

Article

Match Algorithms for Scientific Names in FlorItaly, the Portal to the Flora of Italy

Matteo Conti, Pier Luigi Nimis and Stefano Martellos * 

Department of Life Sciences, University of Trieste, via L. Giorgieri 10, I-34127 Trieste, Italy; matt.ciao@gmail.com (M.C.); nimis@units.it (P.L.N.)

* Correspondence: martelst@units.it

Abstract: Scientific names are not part of everyday language in any modern country, and their input as strings in a query system can be easily associated with typographical errors. While globally unique identifiers univocally address a taxon name, they can hardly be used for querying a database manually. Thus, matching algorithms are often used to overcome misspelled names in query systems in several data repositories worldwide. In order to improve users' experience in the use of FlorItaly, the Portal to the Flora of Italy, a near match algorithm to resolve misspelled scientific names has been integrated in the query systems. In addition, a novel tool in FlorItaly, capable of rapidly aligning any list of names to the nomenclatural backbone provided by the national checklists, has been developed. This manuscript aims at describing the potential of these new tools.

Keywords: checklist; database; nomenclature; synonym; taxon



Citation: Conti, M.; Nimis, P.L.; Martellos, S. Match Algorithms for Scientific Names in FlorItaly, the Portal to the Flora of Italy. *Plants* **2021**, *10*, 974. <https://doi.org/10.3390/plants10050974>

Academic Editor: Filippo Maggi

Received: 30 March 2021

Accepted: 8 May 2021

Published: 13 May 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Nomina si nescis, perit et cognitio rerum. If we ignore the names, we also lose the knowledge of things. This sentence by Linnaeus [1] highlights the relevance of names in biology (as in any human interaction). Scientific names provide access to a wealth of information collected during centuries of research [2]. For accessing the information on a taxon, it is necessary to know its currently accepted name. However, since any binomial—via the genus name—expresses a hypothesis on the phylogenetic affinities of a given taxon, scientific names do change over time, as our knowledge progresses [3–5]. While globally unique identifiers (GUIDs [6]), as the LSIDs [7] can be used to univocally address a taxon name, are useful in web services which allow machine–machine communication, they can hardly be used for querying a database manually. On the other hand, the use of taxon names for querying a database system could lead to several critical issues [8,9]. Among them, synonymies were once a problem, which is, however, at least partly addressed by means of nomenclatural thesauri, which can be invoked for retrieving the accepted name and all its synonyms (e.g., The Catalogue of Life [10]). Today, all modern online checklists have their own nomenclatural thesaurus or are connected to a web service or tool which can at least resolve synonyms [11]. However, another major issue is due to the fact that a taxon name is a string of text (in Latin), so that typing it into a query system can easily lead to typographical errors. Such an issue is normally addressed by means of matching algorithms (e.g., Taxamatch [12]), which find the closest match to any input string with one or more typos.

- Match algorithms applied on taxon names are normally made of two components: A parser, which divides the string into its components (i.e., genus, species, authority, etc.). Some relevant examples are: gnparser [13], the name parser of GBIF [14], and the REGEXP parser [15].
- A matcher, which compares the output of the parser with a thesaurus of names. Among matchers, some are based upon phonetic similarity (such as Soundex [16]),

while others use metrics for measuring orthographic differences among strings, such as the Levenshtein distance. An interesting example is the Taxamatch algorithm [17], which combines the two approaches.

Such algorithms are currently used in several data repositories, such as the GBIF (parser: GBIF name parser, matcher: Taxamatch), the iPlant Taxonomic Name Resolution Service (parser: gnparser, matcher: a modified version of Taxamatch), Euro+Med (parser: homebrew, matcher: a modified version of Taxamatch), WoRMS (parser: gnparser, matcher: a modified version of Taxamatch), and BiOnym (parser: REGEXP, matcher: several combined algorithms [15]).

As far as the vascular flora of Italy is concerned, FlorItaly, the Portal to the Flora of Italy (<http://dryades.units.it/floritaly>; accessed on 29 March 2021) [18], is built on the basis of the checklists by Bartolucci et al. [19] and Galasso et al. [20] and their subsequent updates, which are published twice per year [21–32]. FlorItaly hosts a rich thesaurus of names, and each query is parsed through it in order to provide all relevant results. In order to provide further help to users, it was decided to enrich FlorItaly with a novel tool, capable of retrieving a list of accepted names for any list of scientific names (input as a plain text or CSV file), deriving, e.g., from old literature, phytosociological relevés, lists of herbarium samples, image archives, or any existing database containing the names of plants occurring in Italy. This new tool, that will permit to rapidly align any list of names to the nomenclatural backbone provided by the national checklists, was envisaged as a relevant step towards the aggregation into FlorItaly of other resources, such as digitized herbaria, image archives, databases of vegetation relevés, etc. A near match algorithm was also adopted to support the query system of FlorItaly in order to resolve misspelled scientific names. This algorithm was included in the three query systems of FlorItaly (basic, standard, and advanced), which now warn users if a misspelled name was input and provide results for the closest matching string in the thesaurus of scientific names.

This paper aims at describing the algorithms which have been integrated into the Portal to the Flora of Italy, detailing their logic, aims, and uses.

2. Materials and Methods

All software was written in PHP 5.6 (but is compatible with PHP 7.x) and operates on a MySQL database. The reference database (as on 28th February 2021) contains 10,928 accepted names and 13,300 synonyms. All names in the reference database are written as whole strings, i.e., they are not split into elements (genus, species, authority, subspecies, etc.).

2.1. Data Preparation

Accepted names and synonyms are used in their original form (a single string) by the algorithm integrated in the query interfaces of FlorItaly (see below), while they are used in a parsed form by the algorithm which performs list matches. Each time the checklist is updated (see Martellos et al. [11]), all names are automatically parsed and stored in their parsed form in a data table, together with a phonetic equivalent for each epithet (generated by using a modified version of an algorithm developed by Rees in 2007 and included into Taxamatch [17]). As an example, the name *Huperzia selago* (L.) Bernh. ex Schrank & Mart. subsp. *selago* is parsed into the following components: *Huperzia* (genus), HIPIRSA (genus, phonetical), *selago* (species), SILAGA (species, phonetical), L. Bernh. ex Schrank & Mart. (species authority), subs. (infraspecific rank) *selago* (infraspecific rank), and SILAGA (infraspecific rank, phonetical).

2.2. The Algorithms

Two algorithms were developed, aiming at matching an input string or a list of taxa to the reference database, either returning the closest match (in the case of the algorithm integrated in the three query interfaces) or a list of accepted names (in the case of the algorithm operating on lists of names).

2.2.1. Input String

The algorithm that is integrated in the query interfaces operates by comparing the input string to accepted names and synonyms in the reference checklist. If no exact match is returned, it continues by searching for the closest match, i.e., the string in accepted names or synonyms which is closer to the input string in terms of the minimum number of substitutions, removals, or insertions. This algorithm operates by two major constraints: (1) it allows matching strings which are longer or shorter than the original one by 1 character, i.e., with one insertion or removal only, and (2) it never inserts a blank space, nor replaces a blank space in the input string.

2.2.2. Lists of Taxa

The algorithm operating on a list of taxa follows four steps: normalization, pre-processing, parsing, and matching.

(1) Normalization

In the normalization step, the input string is elaborated as follows: (a) special characters (question marks, exclamation marks, hashtags, underscores, quotes, and double quotes) are removed; (b) numbers—if present—are removed (the reference database does not contain years associated to authorities); (c) letters with diacritical marks are replaced with their ASCII equivalents; (d) a blank space (if not present) is added after each dot; (e) indications of uncertain identification such as “cf.” and “aff.” are removed; and (f) double blank spaces, along with leading and trailing blank spaces, are removed.

(2) Pre-Processing

In the pre-processing step, the normalized string is compared to the reference database. If there is an exact match, it is returned with a maximum score (100%).

(3) Parsing

In the parsing step, each element of the input string is identified, separated, and classified. As an example, the string “*Achillea millefolium* L. subsp. *sudetica* (Opiz) Oborny” is parsed into genus (“*Achillea*”), species epithet (“*millefolium*”), species authority (“L.”), infraspecific rank (“subsp.”), infraspecific epithet (“*sudetica*”), and infraspecific epithet authority (“(Opiz) Oborny”). The algorithm makes use of a set of rules that exploit the properties of PHP arrays, along with REGEX (Regular Expressions), to identify the different elements of the input string. The algorithm is case-sensitive and takes into account the combination of uppercase and lowercase letters to correctly identify the elements of a taxon name. However, it can operate in case-insensitive mode if the input string is entirely uppercase or lowercase or when “force case insensitive parsing” is selected among the advanced options by the user.

To cope with issues related to the presence in the authors’ names of strings which are the same as those used as infraspecific rank indicators, a set of REGEX rules were implemented. If even this verification process fails, the ambiguous string (“v.”, “f.”, “s.”, or “c.”) is considered as the rank indicator of the infraspecific taxon. Thus, the result of the parsing algorithm will be flagged as potentially wrong, and another verification step is applied after the matching step (see below).

(4) Matching

In this step, the parsed string is matched against the parsed reference database using a home-brew version of Taxamatch, which makes use of a modified Damerau–Levenshtein distance algorithm, together with a phonetic algorithm. The Levenshtein distance algorithm calculates the minimum number of deletions, insertions, or substitutions of a single character needed to transform a string into a second string, returning an edit distance value (ED). As an example, “*milefolius*” and “*millefolium*” have an ED of 2 (one insertion and one substitution). The modified Damerau–Levenshtein distance (MDLD) allows even

multiple character transpositions at the cost of the number of transposed characters. As an example, “*canadsiens*” and “*canadensis*” have an ED of 2 (two characters’ transposition) instead of 4 (four characters’ substitution) as per the Levenshtein distance algorithm. Then, the phonetic algorithm transforms the input string elaborated by the MDLD algorithm into a simplified phonetic equivalent. If this matches the phonetic equivalent of one name in the reference database, they are considered as phonetic matches. Thus, every parsed element of the input string (except the authority and the infraspecific rank indicators) is matched against the reference database orthographically and phonetically [17]. To be considered as a match, a name in the reference database must pass at least one of the two tests.

Following the infraspecific epithet match, if the input name is flagged as potentially wrongly parsed (see (3) in Section 2.2.2), and the best match has an infraspecific epithet $ED \geq 2$ (or there are no matches at the infraspecific epithet level), the algorithm foresees a further verification. If (a) the string leading to a parsing error is present in at least one of the taxon name authorities in the reference database which passed the species epithet match, and (b) both in the input string and in the taxon name authority of the reference database, the word following the one that can lead to a parsing error is the same, or begins with the same letter, and has an $ED = 1$ (calculated using the MDLD), then the string that can lead to a parsing error is considered as part of the authority, and the portion of the input string following the specific epithet is parsed for a second time. Otherwise, no changes are applied. Authors’ names are matched against the reference database using an *ngram* algorithm [17], returning a similarity score between 0 and 1.

In the matching step, a *ComponentScore* ranging from 0 to 1 is assigned to each parsed portion of the input string.

$$ComponentScore = 1 - (ED / \max ED) \quad (1)$$

ED is the editing distance calculated with the MDLD, while $\max ED$ is the maximum possible value of ED that is equal to the length of the longest of the two strings which are matched (the input and the reference).

As far as infraspecific epithets are concerned, the *ComponentScore* is decreased by 0.3 when the rank indicator is not the correct one. As an example, in the case of *Achillea millefolium* L. subsp. *millefolium*, if the input string is *Achillea millefolium* L. var. *millefolium*, the infraspecific taxon will receive a *ComponentScore* of 0.7 ($1 - 0.3$).

The final score of the “taxon name” portion of the input string (*TaxonScore*) is given by the average of the score of each name component. As far as authorities are concerned, another score (*AuthorScore*) ranging from 0 to 1 is calculated using an *ngram* algorithm [17].

The final score, called *ReturnedScore*, ranges from 0 to 100.

$$ReturnedScore = (TaxonScore \times 0.9 + AuthorScore \times 0.1) \times 100 \quad (2)$$

(5) Algorithm Settings

Users can run the algorithm in standard or advanced mode. In standard mode, the algorithm always performs fuzzy matching. In addition, infraspecific epithets are recognized only when preceded by a commonly adopted rank indicator [33]. No issues arise if the rank indicator does not have a dot at the end or if it contains typos (as an example, to indicate the rank “subspecies”, the strings subsp., Subsp., and susp are all accepted).

Further settings can be adjusted by users: (a) selecting a custom infraspecific identifier for subspecies, variety, form, and cultivar, in the case that those used in the input dataset are not supported by the algorithm; (b) enabling or disabling the phonetic match algorithm; (c) selecting the threshold score (from 1 to 100) under which a result is not returned: if 100 is selected, only taxon names that are an exact match are returned; and (d) forcing the parsing algorithm to be case-insensitive (useful if the input names do not follow the ICBN uppercase and lowercase convention).

3. Results

The algorithm integrated in the query systems of FlorItaly aims at improving effectiveness and usability of the system. Each of the three query interfaces (basic, standard, and advanced) allows to combine a string in the taxon name with several parameters checked or selected from closed lists. As an example, in the advanced query interface, users can query for all taxa known to occur in the region Friuli Venezia Giulia as archaeophytes, whose name contains the string “*officinalis*”. The result is a list of two taxa, *Calendula officinalis* L. and *Galega officinalis* L. However, while typing a string, it is possible to input one or more typos, thus biasing the results of the query. As an example, if the input string is “*officialis*” (an “*n*” is missing, Figure 1), the result would be an empty list, and users could be led to the conclusion that the two taxa are not known to occur in the region. The new algorithm allows to overcome this error, finding the nearest match for the input string and then performing the query on it. Thus, the output (Figure 2) highlights that the query string has been replaced by a “corrected” string by the algorithm (“*officinalis*”), thus making users aware that the string they input was not present in the database.

The algorithm operating on lists of taxa aims at providing a novel service in FlorItaly, i.e., the possibility to retrieve a list of accepted names after input of a list of scientific names. Users can register to the system and store up to three lists of names in their workspace. Thus, they can operate on them in different moments for resolving uncertain matches. A list of names (e.g., deriving from the digitization of vegetation relevés taken from a scientific paper) can be input either as a text (each name in a new line) or as CSV (comma-separated values) files. In the latter case, users must specify, through a graphic interface, which column in the input file hosts the taxon names. Furthermore, users can:

1. Select the character used as a separator (comma, colon, semicolon, etc.);
2. Select the terms used for infraspecific taxonomic ranks, such as subspecies, variety, and form;
3. Allow the inclusion/exclusion of phonetic matches;
4. Increase the tolerance of the algorithm, i.e., accepting a wider range of near matches than those allowed by the default settings. Increasing the degrees of freedom of the algorithm increases the time required for performing the query.

After the upload, the system performs a SELECT DISTINCT query on the column hosting the names, thus removing duplications. The algorithm will then compare each name in the list with the thesaurus of FlorItaly and return a page (Figure 3) listing the number of names (a) in the original list, (b) with no match, (c) with unambiguous match, and (d) with ambiguous match. The time required for uploading and matching a list depends on the size of the input. Normally, it ranges from 30 s to 5 min. The upper limit in the input files is 5000 names. Users can hide/show any part of the list and can operate on ambiguous matches only, selecting a match among those proposed by the algorithm (each with the corresponding score). At the end of the process, users can either download a CSV file with four columns (original names, accepted names, synonyms—when the input name matches a synonym and not an accepted name—and the match score) or, if they are logged in, save and continue the process in another moment.

Portale della Flora d'Italia
Portal to the Flora of Italy

HOME CERCA/QUERY INFO CREDITI/CREDITS CONTATTI/CONTACTS

Nome scientifico / Scientific name (*)
officialis

Famiglia / Family:

Con immagini / with images: si / yes no

Default = Italia, o scegliere una o più regioni
Default = Italy, or select one or more regions

Friuli Venezia Giulia
 Veneto
 Trentino-Alto Adige
 Lombardia
 Piemonte
 Valle d'Aosta
 Liguria
 Emilia-Romagna
 Toscana
 Umbria
 Marche
 Abruzzo
 Molise
 Lazio
 Campania
 Calabria
 Basilicata
 Puglia
 Sicilia
 Sardegna

Escludi / exclude - **Includi / include** - reset

record storici / historical records
 estinte / extinct
 record erronei / wrong records
 record dubbi / doubtful records
 data deficient

alloctone invasive / invasive alien
 alloc. naturalizz. / naturalized alien
 alloc. casuali / casual alien
 altre alloctone / other alien
 criptogeniche / cryptogenic

neofite / neophytes
 archeofite / archaeophytes

ferale / feral
 culton

endemiche Italiane / Italian endemic
 endem. esclusive / exclusive endem.

taxa incerti / doubtful taxa

PULISCI I CAMPI / RESET - CERCA / SEARCH

(*) Attenzione: stringhe con meno di 4 caratteri saranno ignorate
Warning: strings shorter than 4 characters will be ignored

Portale della flora d'Italia
Portal to the flora of Italy

Figure 1. Results of a query in the advanced query interface of FlorItaly. The input name is incorrectly spelled (“officialis” instead of “officinalis”).



Nessun risultato trovato per / No results found for: **officialis**

Risultato più simile / Most similar result: **officinalis**

Parametri di ricerca / Query parameters:

Taxon = officialis

Senza immagini / Without images

Distribuzione / Distribution: Friuli Venezia Giulia;

Includi / Include archeofite / archaeophytes

Nomi accettati / Accepted names: 2

- [Calendula officinalis L.](#)
- [Galega officinalis L.](#)



Figure 2. Advanced query result from FlorItaly. Since no exact matches were found for the misspelled input string “*officialis*”, the near match algorithm returned the most similar string in the database (“*officinalis*”) and performed the query by using it. Users are provided with a warning at the beginning of the results page.

To test the effectiveness of this algorithm, three heterogeneous datasets in CSV format were used:

- Dataset A: 890 names, without authorities, and “s.” as the infraspecific identifier for subspecies;
- Dataset B: 2981 names, all written in uppercase;
- Dataset C: 304 names written following the current code of nomenclature (ICBN).
- The results are reported in Table 1.

Table 1. Results of the list matching algorithm on three different datasets (see text). For each dataset, other than the total number of taxa, positive matches (divided into unambiguous and ambiguous, the latter requiring user input) and no matches are reported.

| Dataset | Total Names | Unambiguous Matches | Ambiguous Matches | No Matches |
|---------|-------------|---------------------|-------------------|------------|
| A | 890 | 609 (68.43%) | 271 (30.45%) | 10 (1.12%) |
| B | 2981 | 2649 (88.86%) | 296 (9.93%) | 36 (1.21%) |
| C | 304 | 293 (96.38%) | 10 (3.29%) | 1 (0.33%) |

The source code developed during this study is available on GitHub by querying for “FlorItaly name match”.

Portale della Flora d'Italia
Portal to the Flora of Italy

HOME CERCA/QUERY INFO CREDITI/CREDITS CONTATTI/CONTACTS

Nomi inseriti / Inserted names: 890
 Nessun allineamento / No match: 10 Mostra / Show Nascondi / Hide
 Allineamenti univoci / Unambiguous matches: 777 Mostra / Show Nascondi / Hide
 Allineamenti ambigui / Ambiguous matches: 103 Mostra / Show Nascondi / Hide

RISULTATI / RESULTS

Nome inserito / Input name: Achillea maritima
 Allineamenti suggeriti / Suggested matches: Seleziona / Select

Nome inserito / Input name: Agrimonia eu...
 Allineamenti suggeriti / Suggested matches: Seleziona / Select

Nome inserito / Input name: Aira elegantissima
 Allineamenti suggeriti / Suggested matches: Seleziona / Select

Nome inserito / Input name: Allium roseum
 Allineamenti suggeriti / Suggested matches: Seleziona / Select

Nome inserito / Input name: Allium subhirsutum
 Allineamenti suggeriti / Suggested matches: Seleziona / Select

Modalità suggerita:
 Achillea maritima (L.) Ehrend. & Y.P.Guo subsp. maritima [PUNTEGGIO DI SIMILARITÀ / SIMILARITY SCORE: 81%]
 Achillea nana L. [PUNTEGGIO DI SIMILARITÀ / SIMILARITY SCORE: 33%]

Figure 3. Result page after the input of a CSV file containing a list of names. The interface highlights the number of names in the list and the number of no matches, unambiguous, and ambiguous matches. Each part of the list can be hidden. As far as ambiguous matches are concerned, users can select a match among those proposed by the algorithm (each listed with its matching score). When strings are too long, holding the pointer on the incomplete string for a few seconds allows to view the full string in a text pop-up.

4. Discussion

Scientific names written in Latin are not part of everyday language in any modern country. Thus, their input as strings in a query system can generate several issues [34] and can be associated with typos. In the authorities' portion of the name, furthermore, different abbreviations or even the inclusion or exclusion of a blank space could lead to biases in the results of a query. The query system of FlorItaly was originally developed as an exact matcher, with no automated correction of the input string(s). However, in order to improve users' experience, it was decided to switch from an exact to a near match algorithm. This new algorithm has been integrated in the three query systems of FlorItaly, which now warn users if a misspelled string was input and provide results for the string which is the closest match in the reference database. As an example, for retrieving the data for *Ailanthus altissima* Mill., a user can query FlorItaly for the string "lanthus alt". The only name matching the query is *A. altissima*. However, if the input string was "lantus alt", with an omission error, the old query system did not return any result. The novel algorithm, on the contrary, corrects the input string into "lanthus alt", warns the user, and returns the scientific name in the database which is the nearest match, i.e., *A. altissima*.

Other than querying FlorItaly by taxon name, a new service is provided by the possibility of comparing lists with the reference database. The new service can be useful, especially in the framework of the digitization of previously published botanic data, as well as herbarium specimens—efforts which are being carried out by the botanic community in Italy and abroad. Since old publications, as well as specimen labels, normally do contain outdated nomenclature, a system for resolving all the names in a batch process can be a relevant time saver to researchers. The algorithms integrated in FlorItaly are based upon Taxamatch [17]. However, while Taxamatch matches a scientific name up to the species level, the algorithms in FlorItaly support up to two infraspecific epithets, following the TNRS approach [9,35]. Furthermore, the algorithms were developed to address some matching issues which occur in the TNRS [36], especially when: (a) input names are entirely in upper- or lowercase; (b) the word "f." is present (it could be part of the author's name—

filius—or the taxonomic rank “form”); or (c) epithets of species, or infraspecific ranks, are input with uppercase initials. These issues are related to the fact that the principal parsing algorithms, *gnparser* and *gbif parser*, are case-sensitive and use the presence of upper- and lowercase characters to properly parse the input strings. This approach allows for an accuracy of 99% when input names are written following the nomenclatural code rules. However, when this is not the case, part of the accuracy is lost [13,37]. Another relevant issue occurs when parsing a scientific name in which part of the authority can be the same as an infraspecific rank indicator. In the case of names written following the ICBN uppercase and lowercase rules, this could happen for the string “f.” only, which can either refer to “filius” (part of the authority) or “form” (infraspecific rank). However, when names are not written following the ICBN uppercase and lowercase rules, as in the case of names written entirely in uppercase, “V.”, “F.”, “S.”, and “C.” can lead to incorrect parsing, since they can either be part of the authority or refer to infraspecific ranks (“variety”, “form”, “subspecies”, and “cultivar”). Note that “C.” and “S.” are treated as infraspecific rank indicators by the algorithms in FlorItaly only if they are manually selected by the user as infraspecific rank indicators, since they are not commonly used [33]. In order to properly address these issues, algorithms were developed mainly focusing on extending the already effective performance of Taxamatch [17] for handling infraspecific ranks. This is especially relevant in Italy, since lists of names—which can derive, e.g., from digitization efforts—derive from several older backbones, such as the floras by Fiori [38], Zangheri [39], or Pignatti [40], who made extensive use of infraspecific ranks that often are not written following the ICBN standard.

The inclusion into FlorItaly of the two algorithms aims at providing the scientific community with further services for improving data retrieval through the use of taxon names as identifiers. There is, especially in Italy, a wealth of different resources hosting digitized data on plants. Some of them are already aggregated, and their output is provided together with checklist data in the taxon pages of FlorItaly [18]. However, this was possible only because these resources share a common nomenclatural backbone. Thus, providing a simple, fast, and reliable tool for aligning any list of names to the nomenclatural backbone, provided by the national checklists, is a relevant step towards the aggregation of other resources, such as digitized herbaria, image archives, databases of vegetation relevés, etc. The aggregation of databases of phytochemical and ethnobotanic data is already ongoing, together with datasets of primary biodiversity data. The development of a web service to provide data from FlorItaly to other resources, such as the National Biodiversity Network of the Ministry of Environment, is one of the relevant developments planned for the future.

These developments, together with the incoming novel National Center for Botanic Data of the European Research Infrastructure LifeWatch, which will be based at the University of Bologna, aim at providing the scientific community with a unique entry point for botanic data concerning the vascular flora of Italy.

Author Contributions: Conceptualization, M.C., P.L.N. and S.M.; methodology, M.C. and S.M.; software, M.C. and S.M.; validation, S.M. and P.L.N. data curation, M.C.; writing—original draft preparation, S.M. and M.C.; writing—review and editing, P.L.N., S.M. and M.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data are available online in FlorItaly (<http://dryades.units.it/floritaly>, accessed on 29 March 2021).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Linnaeus, C. *Critica Botanica*; Apud Conradum Wishoff: Lugduni Batavorum, The Netherlands, 1737.
2. Patterson, D.; Cooper, J.; Kirk, P.; Pyle, R.; Remsen, D. Names are key to the big new biology. *Trends Ecol. Evol.* **2010**, *25*, 686–691. [[CrossRef](#)] [[PubMed](#)]
3. Nimis, P.L. A tale from Bioutopia. *Nature* **2001**, *413*, 21. [[CrossRef](#)]
4. Kennedy, J.; Kukla, R.; Paterson, T.; Ludäscher, B.; Raschid, L. Scientific names are ambiguous as identifiers for biological taxa: Their context and definition are required for accurate data integration. In *Data Integration in the Life Sciences, Proceedings of the Second International Workshop, San Diego, CA, USA, 20–22 July 2005*; Lecture Notes in Computer Science 3615; Springer: Berlin/Heidelberg, Germany, 2005; pp. 80–95.
5. Remsen, D. The use and limits of scientific names in biological informatics. *ZooKeys* **2016**, *550*, 207–223. [[CrossRef](#)] [[PubMed](#)]
6. Clark, T.; Martin, S.; Liefeld, T. Globally distributed object identification for biological knowledgebases. *Brief. Bioinform.* **2004**, *5*, 59–70. [[CrossRef](#)]
7. Martin, S.; Hohman, M.M.; Liefeld, T. The impact of Life Science Identifier on informatics data. *Drug Discov. Today* **2005**, *10*, 1566–1572. [[CrossRef](#)]
8. Page, R.D.M. Biodiversity informatics: The challenge of linking data and the role of shared identifiers. *Brief. Bioinform.* **2008**, *9*, 345–354. [[CrossRef](#)]
9. Boyle, B.; Hopkins, N.; Lu, Z.; Garay, J.A.R.; Mozzherin, D.; Rees, T.; Matasci, N.; Narro, M.L.; Piel, W.H.; McKay, S.J.; et al. The taxonomic name resolution service: An online tool for automated standardization of plant names. *BMC Bioinform.* **2013**, *14*, 16. [[CrossRef](#)] [[PubMed](#)]
10. Cachueta-Palacio, M. Towards an index of all known species: The Catalogue of Life, its rationale, design and use. *Integr. Zoöl.* **2006**, *1*, 18–21. [[CrossRef](#)] [[PubMed](#)]
11. Norman, K.E.A.; Chamberlain, S.; Boettiger, C. taxadb: A high-performance local taxonomic database interface. *Methods Ecol. Evol.* **2020**, *11*, 1153–1159. [[CrossRef](#)]
12. Rees, T. TAXAMATCH, a “fuzzy” matching algorithm for taxon names, and potential applications in taxonomic data-bases. In *Proceedings of TDWG; Biodiversity Information Standards (TDWG) and the Missouri Botanical Garden*: St. Louis, MO, USA, 2008; p. 35.
13. Mozzherin, D.Y.; Myltsev, A.A.; Patterson, D.J. “gnparser”: A powerful parser for scientific names based on Parsing Expression Grammar. *BMC Bioinform.* **2017**, *18*, 279. [[CrossRef](#)]
14. GBIF Name Parser. Available online: <https://www.gbif.org/tools/name-parser> (accessed on 20 January 2021).
15. Berghe, E.V.; Coro, G.; Bailly, N.; Fiorellato, F.; Aldemita, C.; Ellenbroek, A.; Pagano, P. Retrieving taxa names from large biodiversity data collections using a flexible matching workflow. *Ecol. Inform.* **2015**, *28*, 29–41. [[CrossRef](#)]
16. Christen, P. A Comparison of Personal Name Matching: Techniques and Practical Issues. In *Proceedings of the Sixth IEEE International Conference on Data Mining—Workshops (ICDMW’06)*, Hong Kong, China, 18–22 December 2006; pp. 290–294.
17. Rees, T. Taxamatch, an Algorithm for Near (‘Fuzzy’) Matching of Scientific Names in Taxonomic Databases. *PLoS ONE* **2014**, *9*, e107510. [[CrossRef](#)]
18. Martellos, S.; Bartolucci, F.; Conti, F.; Galasso, G.; Moro, A.; Pennesi, R.; Peruzzi, L.; Pittao, E.; Nimis, P.L. FlorItaly—The portal to the Flora of Italy. *PhytoKeys* **2020**, *156*, 55–71. [[CrossRef](#)]
19. Bartolucci, F.; Peruzzi, L.; Galasso, G.; Albano, A.; Alessandrini, A.; Ardenghi, N.M.G.; Astuti, G.; Bacchetta, G.; Ballelli, S.; Banfi, E.; et al. An updated checklist of the vascular flora native to Italy. *Plant Biosyst. Int. J. Deal. Asp. Plant Biol.* **2018**, *152*, 179–303. [[CrossRef](#)]
20. Galasso, G.; Conti, F.; Peruzzi, L.; Ardenghi, N.M.G.; Banfi, E.; Celesti-Grapow, L.; Albano, A.; Alessandrini, A.; Bacchetta, G.; Ballelli, S.; et al. An updated checklist of the vascular flora alien to Italy. *Plant Biosyst. Int. J. Deal. Asp. Plant Biol.* **2018**, *152*, 556–592. [[CrossRef](#)]
21. Bartolucci, F.; Domina, G.; Ardenghi, N.M.; Banfi, E.; Bernardo, L.; Bonari, G.; Buccomino, G.; Calvia, G.; Carruggio, F.; Cavallaro, V.; et al. Notulae to the Italian native vascular flora: 5. *Ital. Bot.* **2018**, *5*, 71–81. [[CrossRef](#)]
22. Galasso, G.; Domina, G.; Adorni, M.; Ardenghi, N.M.; Bonari, G.; Buono, S.; Cancellieri, L.; Chianese, G.; Ferretti, G.; Fiaschi, T.; et al. Notulae to the Italian alien vascular flora: 5. *Ital. Bot.* **2018**, *5*, 45–56. [[CrossRef](#)]
23. Bartolucci, F.; Domina, G.; Ardenghi, N.M.; Bacchetta, G.; Bernardo, L.; Buccomino, G.; Buono, S.; Caldararo, F.; Calvia, G.; Carruggio, F.; et al. Notulae to the Italian native vascular flora: 6. *Ital. Bot.* **2018**, *6*, 45–64. [[CrossRef](#)]
24. Galasso, G.; Domina, G.; Alessandrini, A.; Ardenghi, N.M.; Bacchetta, G.; Ballelli, S.; Bartolucci, F.; Brundu, G.; Buono, S.; Busnardo, G.; et al. Notulae to the Italian alien vascular flora: 6. *Ital. Bot.* **2018**, *6*, 65–90. [[CrossRef](#)]
25. Bartolucci, F.; Domina, G.; Alessandrini, A.; Angiolini, C.; Ardenghi, N.M.; Bacchetta, G.; Banfi, E.; Bolpagni, R.; Bonari, G.; Bräuchler, C.; et al. Notulae to the Italian native vascular flora: 7. *Ital. Bot.* **2019**, *7*, 125–148. [[CrossRef](#)]
26. Galasso, G.; Domina, G.; Ardenghi, N.M.; Aristarchi, C.; Bacchetta, G.; Bartolucci, F.; Bonari, G.; Bouvet, D.; Brundu, G.; Buono, S.; et al. Notulae to the Italian alien vascular flora: 7. *Ital. Bot.* **2019**, *7*, 157–182. [[CrossRef](#)]
27. Bartolucci, F.; Domina, G.; Ardenghi, N.M.G.; Bacaro, G.; Bacchetta, G.; Ballarin, F.; Banfi, E.; Barberis, G.; Beccarisi, L.; Bernardo, L.; et al. Notulae to the Italian native vascular flora: 8. *Ital. Bot.* **2019**, *8*, 95–116. [[CrossRef](#)]
28. Galasso, G.; Domina, G.; Andreatta, S.; Angiolini, C.; Ardenghi, N.M.G.; Aristarchi, C.; Arnoul, M.; Azzella, M.M.; Bacchetta, G.; Bartolucci, F.; et al. Notulae to the Italian alien vascular flora: 8. *Ital. Bot.* **2019**, *8*, 63–93. [[CrossRef](#)]

29. Bartolucci, F.; Domina, G.; Andreatta, S.; Angius, R.; Ardenghi, N.M.G.; Bacchetta, G.; Ballelli, S.; Banfi, E.; Barberis, D.; Barberis, G.; et al. Notulae to the Italian native vascular flora: 9. *Ital. Bot.* **2020**, *9*, 71–86. [[CrossRef](#)]
30. Galasso, G.; Domina, G.; Adorni, M.; Angiolini, C.; Apruzzese, M.; Ardenghi, N.M.G.; Assini, S.; Aversa, M.; Bacchetta, G.; Banfi, E.; et al. Notulae to the Italian alien vascular flora: 9. *Ital. Bot.* **2020**, *9*, 47–70. [[CrossRef](#)]
31. Bartolucci, F.; Domina, G.; Bagella, S.; Barberis, G.; Briozzo, I.; Calbi, M.; Caria, M.C.; Cavallaro, V.; Chianese, G.; Cibeï, C.; et al. Notulae to the Italian native vascular flora: 10. *Ital. Bot.* **2020**, *10*, 47–55. [[CrossRef](#)]
32. Galasso, G.; Domina, G.; Azzaro, D.; Bagella, S.; Barone, G.; Bartolucci, F.; Bianco, M.; Bolzani, P.; Bonari, G.; Boscutti, F.; et al. Notulae to the Italian alien vascular flora: 10. *Ital. Bot.* **2020**, *10*, 57–71. [[CrossRef](#)]
33. Sigovini, M.; Keppel, E.; Tagliapietra, D. Open Nomenclature in the biodiversity era. *Methods Ecol. Evol.* **2016**, *7*, 1217–1225. [[CrossRef](#)]
34. Jansen, F.; Dengler, J. Plant names in vegetation databases—A neglected source of bias. *J. Veg. Sci.* **2010**, *21*, 1179–1186. [[CrossRef](#)]
35. iPlant Modified Version of the SilverBiology PHP/MySQL Port of Taxamatch. Available online: <https://github.com/iPlantCollaborativeOpenSource/TNRS> (accessed on 20 January 2021).
36. iPlant Taxonomic Name Resolution Service. Available online: <http://tnrs.iplantcollaborative.org> (accessed on 20 January 2021).
37. Global Name Parser/Gnparser (v0.14.4). Available online: <https://parser.globalnames.org> (accessed on 20 January 2021).
38. Fiori, A. *Nuova Flora Analitica d'Italia*; Tipografia M. Ricci: Firenze, Italy, 1923.
39. Zangheri, P. *Flora Italica*; Cedam: Padova, Italy, 1976.
40. Pignatti, S. *Flora d'Italia*; Edagricole: Milano, Italy, 1982.