EBNF grammar for Botanical Scientific Names

G. Hagedorn 1998-2000

Notes

This document has been written on the basis of Bisby (1994) as a first analysis for later definitions in XML. I view it only as a first attempt and publish it to help other people in their analysis. I can not foresee that I have any time myself to work on this. The EBNF grammar presented here is not thoroughly discussed or tested. Also, the analysis should be extended to other groups (animals, bacteria, viruses, etc.), with clear annotations which rules applies in which group.

Please comment in general, but also on the names used for the boring technical definitions. I appreciate any critique about what is wrong. Where can the parser be made more stringent? Does it falsely exclude some cases? What is the best way in EBNF to define alternative versions (currently denoted by superscript numbers), which can be turned on or off depending on the demands on the stringency? I think the exercise is useful, but it also shows the limitations. Perhaps one can use EBNF better than this...

Also, since EBNF was developed for describing and parsing programming languages, I am not sure if a standard parser, when fed with this type of EBNF can do anything with it at all...

@ = Marker for points in need of discussion

Introduction

Version 10^{-3}, G. Hagedorn 1998-2000

EBNF can provide a context free syntactical definition. Symbolic names are defined by expansion rules, '::=' denotes the definition. Literal strings are enclosed in single quotes, a vertical bar ('|') denotes a choice of several options, parentheses '( )' define hierarchies of the operators (e.g. to keep options together), square brackets ('[ ]') denotes 0 or 1 occurrence, curly braces ('{ }') 1 to many occurrences, and thus '[{}]' 0 to many occurrences. The XLM document type definition uses a different, but equivalent notation.

A special feature of this presentation is that for some EBNF rules alternative variants, implementing restrictive or lenient rules are given. These alternative rules are identified by superscript numbers (see e. g. HybridMarker). Only a single of these rules should be used for a given purpose.

This document can be useful to program a set of validation rules for data entry of scientific names, or a parser to check existing lists of scientific names for plausibility. However, if a scientific name passes this test, it is not necessarily correct. Many rules, e.g. gender conformance of the epithet with the gender usage of the genus (which is not necessarily identical with the Latin or Greek gender of the genus!) can not easily be expressed in EBNF. [@Or is it possible? Perhaps one could make a list of gender changing and gender invariant epithet endings: GenderinvariantEnding = ("cola" | …); GenderchangingEnding = everything else…]

An EBNF definition of scientific organism names is especially useful to deal with the tiresome job of checking and comparing imported lists of scientific names. It points to cases where letters and digits are misplace (e.g. , use of zero for capital O, one for lower case l, errors frequently introduced when scanning printed material and using OCR), whether Author abbreviations are followed by a blank or not ('A.B.Jacks.' versus 'A. B. Jacks.'). This document is not intended to supercede a publication like Bisby (1994), but it could supplement a revised publication.

1. General technical definitions:

Digit ::= '0'..'9'

Year ::= Digit Digit Digit Digit

LowerCaseNonDiacriticalLetter ::= 'a'..'z'
(*Note:* excluding punctuation or blank)

UpperCaseNonDiacriticalLetter ::= 'A'..'Z'

LowerCaseDiacriticalLetter ::= 'á' | 'à' | 'â' | 'ä' | 'ã' | 'å' | 'æ' | 'ç' … (etc.!)
(@*Note:* currently informal, provide complete version for electronic versions of this document)

UpperCaseDiacriticalLetter ::= 'Á' | 'À' | 'Â' | 'Ä' | 'Ã' | 'Å' | 'Æ' | 'Ç' … (etc.!)

LowerCaseLetter ::= LowerCaseNonDiacriticalLetter | LowerCaseDiacriticalLetter

UpperCaseLetter ::= UpperCaseNonDiacriticalLetter | UpperCaseDiacriticalLetter

LowerCaseWord ::= LowerCaseLetter {LowerCaseLetter}

UpperCaseWord ::= UpperCaseLetter {LowerCaseLetter}

AnyCaseWord ::= UpperCaseWord | LowerCaseWord
(*Note:* These definitions require a word to have at least 2 letters!
@Question: Are single letter Genera or epithets allowed?)

AbbreviationChar ::= '.'

QuoteSingle ::= '''

QuoteDouble ::= '"'

Apostrophe ::= '''
(*Note:* may or may not be identical with QuoteSingle)

Blank ::= ' '
(*Note:* for parsers allowing formatted/rich text, the code for 'non-breaking blank' should be added here)

2. Special elements:

ScientificNameLetter[1] ::= LowerCaseNonDiacriticalLetter | 'ë' | 'ï' | '-'

ScientificNameLetter[2] ::= LowerCaseNonDiacriticalLetter
(*Note:* This definition is introduced to allow the only diacritical letters frequently accepted in scientific names. The hyphen is frequently used in scientific names, but it can be omitted to form a correct scientific name. The second version of the rule disallows these symbols. @*Question:* Are there any further letters? Æ/æ, Œ/œ?

LowerCaseScientificWord ::= LowerCaseNonDiacriticalLetter {ScientificNameLetter}

UpperCaseScientificWord ::= UpperCaseNonDiacriticalLetter {ScientificNameLetter}

AuthorConcatenationChar ::= ', '

AuthorFinalConcatenationChar ::= Blank ('&' | 'and') Blank

HybridMarker[1] ::= ('x' | '×') Blank
(*Note:* the lowercase letter X or the multiplication sign of the character set used)

HybridMarker[2] ::= ('x' | '×' | '+') Blank
(*Note:* variant, including the plus symbol used to identify interspecific chimaeras)

AggregateMarker ::= 'agg.'
(@*Note:* Are there other accepted forms of this?)

InfraspecificFspString[1] ::= 'fsp.' | 'f. sp.' | 'fm. sp.'

InfraspecificFspString[2] ::= 'f. sp.'
(Note: the formae specialis (= 'fsp.', 'f.sp.', 'fm.sp.') are not formally covered by the botanical code, but are often used with author citations as if they were a formally recognized subspecific rank. Formae specialis apply only to fungi [@*Question:* correct?])

InfraspecificPVString[1] ::= 'pvar' | 'pv' | 'pathovar'
(Note: pathovars are for bacteria what formae specialis are in fungi. In bacteria also serovar and biovar is used, but I have never seen this included in a name. But then I am a mycologist...)

InfraspecificPVString[2] ::= 'pvar'
(Note: version restricted to a standard abbreviation)

InfraspecificRankString[1] ::= 'ssp.' | 'subsp.' | 'var.' | 'f.' | 'fm.' | 'fa.' | 'forma' | 'subf.' | InfraspecificFspString | InfraspecificPVString

InfraspecificRankString[2] ::= 'subsp.' | 'var.' | 'f.' | 'subf.'
(*Note:* restrictive version, use of 'ssp.', 'fm.', 'fa.', and 'forma' is not recommended; InfraspecificFspString occurs only in a subset of names, may be excluded)

CultivarMarker ::= 'cv.'

3. Authors: [@Perhaps rather call it citation?]

Compare Bisby 1994 and Brummitt & Powell 1992

AuthorFirstNameAbbrev[1] ::= UpperCaseLetter {LowerCaseLetter} AbbreviationChar [Blank]

AuthorFirstNameAbbrev[2] ::= UpperCaseLetter {LowerCaseLetter} AbbreviationChar Blank

AuthorFirstNameAbbrev[3] ::= UpperCaseLetter {LowerCaseLetter} AbbreviationChar
(*Note 1:* Only one of the three rule variants should be implemented, depending upon whether the blank after the abbreviation character is optional, required, or not desired (e.g. if 'A.B.Jacks.' is the preferred form). In correct typesetting, a blank should be present.
Note 2: Even in modern recommendations, up to three letter firstname abbreviations occur. Perhaps the concept of a AuthorFirstNameAbbrev is impractical, since many abbreviated first names are indistinguishable from abbreviated parts of multi-part last names, e.g. Cec. Roux', where Cec. stands for Cecilia.
Note 3: The first name may remain unabbreviated in lists of standardized author names. For the purpose of this document it is then treated like a AuthorLastName.

AuthorLastName ::= [(LowerCaseLetter | UpperCaseLetter) Apostrophe [Blank]]
[[(AnyCaseWord [Apostrophe LowerCaseWord] [AbbreviationChar] Blank)]
(AnyCaseWord [Apostrophe LowerCaseWord] [AbbreviationChar] Blank)]
UpperCaseWord [AbbreviationChar]
Note 1: A maximum of 3 words is allowed for lastname (using {} would allow an infinite number!)
Note 2: The apostrophe can be part of a prefix (d', D', O'), in which case it is always in the first position and after a single upper or lower case letter. Unfortunately, some names (or transcriptions of names) have an apostrophe as part of the name ('Ben Ze'ev', 'Ola'h') so that a second apostrophe rule is necessary.

AuthorSuffix ::= [','] Blank ('f.' | 'fil.' | 'jr.' | 'jun.' | 'sen.' | 'I' | 'II' | 'III' | 'IV')
Note: The leading comma can be omitted or enforced, to avoid variants like 'X, jun.' versus 'X jun.'
(*@Question:* any other lower case suffixes than 'f.' or 'fil.' for filius, 'jun.'/'sen.'? *Note:* 'f.' may occur together with 'L.'! More restrictive rule may either omit the ', ' in front of the suffix or enforce it to achieve a higher uniformity.)

Author ::= 'L.' | ([{AuthorFirstNameAbbrev}] AuthorLastName) [AuthorSuffix]
(*@Question:* Are other single letter abbreviations than L. possible? Insects: F. = Fabricius. This rule requires at least 2 characters unless the author is 'L.' An additional rule that authors must come from a list of known authors can be applied.
Note: [{AuthorFirstNameAbbrev}] could perhaps be restricted to 0-3 occurrences!
Test authors: 'L. fil.'. 'A. St.-Hil.', 'C.-J. Duval', 'Émile-Weil', 'Da Silva', 'O'Connel'; 'd'Artagne', 'Fischer von Waldheim', 'von Waldheim', 'A. Fischer von Waldheim'.)

AuthorTeam ::= Author [[{AuthorConcatenationChar Author }] AuthorFinalConcatenationChar Author]
(*Note:* This definition allows 'Author', 'Author & Author', and 'Author, Author, Author & Author', but not 'Author, Author'. Also, this rule is restrictive and does not allow the alternative AuthorFinalConcatenationChar forms 'and' or 'et'. Team abbreviations like 'H. & P. Sydow' are explicitly not accepted, they should be converted to 'H. Sydow & P. Sydow'.)

BibliographicReference ::= @@@ NOT YET DONE!

BibliographicReferenceDetail ::= ':' Blank {Digit}
(*Note:* usually the page number, but may also reference a figure or table.)

ProtologueCitationBot ::= BibliographicReference [BibliographicReferenceDetail]
(*Note:* A protologue citation is rarely used for botanical name, usually only in a complete taxonomic revision.)

ProtologueCitationZoo ::= '(' Year ')' [BibliographicReference [BibliographicReferenceDetail]]
(*Note:* The use of the year is common in zoology. Question: Is it required? In this version the year is obligatory, the reference facultative)

SensuCitation ::= 'sensu' ('str.' | 'stricto' | 'l.' | 'lat.' | 'lato' | AuthorTeam)
(*Note:* usually not used, but the addition of a bibliographic citation would be useful to support the potential taxon concept.)

AuthorString[1] ::= ['(' AuthorTeam ') '] AuthorTeam [(' in ' | ' ex ' | ' : ') AuthorTeam] [ProtologueCitationBot | SensuCitation]
Version 1 for botanical code of nomenclature. In botany the author team of the basionym is placed in parenthesis in front of the author team introducing a new combination
Note: '[ProtologueCitation]' can be omitted if only names with a protologue = bibliographic reference shall be allowed.
Note 2: 'non' is sometimes used to clarify a species concept which deviates from a homonym (species with same name, but different authors) encountered in the current literature. While this is valuable information in a flora, it is superfluous in a computer list and is therefore not included in the rule.
Note 3: 'in' and 'ex' may be omitted. 'in' is sometimes used to indicate that the protologue appeared in a publication by a different author team, 'in' and the following authors or reference can be omitted. 'ex' is used, if the name was initially invalidly published. The authors following the 'ex' are the correct authors of the taxon, but they intended to recognize the work of the authors which introduced the invalid publication. Thus, the 'ex' and the preceeding (sic!) authors can be omitted.

AuthorString[2] ::= ('(' AuthorTeam ') ') | AuthorTeam [(' in ' | ' ex ') AuthorTeam] [ProtologueCitationZoo | SensuCitation] **@@@@@**
*Note:*Version 2 for zoological code of nomenclature. In zoology, the author introducing a new combination is never mentioned.

4a. Scientific Names:

ScientificName ::= Monomial | Binomial | Trinomial

Monomial ::= UpperCaseScientificWord

GenusWithoutAuthor ::= [HybridMarker] Monomial
(*Note:* If possible, a parser should implement the additional rule that the genus must come from a list of known genera)

Genus ::= GenusWithoutAuthor [Blank AuthorString]

Epithet ::= LowerCaseScientificWord

CultivarEpithet ::= (CultivarMarker Epithet) | (QuoteDouble Epithet QuoteDouble) | (QuoteSingle Epithet QuoteSingle)
Note: The parser could optionally be more restrictive, e.g. allowing only double quoted cultivar names

BinomialWithoutAuthor ::= GenusWithoutAuthor Blank (([HybridMarker] Epithet) | (Epithet Blank HybridMarker Epithet) | CultivarEpithet)
(*Note:* Two different types of hybrid formulas exist: 'GenusWithoutAuthor Blank HybridMarker Epithet', example: '*Spartina x townsendii*' and 'GenusWithoutAuthor Blank Epithet Blank HybridMarker Epithet', example: '*Primula veris x vulgaris*')
@Problem: The requirement for existence of authors perhaps shall be set separately for cultivars, hybrids, and normal names!

Binomial ::= BinomialWithoutAuthor Blank AuthorString

Trinomial ::= BinomialWithoutAuthor Blank (
(InfraspecificRankString Blank Epithet Blank AuthorString)

| (AuthorString Blank InfraspecificRankString Blank Epithet)
| ([AuthorString Blank] CultivarEpithet))
Note: 'Blank AuthorString InfraspecificRankString Epithet' occurs in autonyms. *@Question:* How can EBNF define the additional condition be that the species epithet and the subspecific epithet must be identical?

4b. Scientific Names (Alternative version, following Bisby 1994)

Note: This is easier to understand, but probably a wrapped up version like 4a. is more appropriate for a parser, but that depends on the parser software.

Taxon ::= Monomial | Species | SpeciesAggregate | IntergenericHybrid | InterspecificHybrid | SubspecificTrinomial | Cultivar | CultivarGroup

GenusWithoutAuthor ::= Monomial
(*Note:* If possible, a parser should implement the additional rule that the genus must come from a list of known genera)

Species ::= GenusWithoutAuthor Blank Epithet Blank AuthorString

SpeciesAggregate ::= GenusWithoutAuthor Blank Epithet Blank AggregateMarker
Note: "species aggregates do not have an author string" (Bisby 1994) ® but is sensu used? Perhaps add SensuCitation.

IntergenericHybrid ::= HybridMarker Species

InterspecificHybrid ::= GenusWithoutAuthor Blank (HybridMarker Epithet) | (Epithet Blank HybridMarker Epithet) Blank AuthorString

SubspecificTrinomial ::= BinomialWithoutAuthor Blank
((InfraspecificRankString Blank Epithet Blank AuthorString)
| (AuthorString Blank InfraspecificRankString Blank Epithet))
Note 1: A form with Species instead of BinomialWithoutAuthor, i.e. with an additional author string, is frequently used, but it is correct and preferable to omit the species authors.
Note 2: 'Blank AuthorString InfraspecificRankString Epithet' occurs in autonyms. Autonyms are names which arise without explicit publication if a variety is published. If author 'Author2' publishes 'G. alba var. nigra Author2' in the species 'G. alba Author1', the autonym 'G. alba Author1 var. alba' is automatically introduced
@Question: How can EBNF define the additional condition be that the species epithet and the subspecific epithet must be identical?
Note 3: Quadrinomials (e.g. 'Genus species subsp. x var. y') can and should be shortened to the trinomial with the lowest rank (e.g. 'Genus species subsp. x var. y').

Cultivar ::= BinomialWithoutAuthor [Blank AuthorString] Blank CultivarEpithet
Note: In contrast to Bisby 1994, this rule allows to omit the botanical author from a cultivar name. "There is no author string for a cultivar name" (Bisby 1994).

CultivarGroup ::= BinomialWithoutAuthor [Blank AuthorString] '(' {(AnyCaseWord | Digit) [AbbreviationChar] [Blank]} ')' Blank CultivarEpithet
Note: Cultivar groups names may be upper or lower case, and contain digits or punctuation.

References

*Http://clever.net/phrantic/j_alan/comp2.html#EBNF

Bisby, F. A. (1994) Plant names in botanical databases. Hunt Inst. for Bot. Documentation, Carnegie Mellon Univ., Pittsburgh.

Brummitt, R.K. & C.E. Powell 1992. Authors of plant names. Royal Botanic Gardens Kew. (The current electronic version is a TDWG Standard).


SOME RIGHTS RESERVED
XHTML 1.0
CSS