Introduction to Refsort/Ruby and Its Application to Taxonomy

October 25, 2013

Toshio Otaguro, Tsukuba, Japan

Who am I?

Nice to meet you!

- Research engineer in fluid mechanics.
- Specialized in turbulence, sensitive to terms such as “order” or “random”.
- Driven into taxonomy since learning “Theorem of the Ugly Duckling”.
- Having fun with birding since student.
- Blog “Lakeside Diary”
- http://griffin.cocolog-nifty.com/lakesidediary/

Introduction to Refsort/Ruby for Novice Users

What is Refsort/Ruby?

- A sorting filter referring to a dictionary.
- Implemented in C on MS-DOS in 1991.
- Motivated to automate organizing birding logs.
- Re-implemented in Ruby in 2001 as v1.0, and currently v2.46 (Jul-2013).
- Works with Ruby-1.9.1 and above, i.e., on multi-platform and in multi-encoding.
- Works as a filter, can be combined with redirection and pipe.

What can Refsort/Ruby do?

- Sorts lines of text in the order which is predefined in a separate reference file.
- Sorting order is based on the reference file, and independent of character code or numerical value.
- With the reference file listing 31,623 species of IOC List v3.5 (2013), you can sort randomly listed so many species in about 2 seconds on your PC.

Features

- Pros
  - You can easily make a reference file which defines the order of sorting.
  - You can select key fields of sorting flexibly.
  - You can define synonyms in a field.
  - You can compile output fields flexibly.
  - You can output hierarchical milestones by utilizing embedded comments in a reference file.
- Cons
  - You have to install and use the Ruby interpreter.
  - Since Refsort handles only fields listed in a reference file, you have to include all possible objects comprehensively in exact spellings.

How to get Refsort/Ruby

- The Ruby interpreter is available at https://www.ruby-lang.org/, or http://rubyinstaller.org/
- Refsort and compiled reference files are available at http://griffin.cocolog-nifty.com/lakesidediary/

How a reference file looks like

- IOC List v3.5 compiled as a reference file

```
#!E -*- coding: UTF-8 -*-
#!R ","
#
#!m >[TINAMIFORMES]  #2 PHY, "Tinamous (with moas) are the outgroup to the …
#!m >>[Tinamidae]     # [Tinamous]
#!m >>>>[Tinamus]       # "Hermann, 1783"
Tinamus tao, Grey Tinamou   # "Temminck, 1815", SA, Amazonia,
Tinamus tao larensis,   # "Phelps & Phelps Jr, 1949", , "c Colombia, nw Venezuela",
Tinamus tao septentrionalis,    # "Brabourne & Chubb, C, 1913", , "ne Venezuela, …
Tinamus tao tao,    # "Temminck, 1815", , nc Brazil,
Tinamus tao kleei,  # "(Tschudi, 1843)", , sc Colombia to e Bolivia and w Brazil,
Tinamus solitarius, Solitary Tinamou    # "(Vieillot, 1819)", SA, se,
Tinamus solitarius pernambucensis,  # "Berla, 1946", , ec Brazil,
Tinamus solitarius solitarius,  # "(Vieillot, 1819)", , "e Brazil, Paraguay, ne Argentina",
…
```

Text file of 34,455 lines, listing 31,623 species and subspecies

How it works (1)

```
# ---- tamagawa.txt -----
Naumann's Thrush          # on fallen leaves
Eurasian Tree Sparrow     # garden in a private house
Daurian Redstart          # male, in the bush on the river bank
Mallard                   # about 50
Eastern Spot-billed Duck  # more than 100
Eurasian Wigeon           # about 50
Northern Pintail          # about 100
Northern Shoveler         # male & female
Japanese Wagtail
Black Kite                # above the river
Black-faced Bunting       # in the bush on the river bank
Brambling                 # didn't see well
Black-headed Gull         # about 20
Little Grebe              # upstream of the dam
Bull-headed Shrike        # singing male
```

This file should be encoded in ASCII.

Text after "#" is treated as comment.

How it works (2)

[Console Input]

>ruby refsort.rb -f ioclist_v35a.ref -nq -r1 -l ”,” tamagawa.txt↲

Invoke the Refsort　reference file　Refsort options　input file

[Console Output] (Standard Error Output)

!R 39: -R ”,” redefined
Ioclist_v35a.ref: total of 10657 records　reference file records
tamagawa.txt: total of 15 records　input file records
tamagawa.txt: total of 15 identified records　identified records
tamagawa.txt: total of 15 unique records　unique records

How it works (3)

- Sorts whole input lines including comments.
- Adds line number as directed by the option.

[Console Output] (Standard Output)

```
 1 Eurasian Wigeon          # about 50
 2 Mallard                  # about 50
 3 Eastern Spot-billed Duck # more than 100
 4 Northern Shoveler        # male & female
 5 Northern Pintail         # about 100
 6 Little Grebe             # upstream of the dam
 7 Black Kite               # above the river
 8 Black-headed Gull        # about 20
 9 Bull-headed Shrike       # singing male
10 Naumann's Thrush         # on fallen leaves
11 Daurian Redstart         # male, in the bush on the river bank
12 Eurasian Tree Sparrow    # garden in a private house
13 Japanese Wagtail
14 Brambling                # didn’t see well
15 Black-faced Bunting      # in the bush on the river bank
```

Record and Field for Sorting Enthusiasts

Record and field

- Record is a basic unit of information processing.
  - In text processing, line is naturally the record.
  - Line is a string delimited by a newline code.
  - Newline code is dependent on the platform…CR/LF, LF, CR
  - Within a record, field is defined as a sub-structure
- Field is a string delimited by specific symbols.
  - Usually blank characters (space/tab) or a comma are used as delimiters.
  - Different delimiters can be defined at a same time.

How record and field look like

[ioclist_v35a.ref]

Tinamus tao, Grey Tinamou↲ 0th record

Tinamus solitarius, Solitary Tinamou↲ 1st record

Tinamus osgoodi, Black Tinamou↲ 2nd record

0th field 1st field

- Record delimiter is newline.
  - Newline is dependent on the platform.
- Field delimiter is comma.
  - Blank spaces adjacent to the delimiter are ignored.

Field delimiter in Refsort

- Default field delimiters are space and tab.
- Delimiters can be defined in the command line or can be redefined in reference/input files.
- Delimiters can be modified in the middle of reference and input files.
- Comma is recommended as the field delimiter of reference and input files.
- A pair of double quotation marks “” can flexibly delimit fields.

How to specify target fields

- Key fields can be specified with -r and -i options.

[Example]

>ruby refsort.rb -f ioclist_v35a.ref -q -r1 -l ”,” tamagawa.txt

- Since English names are placed in the 1st field of the reference file, '-r1' is the proper field specification.
- Input key fields can be specified with -i option.
- Without field specification, it is assumed that options '-r0' and '-i0' are specified.
- Ordered group of fields such as '-r 2,3,0-1' is accepted. If a field doubly specified, only the first one is accepted.
- Note that specifying different fields does not invoke multi-key sorting, but does sorting with a combination of those fields.

Synonyms and aliases

- Different synonyms can be placed in one field.
- Since each synonym is treated equally, they have the same sorting priority.
- The delimiter of synonyms is “|”. Adjacent spaces are ignored.

[Example]

Gavia stellata, Red-throated Diver | Red-throated Loon

- Accepts diversity of naming between U.K. and U.S.A.
- Synonyms or variations in scientific and common names are gracefully accepted, in particular in the course of taxonomical transition.

Compiled Reference Files

JPBirdList_v70p2.ref

- Compiled with “The Check-List of Japanese Birds, 7th ed. (2012)”
- Total of 1,145 species and subspecies
- Field {sname, ename, jname # comment}
  - sname: scientific name
  - ename: common English name
  - jname: Japanese name in Katakana characters
- Encoded in Windows-31J and UTF-8
  - with file name suffixes of w and u, respectively.

wbird_sa091a.ref

- World Bird List according to Sibley-Ahlquist taxonomy with Japanese names added.
- Reference for Japanese names
  - "World Bird List" by Study Group of World Bird Names (1999-2002) http://www.eonet.ne.jp/~saezuri/
  - Some names were suggested by myself.
- Total of 9,996 species
- Field {sname, ename, jname # comment}
- Encoded in Windows-31J and UTF-8
  - with file name suffixes of w and u, respectively.
- Currently not maintained

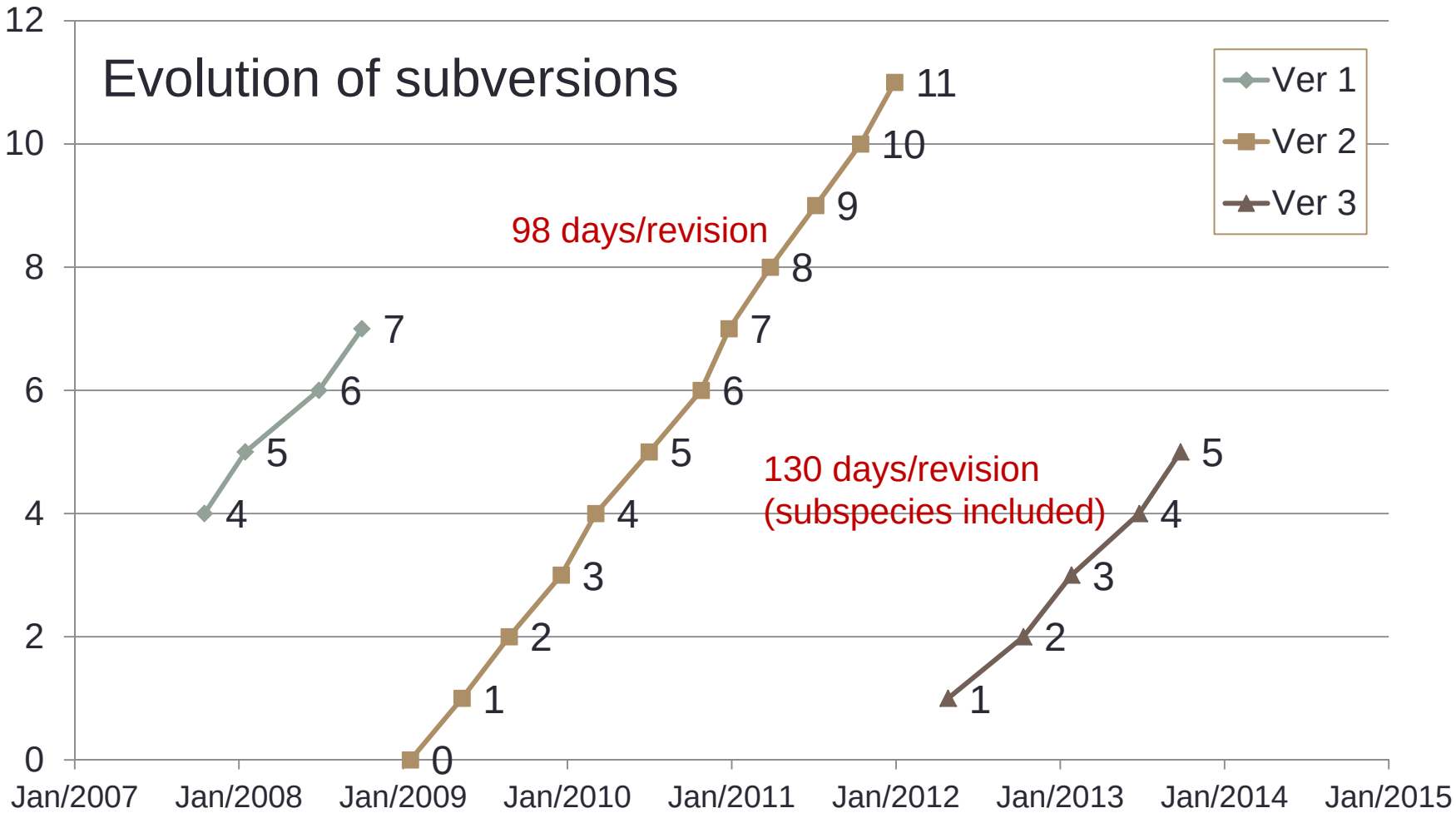
ioclist_v35.ref

- Compiled directly from The Master List of “IOC World Bird List v3.5”.
- Total of 31,623 species and subspecies
- Field {sname, ename # comment}
- Supplemental information in the Master List is quoted in the comment field.
- Encoded in US-ASCII and UTF-8
  - with file name suffixes of a and u, respectively.
  - The version in UTF-8 is genuine since the Master List contains accents and umlauts.

ioclist_v35j.ref

- IOC World Bird List v3.5 with Japanese names.
- Most Japanese names were cited from the previously mentioned Sibley-Ahlquist List.
- Total of 31,671 species and subspecies
  - Many blanks in English and Japanese names for subspecies.
- Field {sname, ename, jname # comment}
- Supplemental information in the Master List is quoted in the comment field.
- Encoded in Windows-31J and UTF-8.
  - with file name suffixes of w and u, respectively.
  - The version in UTF-8 is genuine since the Master List contains accents and umlauts.

History of IOC list



Features of IOC list

- Based upon the philosophy of “Unification” and “Uniqueness”.
- Seeks for uniqueness, does not tolerate diversity.
- American names are prior to British. (poor Diver)
- Differences in spelling between U.S.A. and U.K. are unified with a rule enacted.
- On the other hand, European alphabets (Umlauts and accents) are used in English names, e.g., “Alström’s Warbler”.
- Discussion to be raised for coexisting with diversity.
- Refsort has been developed with a philosophy of tolerating diversity.

Practical Examples for Birders

Simple example

```
# ----- test.txt -----
Mew Gull                    # misidentified?
Naumann's Thrush            # clear voice
Grey-backed Gull            # correct?
Japanese Wagtail
Brown-eared Bulbul
Mallard
Eurasian Skylark            # hovering high
Common Snipe
Ring-necked Wagtail         # exits?
Large-billed Crow           # many
Meadow Bunting
```

Sorting result

```
>ruby refsort.rb -f ioclist_v35a.ref -nq -r1 -l “,” test.txt↲
!R 39: -R "," redefined
ioclist_v35a.ref: total of 10657 records
!I 4: "Grey-backed Gull" not found in ioclist_v35a.ref
!I 10: "Ring-necked Wagtail" not found in ioclist_v35a.ref
test.txt: total of 11 records
test.txt: total of 9 identified records
test.txt: total of 9 unique records
    1 Mallard
    2 Common Snipe
    3 Mew Gull                  # misidentified?
    4 Large-billed Crow         # many
    5 Eurasian Skylark          # hovering high
    6 Brown-eared Bulbul
    7 Naumann's Thrush          # clear voice
    8 Japanese Wagtail
    9 Meadow Bunting
```

unidentified input record

ditto

Output with embedded milestones

```
>ruby refsort.rb -f ioclist_v35a.ref -mnq -r1 -l “,” test.txt↲
```

add -m option to output the embedded milestones

```
[ANSERIFORMES]
[Anatidae]          # [Ducks, Geese and Swans]
[Anas]              # "Linnaeus, 1758"
    1 Mallard
[CHARADRIIFORMES]
[Scolopacidae]      # [Sandpipers, Snipes]
[Gallinago]         # "Brisson, 1760"
    2 Common Snipe
[Laridae]           # [Gulls, Terns and Skimmers]
[Larus]             # "Linnaeus, 1758"
    3 Mew Gull                  # misidentified?
[PASSERIFORMES]
[Corvidae]          # [Crows, Jays]
[Corvus]            # "Linnaeus, 1758"
    4 Large-billed Crow         # many
[Alaudidae]         # [Larks]
[Alauda]            # "Linnaeus, 1758"
    5 Eurasian Skylark          # hovering high
……
```

How to deal with naming fluctuation (1)

- English names consist of indefinite number of words separated by a space or by a hyphen, and the word after a hyphen may be either capitalized or not.

```
# ----- variation.txt -----
Pacific Reef Heron
Malayan Night Heron
Mountain Hawk-Eagle
Swinhoe's Storm-petrel
Collared Scops-owl
```

```
>ruby refsort.rb -f ioclist_v35a.ref -nq -r2 -l “,” variation.txt
!R 39: -R "," redefined
ioclist_v35a.ref: total of 10657 records
!I 5: "Swinhoe's Storm-petrel" not found in ioclist_v35a.ref
!I 6: "Collared Scops-owl" not found in ioclist_v35a.ref
variation.txt: total of 5 records
variation.txt: total of 3 identified records
variation.txt: total of 3 unique records
    1 Malayan Night Heron
    2 Pacific Reef Heron
    3 Mountain Hawk-Eagle
```

How to deal with naming fluctuation (2)

- -b option identifies hyphens with spaces, and does -c option lowercase letters with uppercase.

```
>ruby refsort.rb -f ioclist_v35a.ref -nqbc -r2 -l "," variation.txt↲

!R 39: -R "," redefined
ioclist_v35a.ref: total of 10657 records
variation.txt: total of 5 records
variation.txt: total of 5 identified records
variation.txt: total of 5 unique records
    1 Swinhoe's Storm-petrel
    2 Malayan Night Heron
    3 Pacific Reef Heron
    4 Mountain Hawk-Eagle
    5 Collared Scops-owl
```

How to direct output fields

- Output the entire input line followed by the 0th field of the reference file, delimited by a comma followed by a tab.

```
>ruby refsort.rb -f ioclist_v35a.ref -qr1 -l “,” -O “,¥t” -o ia,r0 test.txt↲

Mallard,                →Anas platyrhynchos
Common Snipe,           →Gallinago gallinago
Mew Gull,               →Larus canus
Large-billed Crow,      →Corvus macrorhynchos
Eurasian Skylark,       →Alauda arvensis
Brown-eared Bulbul,     →Hypsipetes amaurotis
Naumann‘s Thrush,       →Turdus naumanni
Japanese Wagtail,       →Motacilla grandis
Meadow Bunting,         →Emberiza cioides
```

output directives

output field delimiter

Quo vadis, Taxonomy?

Does species exist?

- Classification is deeply related to a human action "recognition."
- Philosophical argument since Classical Greece between "Realismus" and "Nominalisme"
  - Does class exist? Or, class is simply the name of a container for convenience?
- "Theorem of the Ugly Duckling"
  - *If we decide to measure similarity of two objects by the number of all possible predicates that are shared by the two objects, then we can say that "Any arbitrary two objects are equally similar." (S. Watanabe, 1969)*
  - *An ugly duckling is just as similar to a swan as two swans are to each other.*
  - It is likely we put some bias or weights on some particular predicates in order to define classes.

Taxonomy for me

- On balance, I am for nominalism.
  - Species does not exist, nor does genus, etc.
  - Boundary of species is ambiguous and spread, and it always fluctuates.
- Species is a class which man created for convenience of recognition.
  - Classification is an adaptive behavior which man acquired in its history of evolution.
  - Species as well as phylogeny are convenience, but not existence.
  - Does such convenient taxonomy coexist with evolutionary phylogenetics?

Requewt for Comments and Commitment

Comments and commitment requested

- Send your comments and requests for Refsort/Ruby.
- Help edit and maintain reference files.
- Apply Refsort/Ruby to other fields than taxonomy.
- Develop user-friendly software systems embedding Refsort/Ruby as a key component.