

Algorithmic Approaches for Biological Data, Lecture #8

Katherine St. John

City University of New York
American Museum of Natural History

17 February 2016



- More on Pattern Finding: Regular Expressions

Outline



- More on Pattern Finding: Regular Expressions
- Processing CSV files

Recap: Regular Expressions

Recap: Regular Expressions

| Regular Expression | Description of Matching Strings |
|--------------------|--|
| [ACGT]* | A DNA sequence— any string consisting only of A, C, G, and T |

Recap: Regular Expressions

| Regular Expression | Description of Matching Strings |
|----------------------|--|
| <code>[ACGT]*</code> | A DNA sequence— any string consisting only of A, C, G, and T |
| <code>[ACGU]*</code> | A RNA sequence— any string consisting only of A, C, G, and U |

Recap: Regular Expressions

| Regular Expression | Description of Matching Strings |
|----------------------|--|
| <code>[ACGT]*</code> | A DNA sequence– any string consisting only of A, C, G, and T |
| <code>[ACGU]*</code> | A RNA sequence– any string consisting only of A, C, G, and U |
| <code>[AT]+</code> | <i>1 or more repeats of AT: AT, ATAT, ATATAT,...</i> |

Recap: Regular Expressions

| Regular Expression | Description of Matching Strings |
|--|--|
| <code>[ACGT]*</code> | A DNA sequence— any string consisting only of A, C, G, and T |
| <code>[ACGU]*</code> | A RNA sequence— any string consisting only of A, C, G, and U |
| <code>[AT]+</code> | <i>1 or more repeats of AT: AT, ATAT, ATATAT,...</i> |
| <code>ATG[ATGC]{30,1000}A{5,10}</code> | <i>A sequence beginning with ATG and ending with 5 to 10 A's. Overall length is 38 to 1013 bp.</i> |

Recap: Regular Expressions

| Regular Expression | Description of Matching Strings |
|---------------------------------|--|
| $[ACGT]^*$ | A DNA sequence— any string consisting only of A, C, G, and T |
| $[ACGU]^*$ | A RNA sequence— any string consisting only of A, C, G, and U |
| $[AT]^+$ | <i>1 or more repeats of AT: AT, ATAT, ATATAT,...</i> |
| $ATG[ATGC]\{30,1000\}A\{5,10\}$ | <i>A sequence beginning with ATG and ending with 5 to 10 A's. Overall length is 38 to 1013 bp.</i> |
| $([ACGT]\{3\})^*$ | A DNA sequence that exactly breaks into codons (3-letter sequences). |

Recap: Regular Expressions

| Regular Expression | Description of Matching Strings |
|--|--|
| $[ACGT]^*$ | A DNA sequence— any string consisting only of A, C, G, and T |
| $[ACGU]^*$ | A RNA sequence— any string consisting only of A, C, G, and U |
| $[AT]^+$ | <i>1 or more repeats of AT: AT, ATAT, ATATAT,...</i> |
| $ATG[ATGC]\{30,1000\}A\{5,10\}$ | <i>A sequence beginning with ATG and ending with 5 to 10 A's. Overall length is 38 to 1013 bp.</i> |
| $([ACGT]\{3\})^*$ | A DNA sequence that exactly breaks into codons (3-letter sequences). |
| $ATG([ATGC]\{3\})\{30,1000\}A\{5,10\}$ | An open reading frame (ORF): a sequence that starts with the start codon ATG, followed by any number of codons, and ending with a stop codon (TAA, TAG, or TGA). |

- Can be used as a conditional test:



- Can be used as a conditional test:

```
if re.search(r"GC[ATGC]GC", dna):  
    print "Pattern found!"
```





- Can be used as a conditional test:

```
if re.search(r"GC[ATGC]GC", dna):  
    print "Pattern found!"
```

- Search returns a match object:



- Can be used as a conditional test:

```
if re.search(r"GC[ATGC]GC", dna):  
    print "Pattern found!"
```

- Search returns a match object:

```
dna = "ACTCGTACGAAAGCTGCTTATACGCGCG"  
m = re.search(r"GC[ATGC]GC", dna)  
print "The matching string is",  
m.group()  
print "Match starts at", m.start()  
print "Match ends at", m.end()
```



- Can be used as a conditional test:

```
if re.search(r"GC[ATGC]GC", dna):  
    print "Pattern found!"
```

- Search returns a match object:

```
dna = "ACTCGTACGAAAGCTGCTTATACGCGCG"  
m = re.search(r"GC[ATGC]GC", dna)  
print "The matching string is",  
m.group()  
print "Match starts at", m.start()  
print "Match ends at", m.end()
```

RE findall()



- `findall()` returns a list of strings: `import re`

```
dna = "ACTGCATTATATCGTACGAAATTATACGCGCG"  
runs = re.findall("[AT]4,100", dna)  
print runs
```


RE findall()



- `findall()` returns a list of strings: `import re`

```
dna = "ACTGCATTATATCGTACGAAATTATACGCGCG"  
runs = re.findall("[AT]4,100", dna)  
print runs
```

output:

```
['ATTATAT', 'AAATTATA']
```

Regular Expressions Overview

Python 2.7 Regular Expressions

Non-special chars match themselves. Exceptions are special characters:

```
\  Escape special char or start a sequence.
.  Match any char except newline, see re.DOTALL
$  Match end of the string, see re.MULTILINE
^  Match start of the string, see re.MULTILINE
[]  Enclose a set of matchable chars
|  Match either regex # or regex #.
()  Create capture group, & indicate precedence
```

After '()', enclose a set, the only special chars are:

```
}  End the set, if not the last char
-  A range, eg. a-c matches a, b or c
*  Negate the set only if it is the last char
```

Quantifiers (append '?' for non-greedy):

```
[n]  Exactly n repetitions
[m,n] From m (default 0) to n (default infinity)
+  1 or more. Same as [1,]
*  0 or more. Same as [0,]
?  0 or 1. Same as [0,1]
```

Special sequences:

```
\A Start of string
\b Match empty string at word ('w+') boundary
\B Match empty string not at word boundary
\d Digit
\D Non-digit
\w Whitespace [ '\t\r\n\f\v'], see LOCAL, UNICODE
\W Non-whitespace
\w Alphanumeric: [0-9a-zA-Z_], see LOCAL
\W Non-alphanumeric
\Z End of string
\<id> Match prev named or numbered group,
  'c' & 's' are literal, e.g. '\<id>'
  or '\<name>' (not '\<0' or '\<name>')
```

Special character escapes are much like those already escaped in Python string literals. Hence regex '\n' is same as regex '\\n'.

```
\a ASCII Bell (BEL)
\f ASCII Formfeed
\n ASCII Linefeed
\r ASCII Carriage return
\t ASCII Tab
\v ASCII Vertical tab
\\ A single backslash
\wM Two digit hexadecimal character goes here
\OOO Three digit octal char (or just use an
initial zero, e.g. '\0, '\09)
\DD Decimal number 1 to 99, match
previous numbered group
```

Extensions. Do not cause grouping, except 'P'<name>':

```
(?!) Match empty string, sets re.X flags
(?:...) Non-capturing version of regular parens
(?Pname...) Create a named capturing group
(?Pname) Match whatever named prev named group
(?#...) A comment; ignored.
(?=...) Lookahead assertion, match without consuming
(?!...) Negative lookahead assertion
(?<...) Lookbehind assertion, match if preceded
(?>...) Negative lookbehind assertion
(?<id>|n) Match 'y' if group 'id' matched, else 'n'
```

Flags for re.compile(). etc. Combine with '|' :

```
re.I == re.IGNORECASE Ignore case
re.L == re.LOCAL Make 'w, s, d, and \w' locale dependent
re.M == re.MULTILINE Multiline
re.D == re.DOTALL Dot matches all (including newline)
re.S == re.DOTALL Make 'w, s, d, and \w' locale dependent
re.U == re.UNICODE Unicode (unescaped whitespace in parens
is ignored, and 'd' matches non-digits)
```

Module level functions:

```
compile(pattern[, flags]) -> RegexObject
match(pattern, string[, flags]) -> MatchObject
search(pattern, string[, flags]) -> MatchObject
findall(pattern, string[, flags]) -> list of strings
finditer(pattern, string[, flags]) -> iter of MatchObject
split(pattern, string[, maxsplit, flags]) -> list of strings
splitlines(pattern, string[, maxsplit, flags]) -> list of strings
sub(pattern, repl, string[, count, flags]) -> string
subn(pattern, repl, string[, count, flags]) -> string
submatch(pattern, repl, string[, count, flags]) -> string
submatch(pattern, repl, string[, count, flags]) -> string
```

RegexObjects (returned from compile()):

```
match(string[, pos, endpos]) -> MatchObject
search(string[, pos, endpos]) -> MatchObject
findall(string[, pos, endpos]) -> list of strings
finditer(string[, pos, endpos]) -> iter of MatchObject
split(string[, maxsplit]) -> list of strings
sub(pattern, repl, string[, count]) -> string
subn(pattern, repl, string[, count]) -> string
submatch(pattern, repl, string[, count]) -> string
submatch(pattern, repl, string[, count]) -> string
groupindex {} Map group names to ints
pattern {} String, passed to compile()
```

MatchObjects (returned from match() and search()):

```
group([group1, ...]) -> string, Matchable & group expansion
groupdict([default]) -> dict of all groups, non-matching defaults
groupindex([default]) -> {} Map group names to ints, non-matching defaults
start([group]) -> int, Match start of matching match by group
end([group]) -> int, Group default to 0, the whole match
span([group]) -> tuple (match start [group], match end [group])
pos int, passed to search() or match()
endpos int, passed to search() or match()
lastindex int, Index of last matched capturing group
lastgroup string, Name of last matched capturing group
re string, as passed to search() or match()
string string, as passed to search() or match()
```

Gleaned from the python 2.7 're' docs.

<http://docs.python.org/library/re.html>

<https://github.com/tartley/python-regex-cheatsheet>

Version: v0.3.3

Regular Expressions: Matches

Non-special chars match themselves. Exceptions are special characters:

| | |
|------------------|---|
| <code>\</code> | Escape special char or start a sequence. |
| <code>.</code> | Match any char except newline, see <code>re.DOTALL</code> |
| <code>^</code> | Match start of the string, see <code>re.MULTILINE</code> |
| <code>\$</code> | Match end of the string, see <code>re.MULTILINE</code> |
| <code>[]</code> | Enclose a set of matchable chars |
| <code>R S</code> | Match either regex R or regex S. |
| <code>()</code> | Create capture group, & indicate precedence |

After '[', enclose a set, the only special chars are:

| | |
|----------------|---|
| <code>]</code> | End the set, if not the 1st char |
| <code>-</code> | A range, eg. <code>a-c</code> matches <code>a</code> , <code>b</code> or <code>c</code> |
| <code>^</code> | Negate the set only if it is the 1st char |

Quantifiers (append '?' for non-greedy):

| | |
|--------------------|--|
| <code>{m}</code> | Exactly <code>m</code> repetitions |
| <code>{m,n}</code> | From <code>m</code> (default 0) to <code>n</code> (default infinity) |
| <code>*</code> | 0 or more. Same as <code>{,}</code> |
| <code>+</code> | 1 or more. Same as <code>{1,}</code> |
| <code>?</code> | 0 or 1. Same as <code>{,1}</code> |

Regular Expressions: Special Characters

Special character escapes are much like those already escaped in Python string literals. Hence regex `'\n'` is same as regex `'\\n'`:

```
\a  ASCII Bell (BEL)
\f  ASCII Formfeed
\n  ASCII Linefeed
\r  ASCII Carriage return
\t  ASCII Tab
\v  ASCII Vertical tab
\\  A single backslash
\xHH Two digit hexadecimal character goes here
\OOO Three digit octal char (or just use an
      initial zero, e.g. \0, \09)
\DD  Decimal number 1 to 99, match
      previous numbered group
```

Regular Expressions: Special Sequences

Special sequences:

```
\A Start of string
\b Match empty string at word (\w+) boundary
\B Match empty string not at word boundary
\d Digit
\D Non-digit
\s Whitespace [ \t\n\r\f\v], see LOCALE, UNICODE
\S Non-whitespace
\w Alphanumeric: [0-9a-zA-Z_], see LOCALE
\W Non-alphanumeric
\Z End of string
\g<id> Match prev named or numbered group,
      '<' & '>' are literal, e.g. \g<0>
      or \g<name> (not \g0 or \gname)
```

Challenges

A screenshot of a terminal window with a blue background and white text. The window title is "amnh -- vim -- 47x36". The text inside the terminal is as follows:

```
Contact Us
General Information
Info-rggs@amnh.org

Admissions
Admissions-rggs@amnh.org

Fellowship Opportunities
Fellowships-rggs@amnh.org

Webmaster
Webmaster-rggs@amnh.org

Telephone -- Main Office
212-769-5055

Office Hours
Monday to Friday
9am to 5pm and by appointment

Staff
Dr. John Flynn
Dean
212-769-5055          Beth G. Kneller
Director of Administration
212-769-5143
Maria Rios
Assistant Director for Student Affairs and Fellowships
212-769-5017          Anna Manuel
Departmental Administrator
212-769-5055
```

email.txt

In pairs:

- Design a program that will extract all the email addresses from a file.

Challenges

A screenshot of a terminal window with a blue background and white text. The window title is "amnh -- vim -- 47x36". The text inside the terminal is as follows:

```
Contact Us
General Information
Info-rggs@amnh.org

Admissions
Admissions-rggs@amnh.org

Fellowship Opportunities
Fellowships-rggs@amnh.org

Webmaster
Webmaster-rggs@amnh.org

Telephone -- Main Office
212-769-5055

Office Hours
Monday to Friday
9am to 5pm and by appointment

Staff
Dr. John Flynn
Dean
212-769-5055          Beth G. Kneller
Director of Administration
212-769-5143
Maria Rios
Assistant Director for Student Affairs and Fellowships
212-769-5017          Anna Manuel
Departmental Administrator
212-769-5055
```

email.txt

In pairs:

- Design a program that will extract all the email addresses from a file.
- First, figure out a good regular expression to match email formats.

Challenges

A screenshot of a terminal window with a blue background and white text. The window title is "amnh -- vim -- 47x36". The text inside the terminal is the content of a file named "email.txt", which lists contact information for the American Museum of Natural History (AMNH). The text is organized into sections: "Contact Us", "Admissions", "Fellowship Opportunities", "Webmaster", "Telephone -- Main Office", "Office Hours", and "Staff". The "Staff" section lists several individuals with their titles and phone numbers.

```
Contact Us
General Information
Info-rggs@amnh.org

Admissions
Admissions-rggs@amnh.org

Fellowship Opportunities
Fellowships-rggs@amnh.org

Webmaster
Webmaster-rggs@amnh.org

Telephone -- Main Office
212-769-5055

Office Hours
Monday to Friday
9am to 5pm and by appointment

Staff
Dr. John Flynn
Dean
212-769-5055          Beth G. Kneller
Director of Administration
212-769-5143
Maria Rios
Assistant Director for Student Affairs and Fell
owships
212-769-5017          Anna Manuel
Departmental Administrator
212-769-5055
```

email.txt

In pairs:

- Design a program that will extract all the email addresses from a file.
- First, figure out a good regular expression to match email formats.
- Next, use it to find the first match.

Challenges

A screenshot of a terminal window with a blue background and white text. The window title is "amnh -- vim -- 47x36". The text inside is as follows:

```
Contact Us
General Information
Info-rggs@amnh.org

Admissions
Admissions-rggs@amnh.org

Fellowship Opportunities
Fellowships-rggs@amnh.org

Webmaster
Webmaster-rggs@amnh.org

Telephone -- Main Office
212-769-5055

Office Hours
Monday to Friday
9am to 5pm and by appointment

Staff
Dr. John Flynn
Dean
212-769-5055          Beth G. Kneller
Director of Administration
212-769-5143
Maria Rios
Assistant Director for Student Affairs and Fellowships
212-769-5017          Anna Manuel
Departmental Administrator
212-769-5055
```

email.txt

In pairs:

- Design a program that will extract all the email addresses from a file.
- First, figure out a good regular expression to match email formats.
- Next, use it to find the first match.
- Last, add a loop to find all.

CSV Files

| CATALOG NUMBER | IDENTIFICATION | COUNTRY | STATE | COUNTY | PRECISE LOCATION | OCEAN | ISLAND GROUP | ISLAND | LATITUDE | LONGITUDE | COLLECTOR | COLLECTING DATE | COLLECTING SITE |
|----------------|----------------|-------------|----------|---------|------------------|---------------|--------------|--------|------------|-------------|---------------|-----------------|-----------------|
| DOT 84 | FLUJ | Ceyx | tepidus | Solomon | Isla New Georgia | Vella Lavella | Oula River | camp | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 7 May-04 | PRS-2692 |
| DOT 85 | SKEL | Ceyx | tepidus | Solomon | Isla New Georgia | Vella Lavella | Oula River | camp | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 6 May-04 | PRS-2663 |
| DOT 92 | SKEL | Ceyx | tepidus | Solomon | Isla New Georgia | Vella Lavella | Oula River | camp | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 7 May-04 | PRS-2673 |
| DOT 93 | SKEL | Alcedo | atthis | Solomon | Isla New Georgia | Vella Lavella | Oula River | camp | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 6 May-04 | PRS-2656 |
| DOT 95 | SKEL | Alcedo | atthis | Solomon | Isla New Georgia | Vella Lavella | Oula River | camp | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 5 May-04 | PRS-2637 |
| DOT 98 | SKEL | Ceyx | tepidus | Solomon | Isla New Georgia | Vella Lavella | Oula River | camp | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 8 May-04 | PRS-2693 |
| DOT 102 | FLUJ | Ceyx | tepidus | Solomon | Isla New Georgia | Vella Lavella | Oula River | camp | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 5 May-04 | PRS-2647 |
| DOT 104 | | Todiramphus | | Solomon | Isla New Georgia | Vella Lavella | Oula River | camp | 07 47 30 S | 156 37 30 E | | 4 May-04 | PRS-2630 |
| DOT 109 | | Alcedo | atthis | Solomon | Isla New Georgia | Vella Lavella | Oula River | camp | 07 47 30 S | 156 37 30 E | | 6 May-04 | PRS-2617 |
| DOT 116 | | Ceyx | pusillus | Solomon | Isla New Georgia | Vella Lavella | Oula River | camp | 07 47 30 S | 156 37 30 E | | 7 May-04 | PRS-2667 |

- Very structured– the columns and rows matter.

CSV Files

| CATALOG NUMBER | IDENTIFICATION | COUNTRY | STATE | COUNTY | PRECISE LOCATION | OCEAN | ISLAND GROUP | ISLAND | LATITUDE | LONGITUDE | COLLECTOR | COLLECTING DATE | COLLECTING SITE |
|----------------|----------------|---------------|----------------|---------------------------|------------------|-------|--------------|--------|------------|-------------|---------------|-----------------|-----------------|
| DOT 84 | FLUJ | Ceyx teptidus | Solomon Island | New Georgia Vella Lavella | Oula River camp | | | | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 7-May-04 | PRS-2692 |
| DOT 85 | SKEL | Ceyx teptidus | Solomon Island | New Georgia Vella Lavella | Oula River camp | | | | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 6-May-04 | PRS-2668 |
| DOT 92 | SKEL | Ceyx teptidus | Solomon Island | New Georgia Vella Lavella | Oula River camp | | | | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 7-May-04 | PRS-2675 |
| DOT 93 | SKEL | Alcedo atthis | Solomon Island | New Georgia Vella Lavella | Oula River camp | | | | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 6-May-04 | PRS-2656 |
| DOT 95 | SKEL | Alcedo atthis | Solomon Island | New Georgia Vella Lavella | Oula River camp | | | | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 5-May-04 | PRS-2637 |
| DOT 98 | SKEL | Ceyx teptidus | Solomon Island | New Georgia Vella Lavella | Oula River camp | | | | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 8-May-04 | PRS-2693 |
| DOT 102 | FLUJ | Ceyx teptidus | Solomon Island | New Georgia Vella Lavella | Oula River camp | | | | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 5-May-04 | PRS-2647 |
| DOT 104 | | Todiramphus | Solomon Island | New Georgia Vella Lavella | Oula River camp | | | | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 4-May-04 | PRS-2630 |
| DOT 109 | | Alcedo atthis | Solomon Island | New Georgia Vella Lavella | Oula River camp | | | | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 6-May-04 | PRS-2657 |
| DOT 116 | | Ceyx pusillus | Solomon Island | New Georgia Vella Lavella | Oula River camp | | | | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 7-May-04 | PRS-2667 |

- Very structured– the columns and rows matter.
- To keep that format as a text file:

CSV Files

| CATALOG NUMBER | IDENTIFICATION | COUNTRY | STATE | COUNTY | PRECISE LOCATION | OCEAN | ISLAND GROUP | ISLAND | LATITUDE | LONGITUDE | COLLECTOR(S) | COLLECTING DATE | COLLECTING SITE |
|----------------|----------------|---------|-------------|--------------|------------------|---------------|-----------------|-----------------|------------|-------------|---------------|-----------------|-----------------|
| 1 | DOT 84 | FLUJ | Ceyx | tepidus | Solomon Isia | New Georgia | Vella Lavella | Oula River camp | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 7-May-04 | PRS-2692 |
| 2 | DOT 85 | SKEL | Ceyx | tepidus | Solomon Isia | New Georgia | Vella Lavella | Oula River camp | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 6-May-04 | PRS-2693 |
| 3 | DOT 92 | SKEL | Ceyx | tepidus | Solomon Isia | New Georgia | Vella Lavella | Oula River camp | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 7-May-04 | PRS-2694 |
| 4 | DOT 93 | SKEL | Alcedo | atthis | Solomon Isia | New Georgia | Vella Lavella | Oula River camp | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 6-May-04 | PRS-2695 |
| 5 | DOT 95 | SKEL | Alcedo | atthis | Solomon Isia | New Georgia | Vella Lavella | Oula River camp | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 5-May-04 | PRS-2696 |
| 6 | DOT 98 | SKEL | Ceyx | tepidus | Solomon Isia | New Georgia | Vella Lavella | Oula River camp | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 8-May-04 | PRS-2697 |
| 7 | DOT 102 | FLUJ | Ceyx | tepidus | Solomon Isia | New Georgia | Vella Lavella | Oula River camp | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 5-May-04 | PRS-2698 |
| 8 | DOT 104 | | Todiramphus | Solomon Isia | New Georgia | Vella Lavella | Oula River camp | | 07 47 30 S | 156 37 30 E | | 4-May-04 | PRS-2699 |
| 9 | DOT 109 | | Alcedo | atthis | Solomon Isia | New Georgia | Vella Lavella | Oula River camp | 07 47 30 S | 156 37 30 E | | 6-May-04 | PRS-2700 |
| 10 | DOT 116 | | Ceyx | pusillus | Solomon Isia | New Georgia | Vella Lavella | Oula River camp | 07 47 30 S | 156 37 30 E | | 7-May-04 | PRS-2701 |

- Very structured– the columns and rows matter.
- To keep that format as a text file:
 - ▶ columns separated by commas (',') and

CSV Files

| CATALOG# | IDENTIFICATION | COUNTRY | STATE | COUNTY | PRECISE LOC | OCEAN | ISLAND GROUP | ISLAND | LATITUDE | LONGITUDE | COLLECTOR | COLLECTING | COLLECTING |
|----------|----------------|---------------|-----------------|---------------------------|-----------------|-------|--------------|--------|------------|-------------|---------------|------------|------------|
| DOT 84 | FLUID 11383 | Ceyx lepidus | Solomon Islands | New Georgia Vella Lavella | Oula River camp | | | | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 7-May-04 | PRS-2672 |
| DOT 85 | SKEL | Ceyx lepidus | Solomon Islands | New Georgia Vella Lavella | Oula River camp | | | | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 6-May-04 | PRS-2668 |
| DOT 92 | SKEL | Ceyx lepidus | Solomon Islands | New Georgia Vella Lavella | Oula River camp | | | | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 7-May-04 | PRS-2673 |
| DOT 93 | SKEL | Alcedo atthis | Solomon Islands | New Georgia Vella Lavella | Oula River camp | | | | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 6-May-04 | PRS-2656 |
| DOT 95 | SKEL | Alcedo atthis | Solomon Islands | New Georgia Vella Lavella | Oula River camp | | | | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 5-May-04 | PRS-2637 |
| DOT 98 | SKEL | Ceyx lepidus | Solomon Islands | New Georgia Vella Lavella | Oula River camp | | | | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 8-May-04 | PRS-2692 |
| DOT 102 | FLUID | Ceyx lepidus | Solomon Islands | New Georgia Vella Lavella | Oula River camp | | | | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 5-May-04 | PRS-2647 |
| DOT 104 | | Todiramphus | Solomon Islands | New Georgia Vella Lavella | Oula River camp | | | | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 4-May-04 | PRS-2630 |
| DOT 109 | | Alcedo atthis | Solomon Islands | New Georgia Vella Lavella | Oula River camp | | | | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 6-May-04 | PRS-2657 |
| DOT 116 | | Ceyx pusillus | Solomon Islands | New Georgia Vella Lavella | Oula River camp | | | | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 7-May-04 | PRS-2667 |

- Very structured– the columns and rows matter.
- To keep that format as a text file:
 - ▶ columns separated by commas (',') and
 - ▶ rows separated by new lines ('\n')
- Rows look like:
"DOT 84 FLUID 11383",Ceyx lepidus collectoris,Solomon Islands,New Georgia Group,Vella Lavella Island,Oula River camp,,,07 47 30 S,156 37 30 E,Paul R. Sweet,7-May-04,,PRS-2672,,,"Tissue Fluid "

CSV Files

AMNH-Ornithology-Internet-Export.csv

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O |
|----|-------------|----------------|---------------|--------------|---------------------------|-----------------|-------|--------------|--------|------------|-------------|---------------|-----------------|--------------|-----------------|
| | CATALOG NO. | IDENTIFICATION | COUNTRY | STATE | COUNTY | PRECISE LOC. | OCEAN | ISLAND GROUP | ISLAND | LATITUDE | LONGITUDE | COLLECTOR(S) | COLLECTING DATE | COLLECTOR(S) | COLLECTING DATE |
| 1 | DOT 84 | FLUJ | Ceyx teptidus | Solomon Isia | New Georgia Vella Lavella | Oula River camp | | | | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 7-May-04 | | PRS-2672 |
| 2 | DOT 85 | SKEI | Ceyx teptidus | Solomon Isia | New Georgia Vella Lavella | Oula River camp | | | | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 6-May-04 | | PRS-2663 |
| 3 | DOT 92 | SKEI | Ceyx teptidus | Solomon Isia | New Georgia Vella Lavella | Oula River camp | | | | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 7-May-04 | | PRS-2673 |
| 4 | DOT 93 | SKEI | Alcedo atthis | Solomon Isia | New Georgia Vella Lavella | Oula River camp | | | | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 6-May-04 | | PRS-2655 |
| 5 | DOT 95 | SKEI | Alcedo atthis | Solomon Isia | New Georgia Vella Lavella | Oula River camp | | | | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 5-May-04 | | PRS-2637 |
| 6 | DOT 98 | SKEI | Ceyx teptidus | Solomon Isia | New Georgia Vella Lavella | Oula River camp | | | | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 8-May-04 | | PRS-2692 |
| 7 | DOT 102 | FLUJ | Ceyx teptidus | Solomon Isia | New Georgia Vella Lavella | Oula River camp | | | | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 5-May-04 | | PRS-2647 |
| 8 | DOT 104 | | Todiramphus | Solomon Isia | New Georgia Vella Lavella | Oula River camp | | | | 07 47 30 S | 156 37 30 E | | 4-May-04 | | PRS-2631 |
| 9 | DOT 109 | | Alcedo atthis | Solomon Isia | New Georgia Vella Lavella | Oula River camp | | | | 07 47 30 S | 156 37 30 E | | 6-May-04 | | PRS-2657 |
| 10 | DOT 116 | | Ceyx pusillus | Solomon Isia | New Georgia Vella Lavella | Oula River camp | | | | 07 47 30 S | 156 37 30 E | | 7-May-04 | | PRS-2667 |

- Built-in package for reading CSV files. To use it:

CSV Files

AMNH-Ornithology-Internet-Export.csv

100%

Search in Sheet

Home
 Layout
 Tables
 Charts
 SmartArt
 Formulas
 Data
 Review

Edit
 Font
 Alignment
 Number

Paste
 Bold
 Italic
 Underline
 Text Color
 Fill Color
 Background Color
 Merge
 Wrap Text
 Conditional Formatting
 Styles
 Actions
 Themes

| L1 | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O |
|----|-------------|----------------|-------------|----------|---------|--------------|-------------|---------------|------|------------|------------|-------------|---------------|-----------------|--------------|
| | CATALOG NO. | IDENTIFICATION | COUNTRY | STATE | COUNTY | PRECISE LOC. | OCEAN | ISLAND | GROU | ISLAND | LATITUDE | LONGITUDE | COLLECTOR(S) | COLLECTING DATE | COLLECTOR(S) |
| 1 | DOT 84 | FLUJ | Ceyx | tepidus | Solomon | Isla | New Georgia | Vella Lavella | Oula | River camp | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 7-May-04 | PRS-2672 |
| 2 | DOT 85 | SKEI | Ceyx | tepidus | Solomon | Isla | New Georgia | Vella Lavella | Oula | River camp | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 6-May-04 | PRS-2663 |
| 3 | DOT 92 | SKEI | Ceyx | tepidus | Solomon | Isla | New Georgia | Vella Lavella | Oula | River camp | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 7-May-04 | PRS-2673 |
| 4 | DOT 93 | SKEI | Alcedo | atthis | Solomon | Isla | New Georgia | Vella Lavella | Oula | River camp | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 6-May-04 | PRS-2655 |
| 5 | DOT 95 | SKEI | Alcedo | atthis | Solomon | Isla | New Georgia | Vella Lavella | Oula | River camp | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 5-May-04 | PRS-2637 |
| 6 | DOT 98 | SKEI | Ceyx | tepidus | Solomon | Isla | New Georgia | Vella Lavella | Oula | River camp | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 8-May-04 | PRS-2692 |
| 7 | DOT 102 | FLUJ | Ceyx | tepidus | Solomon | Isla | New Georgia | Vella Lavella | Oula | River camp | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 5-May-04 | PRS-2647 |
| 8 | DOT 104 | | Todiramphus | | Solomon | Isla | New Georgia | Vella Lavella | Oula | River camp | 07 47 30 S | 156 37 30 E | | 4-May-04 | PRS-2630 |
| 9 | DOT 109 | | Alcedo | atthis | Solomon | Isla | New Georgia | Vella Lavella | Oula | River camp | 07 47 30 S | 156 37 30 E | | 6-May-04 | PRS-2657 |
| 10 | DOT 116 | | Ceyx | pusillus | Solomon | Isla | New Georgia | Vella Lavella | Oula | River camp | 07 47 30 S | 156 37 30 E | | 7-May-04 | PRS-2667 |

- Built-in package for reading CSV files. To use it:
 - At top of file, include: `import csv`

CSV Files

| A | B | C | D | E | F | G | H | I | J | K | L | M | N | O |
|----|---------|-------------------|---------------|--------------|-------------|---------------|-----------------|---|------------|-------------|---------------|----------|---|----------|
| 1 | CATALOG | ML IDENTIFICATION | COUNTRY | STATE | COUNTY | PRECISE LOC. | OCEAN | | | | | | | |
| 2 | DOT 84 | FLUJ | Ceyx tepidus | Solomon Isia | New Georgia | Vella Lavella | Oula River camp | | | | | | | |
| 3 | DOT 85 | SKEI | Ceyx tepidus | Solomon Isia | New Georgia | Vella Lavella | Oula River camp | | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 7-May-04 | | PRS-2672 |
| 4 | DOT 92 | SKEI | Ceyx tepidus | Solomon Isia | New Georgia | Vella Lavella | Oula River camp | | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 6-May-04 | | PRS-2663 |
| 5 | DOT 93 | SKEI | Alcedo atthis | Solomon Isia | New Georgia | Vella Lavella | Oula River camp | | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 7-May-04 | | PRS-2673 |
| 6 | DOT 95 | SKEI | Alcedo atthis | Solomon Isia | New Georgia | Vella Lavella | Oula River camp | | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 6-May-04 | | PRS-2655 |
| 7 | DOT 98 | SKEI | Ceyx tepidus | Solomon Isia | New Georgia | Vella Lavella | Oula River camp | | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 5-May-04 | | PRS-2637 |
| 8 | DOT 102 | FLUJ | Ceyx tepidus | Solomon Isia | New Georgia | Vella Lavella | Oula River camp | | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 8-May-04 | | PRS-2692 |
| 9 | DOT 104 | | Todiramphus | Solomon Isia | New Georgia | Vella Lavella | Oula River camp | | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 5-May-04 | | PRS-2647 |
| 10 | DOT 109 | | Alcedo atthis | Solomon Isia | New Georgia | Vella Lavella | Oula River camp | | 07 47 30 S | 156 37 30 E | | 4-May-04 | | PRS-2631 |
| 11 | DOT 116 | | Ceyx pusillus | Solomon Isia | New Georgia | Vella Lavella | Oula River camp | | 07 47 30 S | 156 37 30 E | | 6-May-04 | | PRS-2657 |
| | | | | | | | | | 07 47 30 S | 156 37 30 E | | 7-May-04 | | PRS-2667 |

- Built-in package for reading CSV files. To use it:
 - ▶ At top of file, include: `import csv`
 - ▶ Open the file normally: `f = open("in.csv", "rU")`

CSV Files

AMNH-Ornithology-Internet-Export.csv

Home Layout Tables Charts SmartArt Formulas Data Review View 100% Search in Sheet

Font:Calibri (Body) 12 Alignment:General Number:Format:Cells:Themes:Wrap Text

Paste: B I U Conditional Formatting Styles Actions Themes Aa

| L1 | C | D | E | F | G | H | I | J | K | L | M | N | O | | | |
|----|---------|--------|----------------|---------|-------|-------------|---------------|-------|--------|--------|------------|-------------|---------------|--------------|-----------------|-------------------|
| | CATALOG | NUMBER | IDENTIFICATION | COUNTRY | STATE | COUNTY | PRECISE LOC | OCEAN | ISLAND | GROUPE | ISLAND | LATITUDE | LONGITUDE | COLLECTOR(S) | COLLECTING DATE | COLLECTING METHOD |
| 2 | DOT 84 | FLUJ | Ceyx telegonus | Solomon | Isla | New Georgia | Vella Lavella | Oula | River | camp | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 7-May-04 | PRS-2672 | |
| 3 | DOT 85 | SKEI | Ceyx telegonus | Solomon | Isla | New Georgia | Vella Lavella | Oula | River | camp | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 6-May-04 | PRS-2663 | |
| 4 | DOT 92 | SKEI | Ceyx telegonus | Solomon | Isla | New Georgia | Vella Lavella | Oula | River | camp | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 7-May-04 | PRS-2673 | |
| 5 | DOT 93 | SKEI | Alcedo atthis | Solomon | Isla | New Georgia | Vella Lavella | Oula | River | camp | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 6-May-04 | PRS-2655 | |
| 6 | DOT 95 | SKEI | Alcedo atthis | Solomon | Isla | New Georgia | Vella Lavella | Oula | River | camp | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 5-May-04 | PRS-2637 | |
| 7 | DOT 98 | SKEI | Ceyx telegonus | Solomon | Isla | New Georgia | Vella Lavella | Oula | River | camp | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 8-May-04 | PRS-2692 | |
| 8 | DOT 102 | FLUJ | Ceyx telegonus | Solomon | Isla | New Georgia | Vella Lavella | Oula | River | camp | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 5-May-04 | PRS-2647 | |
| 9 | DOT 104 | | Todiramphus | Solomon | Isla | New Georgia | Vella Lavella | Oula | River | camp | 07 47 30 S | 156 37 30 E | | 4-May-04 | PRS-2630 | |
| 10 | DOT 109 | | Alcedo atthis | Solomon | Isla | New Georgia | Vella Lavella | Oula | River | camp | 07 47 30 S | 156 37 30 E | | 6-May-04 | PRS-2657 | |
| 11 | DOT 116 | | Ceyx pusillus | Solomon | Isla | New Georgia | Vella Lavella | Oula | River | camp | 07 47 30 S | 156 37 30 E | | 7-May-04 | PRS-2667 | |

- Built-in package for reading CSV files. To use it:
 - ▶ At top of file, include: `import csv`
 - ▶ Open the file normally: `f = open("in.csv", "rU")`
"rU" avoids errors with different newlines, and accepts all variants.

CSV Files

| L1 | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O |
|----|--------------|----------------|--------------|-------------|---------------|--------------|-------|--------|-----|--------|------------|-------------|---------------|------------|----------|
| | CATALOG NO. | IDENTIFICATION | COUNTRY | STATE | COUNTY | PRECISE LOC. | OCEAN | ISLAND | GRO | ISLAND | LATITUDE | LONGITUDE | COLLECTOR | COLLECTING | COLLECT |
| 1 | CATALOG NO. | IDENTIFICATION | COUNTRY | STATE | COUNTY | PRECISE LOC. | OCEAN | ISLAND | GRO | ISLAND | LATITUDE | LONGITUDE | COLLECTOR | COLLECTING | COLLECT |
| 2 | DOT 84 FLUJ | Ceyx teptidus | Solomon Isia | New Georgia | Vella Lavella | Oula River | camp | | | | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 7-May-04 | PRS-2672 |
| 3 | DOT 85 SKEL | Ceyx teptidus | Solomon Isia | New Georgia | Vella Lavella | Oula River | camp | | | | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 6-May-04 | PRS-2663 |
| 4 | DOT 92 SKEL | Ceyx teptidus | Solomon Isia | New Georgia | Vella Lavella | Oula River | camp | | | | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 7-May-04 | PRS-2673 |
| 5 | DOT 93 SKEL | Alcedo atthis | Solomon Isia | New Georgia | Vella Lavella | Oula River | camp | | | | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 6-May-04 | PRS-2655 |
| 6 | DOT 95 SKEL | Alcedo atthis | Solomon Isia | New Georgia | Vella Lavella | Oula River | camp | | | | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 5-May-04 | PRS-2637 |
| 7 | DOT 98 SKEL | Ceyx teptidus | Solomon Isia | New Georgia | Vella Lavella | Oula River | camp | | | | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 8-May-04 | PRS-2692 |
| 8 | DOT 102 FLUJ | Ceyx teptidus | Solomon Isia | New Georgia | Vella Lavella | Oula River | camp | | | | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 5-May-04 | PRS-2647 |
| 9 | DOT 104 | Todiramphus | Solomon Isia | New Georgia | Vella Lavella | Oula River | camp | | | | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 4-May-04 | PRS-2630 |
| 10 | DOT 109 | Alcedo atthis | Solomon Isia | New Georgia | Vella Lavella | Oula River | camp | | | | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 6-May-04 | PRS-2657 |
| 11 | DOT 116 | Ceyx pusillus | Solomon Isia | New Georgia | Vella Lavella | Oula River | camp | | | | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 7-May-04 | PRS-2667 |

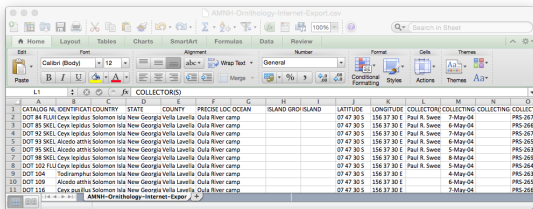
- Built-in package for reading CSV files. To use it:
 - ▶ At top of file, include: `import csv`
 - ▶ Open the file normally: `f = open("in.csv", "rU")`
"rU" avoids errors with different newlines, and accepts all variants.
 - ▶ Create a reader: `reader = csv.DictReader(f)`

CSV Files

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O |
|----|---------|----------------|---------------|--------------|---------------------------|-----------------|-------|--------|------|--------|------------|-------------|---------------|------------|------------|
| | CATALOG | IDENTIFICATION | COUNTRY | STATE | COUNTY | PRECISE LOC. | OCEAN | ISLAND | GRP. | ISLAND | LATITUDE | LONGITUDE | COLLECTOR | COLLECTING | COLLECTION |
| 1 | DOT 84 | FLUJ | Ceyx teptidus | Solomon Isia | New Georgia Vella Lavella | Oula River camp | | | | | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 7-May-04 | PRS-2672 |
| 2 | DOT 85 | SKEI | Ceyx teptidus | Solomon Isia | New Georgia Vella Lavella | Oula River camp | | | | | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 6-May-04 | PRS-2663 |
| 3 | DOT 92 | SKEI | Ceyx teptidus | Solomon Isia | New Georgia Vella Lavella | Oula River camp | | | | | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 7-May-04 | PRS-2673 |
| 4 | DOT 93 | SKEI | Alcedo atthis | Solomon Isia | New Georgia Vella Lavella | Oula River camp | | | | | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 6-May-04 | PRS-2655 |
| 5 | DOT 95 | SKEI | Alcedo atthis | Solomon Isia | New Georgia Vella Lavella | Oula River camp | | | | | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 5-May-04 | PRS-2637 |
| 6 | DOT 98 | SKEI | Ceyx teptidus | Solomon Isia | New Georgia Vella Lavella | Oula River camp | | | | | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 8-May-04 | PRS-2692 |
| 7 | DOT 102 | FLUJ | Ceyx teptidus | Solomon Isia | New Georgia Vella Lavella | Oula River camp | | | | | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 5-May-04 | PRS-2647 |
| 8 | DOT 104 | | Todiramphus | Solomon Isia | New Georgia Vella Lavella | Oula River camp | | | | | 07 47 30 S | 156 37 30 E | | 4-May-04 | PRS-2630 |
| 9 | DOT 109 | | Alcedo atthis | Solomon Isia | New Georgia Vella Lavella | Oula River camp | | | | | 07 47 30 S | 156 37 30 E | | 6-May-04 | PRS-2657 |
| 10 | DOT 116 | | Ceyx pusillus | Solomon Isia | New Georgia Vella Lavella | Oula River camp | | | | | 07 47 30 S | 156 37 30 E | | 7-May-04 | PRS-2667 |

- Built-in package for reading CSV files. To use it:
 - ▶ At top of file, include: `import csv`
 - ▶ Open the file normally: `f = open("in.csv", "rU")`
"rU" avoids errors with different newlines, and accepts all variants.
 - ▶ Create a reader: `reader = csv.DictReader(f)`
Uses column names in first line of csv file to access row data.

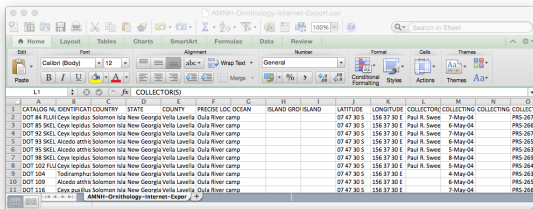
CSV Files



| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O |
|----|----------|----------------|---------|---------|------------|-------------|-------------|-------------|-------------|---------|----------|-----------|-----------|------------|----------|
| | | | | | | | | | | | | | | | |
| 1 | CATALOG# | IDENTIFICATION | COUNTRY | STATE | COUNTY | PRECISE LOC | OCEAN | ISLAND | GRO | ISLAND | LATITUDE | LONGITUDE | COLLECTOR | COLLECTING | COLLECT |
| 2 | DOT 84 | FLUJ | Cey | tepidus | Solomon | Isla | New Georgia | Vella | Lavella | Oula | River | camp | | | PRS-2672 |
| 3 | DOT 85 | SKE | Cey | tepidus | Solomon | Isla | New Georgia | Vella | Lavella | Oula | River | camp | | | PRS-2663 |
| 4 | DOT 92 | SKE | Cey | tepidus | Solomon | Isla | New Georgia | Vella | Lavella | Oula | River | camp | | | PRS-2673 |
| 5 | DOT 93 | SKE | Alcedo | atthis | Solomon | Isla | New Georgia | Vella | Lavella | Oula | River | camp | | | PRS-2655 |
| 6 | DOT 95 | SKE | Alcedo | atthis | Solomon | Isla | New Georgia | Vella | Lavella | Oula | River | camp | | | PRS-2637 |
| 7 | DOT 98 | SKE | Cey | tepidus | Solomon | Isla | New Georgia | Vella | Lavella | Oula | River | camp | | | PRS-2692 |
| 8 | DOT 102 | FLUJ | Cey | tepidus | Solomon | Isla | New Georgia | Vella | Lavella | Oula | River | camp | | | PRS-2647 |
| 9 | DOT 104 | | | | Todivamphu | Solomon | Isla | New Georgia | Vella | Lavella | Oula | River | camp | | PRS-2631 |
| 10 | DOT 109 | | | | Alcedo | atthis | Solomon | Isla | New Georgia | Vella | Lavella | Oula | River | camp | PRS-2657 |
| 11 | DOT 116 | | | | Cey | pusillus | Solomon | Isla | New Georgia | Vella | Lavella | Oula | River | camp | PRS-2657 |

- Built-in package for reading CSV files. To use it:
 - ▶ At top of file, include: `import csv`
 - ▶ Open the file normally: `f = open("in.csv", "rU")`
"rU" avoids errors with different newlines, and accepts all variants.
 - ▶ Create a reader: `reader = csv.DictReader(f)`
Uses column names in first line of csv file to access row data.
 - ▶ Read in lines from reader:

CSV Files



| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O |
|----|----------|----------------|---------|---------|------------|--------------|-------------|-------------|-------------|---------|----------|-----------|--------------|------------|----------|
| | | | | | | | | | | | | | | | |
| 1 | CATALOG# | IDENTIFICATION | COUNTRY | STATE | COUNTY | PRECISE LOC. | OCEAN | ISLAND | GRO | ISLAND | LATITUDE | LONGITUDE | COLLECTOR(S) | COLLECTING | COLLECT |
| 2 | DOT 84 | FLUJ | Cey | tepidus | Solomon | Isla | New Georgia | Vella | Lavella | Oula | River | camp | | | PRS-2672 |
| 3 | DOT 85 | SKEI | Cey | tepidus | Solomon | Isla | New Georgia | Vella | Lavella | Oula | River | camp | | | PRS-2663 |
| 4 | DOT 92 | SKEI | Cey | tepidus | Solomon | Isla | New Georgia | Vella | Lavella | Oula | River | camp | | | PRS-2673 |
| 5 | DOT 93 | SKEI | Alcedo | atthis | Solomon | Isla | New Georgia | Vella | Lavella | Oula | River | camp | | | PRS-2655 |
| 6 | DOT 95 | SKEI | Alcedo | atthis | Solomon | Isla | New Georgia | Vella | Lavella | Oula | River | camp | | | PRS-2637 |
| 7 | DOT 98 | SKEI | Cey | tepidus | Solomon | Isla | New Georgia | Vella | Lavella | Oula | River | camp | | | PRS-2692 |
| 8 | DOT 102 | FLUJ | Cey | tepidus | Solomon | Isla | New Georgia | Vella | Lavella | Oula | River | camp | | | PRS-2647 |
| 9 | DOT 104 | | | | Todivamphu | Solomon | Isla | New Georgia | Vella | Lavella | Oula | River | camp | | PRS-2630 |
| 10 | DOT 109 | | | | Alcedo | atthis | Solomon | Isla | New Georgia | Vella | Lavella | Oula | River | camp | PRS-2657 |
| 11 | DOT 116 | | | | Cey | pusillus | Solomon | Isla | New Georgia | Vella | Lavella | Oula | River | camp | PRS-2657 |

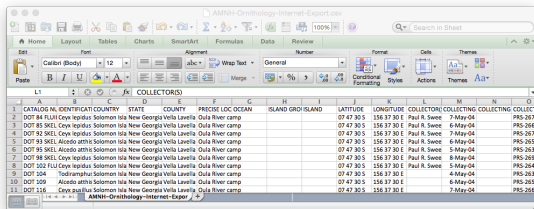
- Built-in package for reading CSV files. To use it:
 - ▶ At top of file, include: `import csv`
 - ▶ Open the file normally: `f = open("in.csv", "rU")`
"rU" avoids errors with different newlines, and accepts all variants.
 - ▶ Create a reader: `reader = csv.DictReader(f)`
Uses column names in first line of csv file to access row data.
 - ▶ Read in lines from reader: `for row in reader:`

CSV Files

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O |
|----|---------|--------|----------------|----------------|-------------|---------------|------------------|--------------|--------|------------|-------------|---------------|-----------------|-------------------|----------|
| | CATALOG | NUMBER | IDENTIFICATION | COUNTRY | STATE | COUNTY | PRECISE LOCATION | ISLAND GROUP | ISLAND | LATITUDE | LONGITUDE | COLLECTOR | COLLECTING DATE | COLLECTION NUMBER | |
| 1 | DOT 84 | FLUJ | Ceyx teptidus | Solomon Island | New Georgia | Vella Lavella | Oula River camp | | | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 7-May-04 | | PRS-2672 |
| 2 | DOT 85 | SKEI | Ceyx teptidus | Solomon Island | New Georgia | Vella Lavella | Oula River camp | | | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 6-May-04 | | PRS-2663 |
| 3 | DOT 92 | SKEI | Ceyx teptidus | Solomon Island | New Georgia | Vella Lavella | Oula River camp | | | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 7-May-04 | | PRS-2673 |
| 4 | DOT 93 | SKEI | Alcedo atthis | Solomon Island | New Georgia | Vella Lavella | Oula River camp | | | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 6-May-04 | | PRS-2655 |
| 5 | DOT 95 | SKEI | Alcedo atthis | Solomon Island | New Georgia | Vella Lavella | Oula River camp | | | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 5-May-04 | | PRS-2637 |
| 6 | DOT 98 | SKEI | Ceyx teptidus | Solomon Island | New Georgia | Vella Lavella | Oula River camp | | | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 8-May-04 | | PRS-2692 |
| 7 | DOT 102 | FLUJ | Ceyx teptidus | Solomon Island | New Georgia | Vella Lavella | Oula River camp | | | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 5-May-04 | | PRS-2647 |
| 8 | DOT 104 | | Todiramphus | Solomon Island | New Georgia | Vella Lavella | Oula River camp | | | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 4-May-04 | | PRS-2630 |
| 9 | DOT 109 | | Alcedo atthis | Solomon Island | New Georgia | Vella Lavella | Oula River camp | | | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 6-May-04 | | PRS-2657 |
| 10 | DOT 116 | | Ceyx pusillus | Solomon Island | New Georgia | Vella Lavella | Oula River camp | | | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 7-May-04 | | PRS-2667 |

- Built-in package for reading CSV files. To use it:
 - ▶ At top of file, include: `import csv`
 - ▶ Open the file normally: `f = open("in.csv", "rU")`
"rU" avoids errors with different newlines, and accepts all variants.
 - ▶ Create a reader: `reader = csv.DictReader(f)`
Uses column names in first line of csv file to access row data.
 - ▶ Read in lines from reader: `for row in reader:`
 - ▶ To access individual entries in a row:

CSV Files



| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O |
|----|---------|----------------|--------------|----------|---------|-------------|---------|---------|---------|---------|----------|-----------|-----------|------------|-------------|
| | CATALOG | IDENTIFICATION | COUNTRY | STATE | COUNTY | PRECISE LOC | OCEAN | ISLAND | GROUP | ISLAND | LATITUDE | LONGITUDE | COLLECTOR | COLLECTING | COLLECTOR'S |
| 1 | DOT 84 | FLUJ | Cey | tepidus | Solomon | Isla | New | Georgia | Vella | Lavella | Oula | River | camp | | PRS-2672 |
| 2 | DOT 85 | SKEI | Cey | tepidus | Solomon | Isla | New | Georgia | Vella | Lavella | Oula | River | camp | | PRS-2663 |
| 3 | DOT 92 | SKEI | Cey | tepidus | Solomon | Isla | New | Georgia | Vella | Lavella | Oula | River | camp | | PRS-2673 |
| 4 | DOT 93 | SKEI | Alcedo | atthi | Solomon | Isla | New | Georgia | Vella | Lavella | Oula | River | camp | | PRS-2655 |
| 5 | DOT 95 | SKEI | Alcedo | atthi | Solomon | Isla | New | Georgia | Vella | Lavella | Oula | River | camp | | PRS-2637 |
| 6 | DOT 98 | SKEI | Cey | tepidus | Solomon | Isla | New | Georgia | Vella | Lavella | Oula | River | camp | | PRS-2692 |
| 7 | DOT 102 | FLUJ | Cey | tepidus | Solomon | Isla | New | Georgia | Vella | Lavella | Oula | River | camp | | PRS-2647 |
| 8 | DOT 104 | | Todivamphuzi | Solomon | Isla | New | Georgia | Vella | Lavella | Oula | River | camp | | | PRS-2630 |
| 9 | DOT 109 | | Alcedo | atthi | Solomon | Isla | New | Georgia | Vella | Lavella | Oula | River | camp | | PRS-2657 |
| 10 | DOT 116 | | Cey | pusillus | Solomon | Isla | New | Georgia | Vella | Lavella | Oula | River | camp | | PRS-2657 |

- Built-in package for reading CSV files. To use it:
 - ▶ At top of file, include: `import csv`
 - ▶ Open the file normally: `f = open("in.csv", "rU")`
"rU" avoids errors with different newlines, and accepts all variants.
 - ▶ Create a reader: `reader = csv.DictReader(f)`
Uses column names in first line of csv file to access row data.
 - ▶ Read in lines from reader: `for row in reader:`
 - ▶ To access individual entries in a row:
`if "Malaysia" in row['COUNTRY']: ...`

CSV Files

AMNH-Ornithology-Internet-Export.csv

100%

Search in Sheet

Home Layout Tables Charts SmartArt Formulas Data Review

Edit Font Alignment Number Format Cells Themes

Paste B I U Bold Italic Underline Color Fill Background Color Merge & Center Conditional Formatting Styles Actions Themes Aa

| L1 | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O |
|----|-------------|----------------|-------------|--------------|---------|-------------|---------------|---------------|-------|-------|------|---|---|---|---------|
| | | | | COLLECTOR(S) | | | | | | | | | | | |
| 1 | CATALOG NO. | IDENTIFICATION | COUNTRY | STATE | COUNTY | PRESICE LOC | OCEAN | | | | | | | | |
| 2 | DOT 84 | FLU1 | Cyca | reipidus | Solomon | Isla | New Georgia | Vella Lavella | Oula | River | camp | | | | PMS-261 |
| 3 | DOT 85 | SKEL | Cyca | reipidus | Solomon | Isla | New Georgia | Vella Lavella | Oula | River | camp | | | | PMS-261 |
| 4 | DOT 92 | SKEL | Cyca | reipidus | Solomon | Isla | New Georgia | Vella Lavella | Oula | River | camp | | | | PMS-261 |
| 5 | DOT 93 | SKEL | Alcedo | atthis | Solomon | Isla | New Georgia | Vella Lavella | Oula | River | camp | | | | PMS-261 |
| 6 | DOT 95 | SKEL | Alcedo | atthis | Solomon | Isla | New Georgia | Vella Lavella | Oula | River | camp | | | | PMS-261 |
| 7 | DOT 98 | SKEL | Cyca | reipidus | Solomon | Isla | New Georgia | Vella Lavella | Oula | River | camp | | | | PMS-261 |
| 8 | DOT 102 | FLU1 | Cyca | reipidus | Solomon | Isla | New Georgia | Vella Lavella | Oula | River | camp | | | | PMS-261 |
| 9 | DOT 104 | | Todiramphus | Solomon | Isla | New Georgia | Vella Lavella | Oula | River | camp | | | | | PMS-261 |
| 10 | DOT 109 | | Alcedo | atthis | Solomon | Isla | New Georgia | Vella Lavella | Oula | River | camp | | | | PMS-261 |
| 11 | DOT 116 | | Cyca | pus | Isla | New Georgia | Vella Lavella | Oula | River | camp | | | | | PMS-261 |

In pairs,

- 1 Write a program that will count female specimens in the CSV file.

CSV Files

AMNH-Ornithology-Internet-Export.csv

100%

Search in Sheet

Edit

Font

Alignment

Number

General

Format

Cells

Themes

Paste

Calibri (Body) 12

Wrap Text

Conditional Formatting

Styles

Actions

Themes

Aa

| L1 | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O |
|----|---------|-----|----------------|-------------|---------|---------|---------|---------|---------|---------|---------|-----------|------------|-------|------|
| | CATALOG | NO. | IDENTIFICATION | COUNTRY | STATE | COUNTY | PMS36 | LOC | OCEAN | | | COLLECTOR | COLLECTING | DATE | COL |
| 1 | DOT | 84 | FLU | Cey | tepidus | Solomon | Isla | New | Georgia | Vella | Lavella | Oula | River | camp | |
| 2 | DOT | 85 | SKEL | Cey | tepidus | Solomon | Isla | New | Georgia | Vella | Lavella | Oula | River | camp | |
| 3 | DOT | 92 | SKEL | Cey | tepidus | Solomon | Isla | New | Georgia | Vella | Lavella | Oula | River | camp | |
| 4 | DOT | 93 | SKEL | Alcedo | atthis | Solomon | Isla | New | Georgia | Vella | Lavella | Oula | River | camp | |
| 5 | DOT | 95 | SKEL | Alcedo | atthis | Solomon | Isla | New | Georgia | Vella | Lavella | Oula | River | camp | |
| 6 | DOT | 98 | SKEL | Cey | tepidus | Solomon | Isla | New | Georgia | Vella | Lavella | Oula | River | camp | |
| 7 | DOT | 102 | FLU | Cey | tepidus | Solomon | Isla | New | Georgia | Vella | Lavella | Oula | River | camp | |
| 8 | DOT | 104 | | Todiramphus | Solomon | Isla | New | Georgia | Vella | Lavella | Oula | River | camp | | |
| 9 | DOT | 109 | | Alcedo | atthis | Solomon | Isla | New | Georgia | Vella | Lavella | Oula | River | camp | |
| 10 | DOT | 116 | | Cey | pus | Isla | Solomon | Isla | New | Georgia | Vella | Lavella | Oula | River | camp |

In pairs,

- 1 Write a program that will count female specimens in the CSV file.
- 2 Write a program that returns the fraction missing a collection date.

CSV Files

AMNH-Ornithology-Internet-Export.csv

100%

Search in Sheet

Home
 Layout
 Tables
 Charts
 SmartArt
 Formulas
 Data
 Review

Edit
 Font
 Alignment
 Number
 General
 Format
 Cells
 Themes

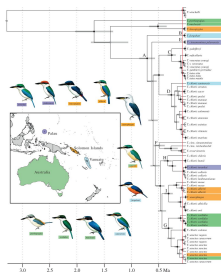
Paste
 B
 I
 U
 Text color
 Background color
 Merge
 Conditional Formatting
 Styles
 Actions
 Themes
 Aa

| L1 | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | |
|----|-------------|----------------|-------------|----------|---------|-----------|-------------|---------------|------|--------|----------|------------|--------------|---------------|------------|----------|
| | CATALOG NO. | IDENTIFICATION | COUNTRY | STATE | COUNTY | PMS36 LOC | OCEAN | ISLAND | GRD | ISLAND | LATITUDE | LONGITUDE | COLLECTOR(S) | COLLECTING | COLLECTING | |
| 1 | DOT 84 | FLU | Ceyx | epidius | Solomon | Isla | New Georgia | Vella Lavella | Oula | River | camp | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 7-May-04 | PKS-2672 |
| 2 | DOT 85 | SKE | Ceyx | epidius | Solomon | Isla | New Georgia | Vella Lavella | Oula | River | camp | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 6-May-04 | PKS-2683 |
| 3 | DOT 92 | SKE | Ceyx | epidius | Solomon | Isla | New Georgia | Vella Lavella | Oula | River | camp | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 7-May-04 | PKS-2673 |
| 4 | DOT 93 | SKE | Alcedo | atthis | Solomon | Isla | New Georgia | Vella Lavella | Oula | River | camp | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 6-May-04 | PKS-2656 |
| 5 | DOT 95 | SKE | Alcedo | atthis | Solomon | Isla | New Georgia | Vella Lavella | Oula | River | camp | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 5-May-04 | PKS-2633 |
| 6 | DOT 98 | SKE | Ceyx | epidius | Solomon | Isla | New Georgia | Vella Lavella | Oula | River | camp | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 8-May-04 | PKS-2692 |
| 7 | DOT 102 | FLU | Ceyx | epidius | Solomon | Isla | New Georgia | Vella Lavella | Oula | River | camp | 07 47 30 S | 156 37 30 E | Paul R. Sweet | 5-May-04 | PKS-2644 |
| 8 | DOT 104 | | Todiramphus | | Solomon | Isla | New Georgia | Vella Lavella | Oula | River | camp | 07 47 30 S | 156 37 30 E | | 4-May-04 | PKS-2630 |
| 9 | DOT 109 | | Alcedo | atthis | Solomon | Isla | New Georgia | Vella Lavella | Oula | River | camp | 07 47 30 S | 156 37 30 E | | 6-May-04 | PKS-2657 |
| 10 | DOT 116 | | Ceyx | pusillus | Solomon | Isla | New Georgia | Vella Lavella | Oula | River | camp | 07 47 30 S | 156 37 30 E | | 7-May-04 | PKS-2667 |

In pairs,

- 1 Write a program that will count female specimens in the CSV file.
- 2 Write a program that returns the fraction missing a collection date.
- 3 Return a list of the species ("IDENTIFICATION" row) in this file.

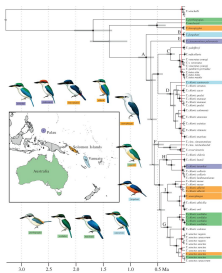
Recap



Anderson *et al.* 2014

- Install anaconda for lab today.

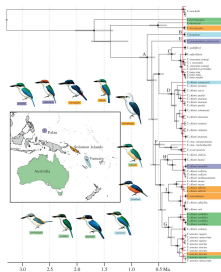
Recap



Anderson et al. 2014

- Install anaconda for lab today.
- Email lab reports to kstjohn@amnh.org

Recap



Anderson *et al.* 2014

- Install anaconda for lab today.
- Email lab reports to kstjohn@amnh.org
- Challenges available at rosalind.info