# VirtHerb - a simple webscraper to collect information on plants from online databases

# Developed by Dr. Matteo Jucker Riva, matteo.jriva@gmail.com

## Intro

information about plants, especially wild species, are scattered across multiple sources and websites, providing unstructured information in different ways and templates. Researchers or interested people often have to manually scan diffe4rten website, verify that the plant exists and copy the relevant information on a separate document and manually merge it. To add to that, most information is actually repeated across the different websites, making the final text redundant and unclear

The aim of ***VirtHerb*** is to collect and integrate information about plants by scraping different databases and web sources and combine the resulting text in a structured pdf.

Workflow of ***VirtHerb***:

1- Scrape the web using a plant scientific name (provided by user).

2- Structure text in sections, optionally together with user defined notes and pictures.

3- Output a simple 1 or 2 page pdf with essential information to be downloaded or shared.

***VirtHerb*** currently scrapes the following sources:

- CABI Invasive species compendium <https://www.cabi.org/ISC>

- World Flora Online <http://www.worldfloraonline.org/>

- JSTOR Global Plants <https://plants.jstor.org/>

- Wikipedia [www.wikipedia.org](http://www.wikipedia.org/)

12.2020 CURRENTLY UNDER DEVELOPMENT

With the integration of a machine learning model for extractive summarization, we aim to extract the most relevant summary sources, reducing repetition and redundant information.

## Text example from the different sources

*See below an example of the text related to one plant across different sources*

User input (plant scientific name): ***Parthenium histerophorus***

### **source 1 (CABI Invasive species compendium)**

Direct link: https://www.cabi.org/isc/datasheet/45573

### Section 1 - Description

Parthenium hysterophorus is an erect, branched, aromatic, annual (or a short-lived perennial),herbaceous plant with a deep taproot. The species reproduces by seed and has vigorous growth. In its neotropical range it grows to 30-90 cm in height (Lorenzi, 1982; Kissmann and Groth, 1992), but up to 1.5 m, or even 2.5 m, in exotic situations (Haseler, 1976; Navie et al., 1996). Shortly after germination the young plant forms a basal rosette of pale green, pubescent, strongly dissected, deeply lobed leaves, 8-20 cm in length and 4-8 cm in width. The rosette stage may persist for considerable periods during unfavourable conditions (such as water or cold stress). As the stem elongates, smaller, narrower and less dissected leaves are produced alternately on the pubescent, rigid, angular, longitudinally-grooved stem, which becomes woody with age. Both leaves and stems are covered with short, soft trichomes, of which four types have been recognized and are considered to be of taxonomic importance within the genus (Kohli and Rani, 1994).

Flower heads are both terminal and axillary, pedunculate and slightly hairy, being composed of many florets formed into small white capitula, 3-5 mm in diameter. Each head consists of five fertile ray florets (sometimes six, seven or eight) and about 40 male disc florets. The first capitulum forms in the terminal leaf axil, with subsequent capitula occurring progressively down the stem on lateral branches arising from the axils of the lower leaves. Thousands of inflorescences, forming in branched clusters, may be produced at the apex of the plant during the season. Seeds (achenes) are black, flattened, about 2 mm long, each with two thin, straw-coloured, spathulate appendages (sterile florets) at the apex which act as air sacs and aid dispersal.

### Section 2 - Ecology

The chromosome number for P. hysterophorus has been reported as 2n=18 in India (Hakoo, 1963) and Australia (Navie et al., 1996); however, both lower (2n=9) and higher (2n=34) chromosome numbers have also been attributed to this species (Kohli and Rani, 1994). Germplasm collections are available at various institutions (Kew Royal Botanic Gardens, 2018; USDA-ARS, 2018). DNA information is available at the Barcode of Life Data System (BOLDS, 2018).

Two distinct races of P. hysterophorus have been identified, the 'South American' and the 'North American' races (Dale, 1981), the former with cream to yellow flowers and the latter with white flowers. Moreover, Towers et al. (1977) indicated that the sesquiterpene lactone, hymenin, which is present in plants from Argentina and Bolivia, is different from the lactone, parthenin, identified from most of the samples collected in India, as well as North and Central America. The detailed genetic comparison between plants from across the introduced range is necessary to ascertain the history of its spread and to develop management programmes.

The species has two biotypes in Australia, which are distinctive based on their introduction, the colonization ability, and their demographic spread (Adkins et al., 1997; Bajwa et al., 2018). The Toogoolawah biotype is confined to less than 10 kilometers from its point of origin. According to Hanif et al. (2011) the Toogoolawah type has a tendency to self-pollination. The Clermont biotype is invasive and has spread into more than 520,522 km2. The two biotypes also have significant genetic and morphological differences. The Clermont biotype has a higher germination rate across different environmental conditions compared with the Toogoolawah biotype, which might be contributing to its high invasive ability (Bajwa et al., 2018). Parker (1989) identified two biotypes with different flowering patterns in Mexico.

Parthenium hysterophorus reproduces by seeds. It is unable to reproduce vegetatively from plant parts or by apomixis, but is a prolific seed producer (15,000-25,000 achenes per plant on average, and up to 100,000 in large plants) (Haseler, 1976; Navie et al., 1996; Mahadevappa, 1997; Gnanavel, 2013), and continues to flower and fruit until senescence. Pollen grains are spheroidal, 15 to 20 μm in size, and have short to medium length spines with an average of 168,192 pollen grains produced in each capitulum (Lewis et al., 1988), adding up to about 624 million per plant (Gnanavel, 2013).

Most seeds germinate within two years if the conditions are suitable (Gnanavel, 2013). The longevity of surface-lying seeds seems to be short with little or no dormancy, but there is evidence that buried achenes can remain viable for at least 4-6 years (Navie et al., 1996). Navie et al. (1998) estimated the half-life of buried seed to be about 6 years. Gnanavel (2013) reports a longer viability for buried seeds, being able to germinate after 8-10 years. Tamado et al. (2002) report that the viability of the seeds was greater than 50% after 26 months of burial in the soil, indicating the potential build-up of a substantial and persistent soil seed bank. Germination occurrs at the mean minimum (10°C) and maximum (25°C) temperatures, as well as over a wide range of fluctuating temperatures (12/2°C - 35/25°C) in light (Tamado et al., 2002). An optimum germination temperature of 22-25°C is reported by Gnanavel (2013). It appears that the species can germinate without light, but a 12 hours photoperiod is ideal for its maximum germination (Bajwa et al., 2018). Seed germination can occur over a wide range of soil pH (2.5-10), with an optimum of 5.5-7.0 (Parsons and Cuthbertson, 1992). It can also germinate under high osmotic and salt stress (Bajwa et al., 2018). Germination may be increased after cold stratification, and with exposure to light (Karlsson et al., 2008). Flowering may begin as early as 4 weeks after seedling emergence (Jayachandra, 1971), and plants continue to flower for extended periods (6-8 months) when conditions permit. Under favourable conditions of adequate moisture and bare soil, four or five generations per year can be completed (Gnanavel, 2013).

Parthenium hysterophorus is an aggressive colonizer of disturbed ground, able to germinate, grow and flower over a wide range of temperatures and photoperiods. Seeds germinate all year-round provided moisture is available and germination rate is extremely high. Four or more successive cohorts of seedlings may be produced in a season (Pandey and Dubey, 1989).

In Australia, P. hysterophorus germinates mainly in spring and early summer. It produces flowers and seeds throughout its life and dies in late autumn (Navie et al., 1996). It can grow at any time of the year as long as there is moisture (Tamado, 2001; Taye, 2002).

Plants emerging during the first (spring) rains usually attain a greater size and have a significantly longer lifespan than those produced in the summer. Soil moisture appears to be the major contributing factor to both the lifespan and to duration of flowering. Plant biomass production increases with increasing temperature up to an optimum day/night temperature regime of 33/22°C (Williams and Groves, 1980). Under unfavourable (dry) conditions, the life cycle may take up to 335 days, compared to 86 days under optimum conditions. Physiological studies have shown that P. hysterophorus has a low photorespiratory activity and has the C3 photosynthetic pathway but with positive C4 tendencies (Patil and Hegde, 1983).

The species also produces chemical compounds that when released in the soil have allelopathic effects on other species (Rubaba et al., 2017).

In Mexico P. hysterophorus is reported as being associated with Bursera and arborescent Ipomoea species (CONABIO, 2018).

Parthenium hysterophorus is able to germinate, grow and flower over a wide range of temperatures and photoperiods. It occurs in the humid and sub-humid tropics, showing a marked preference for black, alkaline, cracking, clay soils of high fertility, but is able to grow on wide variety of soil types from sea level up to 2500 m (Evans, 1987a; Taye, 2002). In Ethiopia, it grows from low to high-mid-altitude areas at 900-2500 m asl (Taye, 2002). High clay content in soils prolonged the rosette stage, enhanced relative growth rates in height and diameter, and hampered root growth, but promoted biomass allocation to shoots (Annapurna and Singh, 2003). Mahadevappa (1997) noted that P. hysterophorus has several built-in properties and efficient behavioural mechanisms that enable it to overcome many ecological adversities and thus continue to survive under stress. The species is not frost tolerant, which is considered as a limiting factor to its distribution in temperate areas (Gnanavel, 2013). It grows best in subtropical regions with mean annual temperatures ranging from 10-25°C and an annual rainfall exceeding 500 mm. Areas receiving less than 500 mm of rainfall are probably unsuitable, although the weed has strong adaptive methods to tolerate both moisture stress (Kohli and Rani, 1994) and saline conditions (Hegde and Patil, 1982). The weed finds access to any type of land but it is especially prolific in disturbed habitats, for example, roadsides and railway tracks, stock-yards, around buildings and on waste land, from where it spreads and invades agricultural systems.

According to Dale (1981), the distribution of P. hysterophorus in Queensland is controlled by factors similar to those limiting the plant in its areas of origin. However, because of differences in land management, soils and climate, the plant covers much greater areas and is a far more significant problem in Australia. The combination of neutral to alkaline clays and absence of competing vegetation provided ideal conditions for the development of large stands of P. hysterophorus. The plant is more common on roadsides in Queensland than in North America, but this can be attributed to more regular disturbance. In natural grasslands, the situation is similar in the two areas, with P. hysterophorus becoming dominant only in the most overgrazed situations. The plant appeared more prominent in cultivated areas in North America, particularly during fallow periods, but in neither area does it significantly affect crop production. In all regions, the densest stands produce a complete ground cover with no other species present.

### Section 3 :Uses

Like many other weeds, P. hysterophorus also has beneficial aspects. Ramaswami (1997) and Seier and Djeddour (2000) reviewed the potential uses of parthenium weed. The application of P. hysterophorus compost and green leaf manure was reported to lower weed populations in rice. This was due to the role of allelopathic compounds present in it (Sudhakar, 1984), an increase in soil moisture due to the build-up of soil organic carbon (Son, 1995), increased soil N, P and K content (Bharati et al., 2001), and a reduced incidence of pests in rice such as stem borers and leaf rollers (Ramaswami, 1997). P. hysterophorus is also potentially a rich source of potash (Parsons and Cuthbertson, 1992).

The allelopathic substances present in P. hysterophorus have been proposed as a source of insecticide (Parsons and Cuthbertson, 1992; Hiremath and Ahn, 1997), herbicide (Mersie and Singh, 1987; Pandy et al., 1993; Batish et al., 2002), fungicide (Ganeshan and Jayachandra, 1993), and nematicide (Azam et al. 2001; Prasad et al., 2002). Dwivedi et al. (2000) and Sharma et al. (2003) reported that P. hysterophorus extract has 95% repellency and oviposition deterrent properties against Callosobruchus chinensis in chick pea grains. A larval mortality of greater than 50% has been recorded in 2% concentration for Helicoverpa armigera (Sundararajan, 2002) and a significant decrease in life span and progeny production of the mustard aphid, Lipaphis erysimi, was reported (Sohal et al., 2002). Extract of P. hysterophorus was also found to significantly inhibit the growth of bacterial spot pathogen (Xanthomonas axonopodis pv. vesicatoria) infecting Capsicum frutescens (Sree and Sreeramulu, 2002).

Additional uses include as a foliar supplementation of the leaf water extract of P. hysterophorus on mulberry leaves stimulated silkworms to feed and utilize them more efficiently, resulting in vigorous growth of larvae, pupa cocoons and silk yield (Patil, 1997; Singhal et al., 1998). Production of oxalic acid (Mane et al., 1986) and biogas (Gunaseelan, 1987; Abubacker et al., 1999; Thakur and Singh, 2000, 2003) from parthenium has been reported. In India, it is considered to be an exploitable source of easily extractable, high quality protein for stock feed (Savangikar and Joshi, 1978).

Parthenium hysterophorus is an annual herb used in its native neotropics as a herbal remedy for various intestinal and skin disorders using a decoction of boiled roots (Dominguez and Sierra, 1970). There are several reports on the antiviral, antifungal, antibacterial, antihelmintic and anti-inflammatory properties of the species (Joshi et al., 2016). It is used for the treatment of wounds, ulcerated sores, fever, anaemia, heart problems and malaria. It has potential medicinal properties due to antitumor (Mew et al., 1982) and antiamoebic activities (Sharma and Bhutani, 1988). It is also reported as used for toothaches, to remove boils and pimples, to relieve constipation, fevers, against insomnia and to treat diabetes (Joshi et al., 2016).

The dry leaf powder of P. hysterophorus causes wilting of Salvinia molesta, which is an invasive species in some water bodies (Joshi et al., 2016). The species can be used for bioremediation of lead and nickel contaminated soils and water sources. The parthenin produced by the species has allelopathic effects, but also can stimulate growth of other plant species at subtoxic doses (Belz, 2016).

### source 2 (World Flora Online)

direct link: <http://www.worldfloraonline.org/taxon/wfo-0000065156;jsessionid=555BDFCA5AC35C6C605AADF418C54332>

**Section 1 – General information**

Herbs, annual, 30-120 cm tall. Leaf blade ovate to elliptic, 3-18 × 1-5(-9) cm, pinnately (1 or)2-lobed, ultimate lobes lanceolate to linear, 3-50 × 2-15 mm, both surfaces sparsely to densely scaberulose and gland-dotted. Synflorescences of open panicles. Capitula obscurely radiate; peduncles 1-8(-15+) mm; outer phyllaries 5(or 6), elliptic-lanceolate, 2-4 mm, inner 5(or 6) ovate to orbicular, 2.5-4 mm. Female florets 5(or 6); corolla limbs reniform or orbicular to oblong, 0.3-1 mm. Disk florets 12-30(-60). Achenes obovoid, 1.5-2(-3.5) mm; pappuslike enations erect, deltate to ovate, 0.5-1 mm. Fl. Apr-Aug. 2n = 34.

Annuals, (10–)30–120+ cm. Leaf blades ovate to elliptic, 30–180+ × 10–50(–90+) mm, (1–)2-pinnately lobed (ultimate lobes lanceolate to linear, 3–50 × 2–15 mm), faces sparsely to densely scabrellous and gland-dotted (seldom with additional erect hairs 1–2 mm). Heads obscurely radiate, borne in open, paniculiform arrays. Peduncles 1–8(–15+) mm. Phyllaries: outer 5(–6), lance-elliptic, 2–4 mm, inner 5(–6) ovate to ± orbiculate, 2.5–4 mm. Pistillate florets 5(–6); corolla laminae reniform or orbiculate to oblong, 0.3–1 mm. Disc florets 12–30[–60]. Cypselae obovoid, 1.5–2(–3.5) mm; pappus-like enations erect, deltate to ovate, 0.5–1 mm (sometimes a third, subulate spur near apex adaxially). 2n = 34.

Annual herb to -50 cm tall; stems erect, much branched, to 4 mm diam., strigillose with hairs 0.3 mm long. Leaves lanceolate above to pinnately dissected below, 2.5-8 cm long, 0.2-4.5 cm wide, acute, the base attenuate, both surfaces strigillose with hairs 0.2-0.4 mm long, the margin entire or deeply lobed. Inflorescence paniculiform, of numerous small heads; peduncles 5-8 mm long. Heads radiate, 2.5-3 mm tall; outer involucre cupulate, 4-5 mm diam., the bracts 5, separate, narrowly ovate, 1.6-2.2 mm long, 1.5-1.8 mm wide, acute, abaxial surface strigose with hairs 0.2 mm long, the margin herbaceous, the inner bracts 5, uniseriate, each attached abaxially to the base of a ray floret, two disc paleas and two included sterile disc florets also attached to the base of the ray floret on the adaxial side (the whole unit termed an "achene complex" (Fig. 56C), each of the two disc florets and enclosing paleas is additionally attached from the base to near the apex of the achene by a thin rib); paleas elliptic, 1 mm long, 0.3 mm wide, apically fimbriate; ray florets 5, the ligules light yellow, ovate, 0.6 mm long, 0.5 mm wide, emarginate; disc florets perfect, 25-50, the corollas light yellow, narrowly funnelform, 0.3 mm diam., the throat 1 mm long, the ovary sterile, filiform, 0.7 mm long. Achenes flattened, 2 mm long, 1.2 mm diam.; pappus of two broad awnis 0.5 mm long. Chromosome number n= 17.

"Annual to 1 m, usually much branched, ± hairy, and often glandular above; lvs pinnatifid or usually bipinnatifid, to 2 × 1 dm; heads small, numerous in an often leafy infl, the disk 3–5 mm wide; achenes obovate, black, 2–2.5 mm; 2n=34. A weed of waste places; trop. Amer., n. to se. U.S. and occasionally to Mass., Mich., and Kans. July–Oct."

**Section 2:Morphology**

Leaves elliptic to ovate, bipinnately lobed, 6–30 cm long, 3–10 cm wide, lobes 1.5–4 mm wide and obtuse, puberulous on both surfaces.

Annual herb to 60 cm high; stems puberulous or scabridulous.

Ray florets 5, whitish, almost round, 0.5–1 mm; disc florets yellowish, narrowly funnel-shaped, ± 2.7 mm long.

Capitula hemispherical, in lax panicles; involucre 2–3 mm long, the phyllaries pubescent.

Achenes black, obovoid, ± 2 mm long, with 2 recurved pappus awns.

Annual herb to 60 cm high; stems puberulous or scabridulous. Leaves elliptic to ovate, bipinnately lobed, 6-30 cm long, 3-10 cm wide, lobes 1.5-4.0 mm wide and obtuse, puberulous on both surfaces. Capitula hemispherical, in lax panicles; involucre 2-3 mm long, the phyllaries pubescent. Ray florets 5, whitish, almost round, 0.5-1.0 mm; disc florets yellowish, narrowly funnel-shaped, ±2.7 mm long. Achenes black, obovoid, ±2 mm long, with 2 recurved pappus awns.

Annual herb, up to 0.6 m high. Leaves deeply bipinnatisect, puberulous, decreasing in size upwards. Heads 3 mm in diameter, on slender peduncles, in lax, many-headed panicle. Flowers with (small) white rays and yellowish disc.

### Source 3 (Wikipedia)

Direct link: https://en.wikipedia.org/wiki/Parthenium\_hysterophorus

**Section 1 – General description**

Parthenium hysterophorus is a species of flowering plant in the aster family, Asteraceae. It is native to the American tropics.[1] Common names include Santa-Maria,[2] Santa Maria feverfew,[3] whitetop weed,[4] and famine weed.[5] In India, it is locally known as carrot grass, congress grass or Gajar Ghas.[6] It is a common invasive species[7] in India,[8] Australia, and parts of Africa.

The pollen grains of Parthenium hysterophorus.

The pollen grains of Parthenium hysterophorus invades disturbed land, including roadsides. It infests pastures and farmland, causing often disastrous loss of yield, as reflected in common names such as famine weed.[9][10][11][12][13] In some areas, heavy outbreaks have been ubiquitous, affecting livestock and crop production, and human health.[14][15] As an invader it first appeared as a contaminant in imported wheat.[16]

The plant produces allelopathic chemicals that suppress crop and pasture plants, and allergens that affect humans and livestock. It also frequently causes pollen allergies.[17]

It is being investigated as a means of removing heavy metals and dyes from the environment, control of aquatic weeds, commercial enzyme production, an additive in manure for biogas production, as a biopesticide, and as green manure and compost.[18]