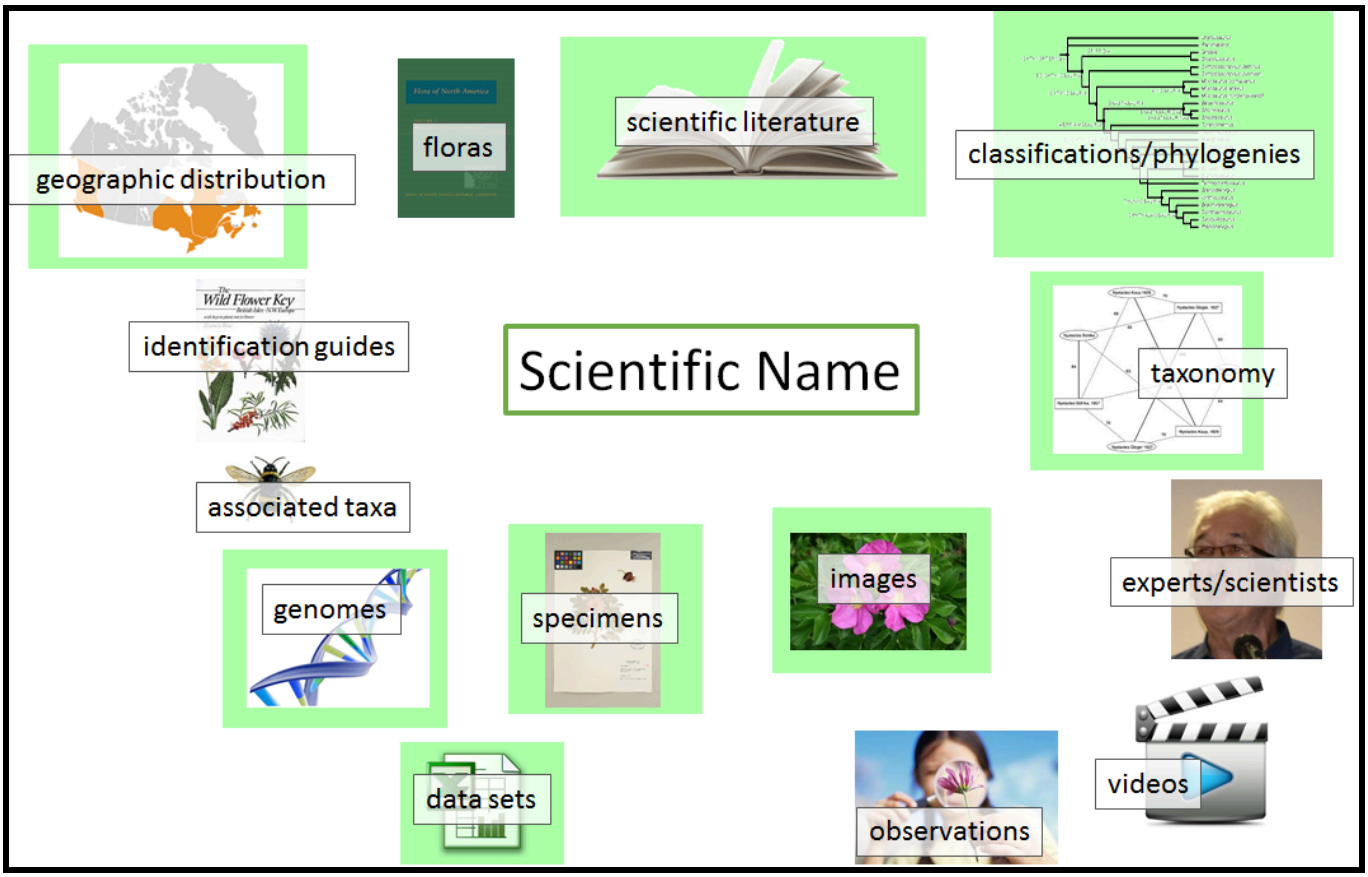

Figure 2 (version 1, but see version above). The scientific name serves as the central point of reference for linking diverse sources of information available from external (or not) data sources. Initially we might want to focus on the resources in green, but others could be added in the future depending on needs and availability.

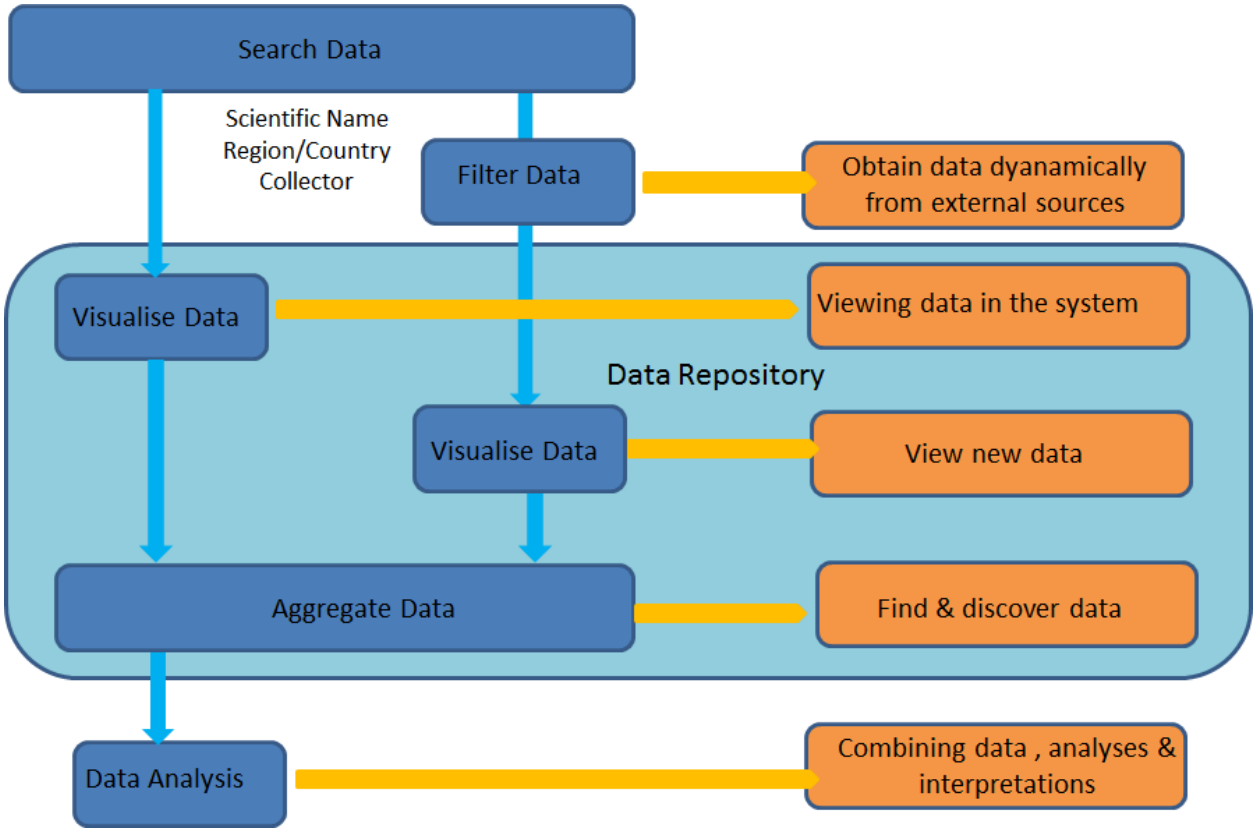

Fig. 3 - Legume Systematics Portal: Components, tools and functions (Back end showing informatics tools and structure required/suggested)