

# 実践・自然史標本の電子化 ～データ入力から発信まで～

国立科学博物館

動物研究部

神保 宇嗣

# 自己紹介

- 神保宇嗣 (じんぼ うつぎ)
- 動物研究部 陸生無脊椎動物グループ
- 専門
  - 蝶・蛾類の分類学
  - 生物多様性情報学



# 「電子化」に対する色々な疑問

- どういうソフトを使えばいいの？
- どういう項目を用意すればいいの？
- どういう所に気をつければいいの？
- どういう使い道があるの？

こういった疑問に対する私の考えを  
お話しします。

# なぜ電子化するのか？

- 自分自身が使う — 標本管理の効率化
  - 検索が簡単
  - コピーが簡単
- みんなで共有する — 様々な活動での利用
  - 日本中・世界中で共有できる
  - 様々な解析をおこなえる

これらが可能であることが重要

# 電子化の実際：S-Netの場合

## ■ 標本情報

学名	<input type="text"/>
和名 (全角カナ)	ハマキ
界名	<input type="text"/>
門名	<input type="text"/>
綱名	<input type="text"/>
目名	<input type="text"/>
科名	<input type="text"/>
属名	<input type="text"/>
種小名	<input type="text"/>
タイプ標本	<input type="text"/>

## ■ 採集に関する情報

採集日付	<input type="text"/> 年 <input type="text"/> 月 <input type="text"/> 日 ~ <input type="text"/>
採集地：国名	<input type="text"/>
採集地：都道府県名	<input type="text"/>
採集地：市区町村名	<input type="text"/>
採集地詳細	皇居
所蔵博物館	<input type="text"/>





自然史標本情報検索

[Topへ](#) | [協力博物館一覧](#)

[分類検索](#)

[詳細情報](#)

[Web検索](#) | [データリンク](#)

学名	Cerace xanthocosma (Diakonoff, 1950)	
一般名 (和名)	ピロードハマキ	
界名	Animalia	動物界
門名		
綱名		
目名	Lepidoptera	鱗翅目
科名	Tortricidae	ハマキガ科
属名	Cerace	
種小名	xanthocosma	
種の命名者	Diakonoff	
亜種以下のランク		
亜種以下のタクソン		
採集地	日本 東京 千代田区 皇居	
採集日	1999年04月18日	
採集者番号		
標本の性別	male	

# 電子化の流れ：

## 設計

- ・ソフトウェアの選択
- ・データ項目の決定

## 作成

- ・データの入力
- ・クリーニング・整形

## 使用

- ・実務での使用
- ・S-Netなどからの公開

# 電子化の流れ

## 設計

- ・ソフトウェアの選択
- ・データ項目の決定

## 作成

- ・データの入力
- ・クリーニング・整形

## 使用

- ・実務での使用
- ・S-Netなどからの公開

# どういうソフトウェアを使うか？

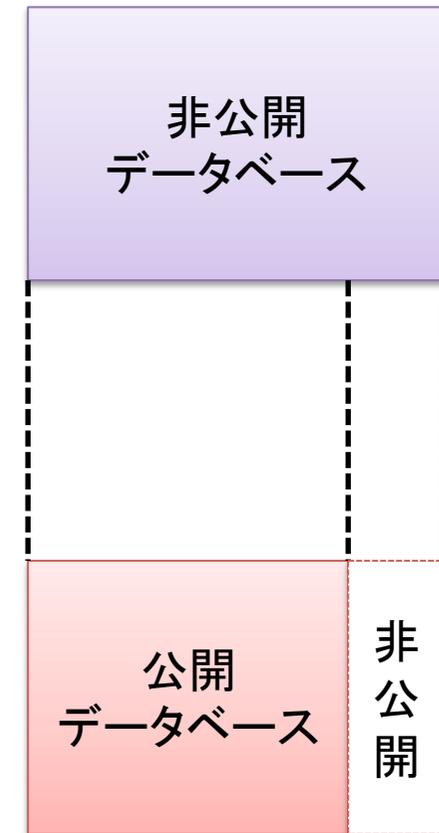
- ワープロソフト…×
  - 項目に分けたデータを扱えない
- データベースソフト
  - データの管理専門のソフト
  - 堅牢。データが壊れることが少ない
  - ある程度の知識が必要
- 表計算ソフト
  - 表形式のものならなんでも扱える
  - 自由度が高いが、その分堅牢では無い
  - 使える人が多い

# どういうソフトウェアを使うか？

- データ入力とクリーニング = 表計算
  - データ入力はアルバイトが行うことも多い
  - 効率よく入力・修正できる機能が充実
- 実務での利用 = データベース・表計算
  - データベースは詳細な検索も得意
  - データベースは必須では無い
- 公開 = データベース (+インターネット)
  - S-Netを「自分のデータベース」として使う
  - 独自に発信するデータベースを持つのも手

# どういう項目を用意するか？

- 自分のデータベース
  - 自分が使うためのデータベース
  - 公開する必要の無い情報
    - 標本箱の番号
  - 公開すると問題のある情報
    - 絶滅危惧種の採集地点の詳細情報
- 公開のデータベース
  - サイエンスミュージアムネットなど
  - 公開可能な情報だけを含む
    - 公開しても問題が無い
    - 公開のデータベースに項目が存在



# データ設計・作成の考え方

できるだけ効率よく正確に電子化したい



機械的な作業は機械にやらせる

# 項目の例

1. 標本の管理番号
2. 種名・分類体系
3. 採集地名・緯度経度
4. 採集日付・採集者
5. その他

# 項目決定のガイド私案

- S-Netにある項目はそのまま使う
  - S-Net項目は世界標準（ダーウィンコア）準拠
    - 多くの人が様々な検討を重ねた結果
    - S-Netを通じ世界で共有可能
    - 手間の軽減
- S-Netの項目には無い・問題がある場合
  - S-Netでは使われてないダーウィンコア項目の採用を検討
    - JBIFポータルのダーウィンコア解説
  - きちんと項目ごとに分ける
  - 入力の際に「ゆらぎ」を出さないような設計



[トップ](#)
[GBIF公開支援](#)
[国内活動](#)
[チュートリアル](#)
[Barcode同定](#)
[有用サイト](#)
[GBIF加盟組織サイト](#)

## Darwin Core について

Darwin Core は、標本、観察データの標準交換形式であり、インターネットを通じてデータ共用するために使用され、GBIFプロジェクトでも採用されています。以下のようにいくつかのバージョンがあり、詳細は XML Schema で定義されています。

- [Darwin Core 最新版 \(XML Schema\)](#)
- [Darwin Core 1.4 \(XML Schema\)](#)
- [Darwin Core 1.2 \(XML Schema\)](#)

## Darwin Core最新版の項目

要素名	説明	データ例
<a href="#">occurrenceID</a>	標本・観察データレコードに割り振られるユニークID	"urn:lsid:nhm.ku.edu:Herps:32", "urn:catalog:FMNH:Mammal:145732"
<a href="#">type</a>	リソースの性質	"StillImage", "MovingImage", "Sound"
<a href="#">modified</a>	データの最終更新日	"1963-03-08T14:07-0600", "2009-02-20T08:40Z", "1906-06", "1971", "2007-03-01T13:00:00Z/2008-05-11T15:30:00Z"
<a href="#">language</a>	リソースの言語	"en", "ja"
<a href="#">rights</a>	リソースのライセンス	"http://creativecommons.org/licenses/by-sa/3.0/"
<a href="#">rightsHolder</a>	リソースのライセンスを保有している個人または機関	"The Regents of the University of California."
<a href="#">accessRights</a>	データアクセスに関するライセンス情報	"not-for-profit use only"
<a href="#">bibliographicCitation</a>	リソースに関する文献情報	"Ctenomys sociabilis (MVZ 165861)" for a specimen
<a href="#">references</a>	参照している関連リソース	"http://mvzarctos.berkeley.edu/guid/MVZ:Mamm:165861"
<a href="#">institutionID</a>	内部で使用している研究機関ID	
<a href="#">collectionID</a>	コレクションを識別するID。標本データの場合 Biodiversity Collections Index のID	"urn:lsid:biocol.org:col:34818"

# 標本目録の電子化

アトキハマキ *Archips audax* Razowski

吹上御苑, 31. VIII. 2000, 1♂; 3. VI. 2004, 2♂3♀;  
25. V. 2005, 1♂. em. 12. IV. 2001, 1♂, em. 11. V.  
2002, 1♀, ex *Camellia sinensis* チャノキ.

マツアトキハマキ *Archips oporana* (Linnaeus)

吹上御苑, 3. VI. 2004, 1♂. 上道灌濠, 31. VIII. 2000,  
1♂.

(大和田ほか, 2006)

# 標本目録の電子化

アトキハマキ *Archips audax* Razowski

吹上御苑 31. VIII. 2000 1♂

3. VI. 2004 2♂3♀

25. V. 2005 1♂

em. 12. IV. 2001 1♂ ex *Camellia sinensis* チャノキ

em. 11. V. 2002 1♀ ex *Camellia sinensis* チャノキ

マツアトキハマキ *Archips oporana* (Linnaeus)

吹上御苑 3. VI. 2004 1♂

上道灌濠 31. VIII. 2000 1♂

# 標本目録の電子化

和名

学名

著者名

アトキハマキ *Archips audax* Razowski

吹上御苑 31. VIII. 2000 1♂

3. VI. 2004 2♂?

25. 5 1♂

地名

採集日

個体数

em. 12. IV. 2001 1♂ ex *Camellia sinensis* チャノキ

em. 11. V. 2002 1♀ ex *Camellia sinensis* チャノキ

マツア *Archips oporata* (Thaëus)

羽化日

寄主

吹上御苑 3. VI. 2004 1♂

上道灌濠 31. VIII. 2000 1♂

# 標本目録の電子化

- 和名→一般名（和名）
- 学名→属名・種小名・亜種以下のランク・タクソン
- 学名の著者→学名の命名者
- 地名→採集場所（国・都道府県・市区町村・詳細）
- 採集日→採集年月日（始め・終わり）
- 個体数→性別 （標本ごとに分ける）
- 羽化日→無し（公開時は採集日に入れることで対応）
- 寄主→無し
  
- 標本番号（機関略号・標本コード・標本番号）を追加

# 標本目録の電子化

	A	B	C	D	E	F	G	H
1	機関略号	標本コード	標本番号	一般名(和名)	属名	種小名	亜種以下のタクソン	学名の命名者
2	NSMT	I-L	33333	アトキハマキ	Archips	audax		Razowski

	I	J	K	L	M	N
1	国	国(日本語)	都道府県	都道府県(日本語)	市区町村	市区町村(日本語)
2	Japan	日本	Tokyo	東京都	Chiyoda-ku	千代田区

	O	P	Q	R	S
1	採集場所詳細	採集場所詳細(日本語)	採集年月日(始め)	採集年月日(終わり)	性別
2	Fukiage-gyoen	吹上御苑		20000831	male

# 電子化の流れ

## 設計

- ・ソフトウェアの選択
- ・データ項目の決定

## 作成

- ・データの入力
- ・クリーニング・整形

## 使用

- ・実務での使用
- ・S-Netなどからの公開

# データの入力

- 多くの場合アルバイトや非常勤が対応
  - 労力を少なくする
  - 不正なデータの入力を出来るだけ防ぐ



1. 入力させない
2. 入力時の不正データチェック

# 入力させない

- 必要な項目だけ入力させる
  - 機械的に入力できる項目はあとから追加
  - 同じ値の場合
    - 機関番号・標本コード
  - 対訳表で対応付けできる項目
    - 学名関連（和名から決定）
- 対訳表が不適切な場合
  - 名前が対訳表とあわない
    - 「標本の学名」 「最新の学名」 両方が必要なことも

# 入力時の不正データチェック

- 「データの入力規制」機能の活用

データの入力規制

設定 入力時メッセージ エラーメッセージ 日本語入力

条件の設定

入力値の種類:  空白を無視する

日付

データ: 次の値の間

次の日付から: 2000/1/1

次の日付まで: 2005/3/31

同じ入力規制が設定されたすべてのセルに変更を適用する

すべてクリア キャンセル OK

範囲指定した日付チェック

データの入力規制

設定 入力時メッセージ エラーメッセージ 日本語入力

日本語入力:

コントロールなし

オン

オフ (英語モード)

すべてクリア キャンセル OK

日本語入力の自動コントロール

# データのクリーニング・整形

- 入力データをチェックし、最終的なデータベースに入れられるようにする過程
  1. 機械的に入力できる項目を追加
  2. 入力文字の問題を解消
  3. 対訳表による入力
  4. 入力の「ゆらぎ」を解消
  5. 項目の結合・分離（事前整形ツールなど）
  6. 追加情報（レッドリスト掲載種など）
  7. 最終チェック（変換ツールなど）

# データのクリーニング・整形

- 入力データをチェックし、最終的なデータベースに入れられるようにする過程
  1. 機械的に入力できる項目を追加
  2. 入力文字の問題を解消
  3. 対訳表による入力
  4. 入力の「ゆらぎ」を解消
  5. 項目の結合・分離（事前整形ツールなど）
  6. 追加情報（レッドリスト掲載種など）
  7. 最終チェック（変換ツールなど）

# クリーニング・整形に使う ソフトウェア

- 表計算ソフト（Excelなど）
- テキストエディタ（サクラエディタ, miなど）
  - 正規表現使用
- 汎用ツール（Open Refineなど）
- 専門ツール（サイエンスミュージアムネット）
  - 事前整形ツール・レッドリストチェッカー・地名辞書など
- アプリケーションの自動機能
  - Excelマクロ, FileMakerスクリプトなど
- 汎用スクリプト言語（Ruby, Perl, Python）

# 入力文字の問題を解消

## 機械的な間違いのチェック

- Excelの関数を使う
  - アルファベットなどの全角→半角
    - ASC関数
  - 項目の前後に付いている空白
    - TRIM関数
  - 連続する空白を一つにする
    - 「空白2個」を「空白1個」に置換
- 「事前整形支援ツール」等を利用

# 対訳表による入力

	A
1	和名
2	ビロードハマキ
3	ビロードハマキ
4	アトキハマキ
5	アトキハマキ
6	アトキハマキ
7	アトキハマキ
8	ビロードハマキ

+

A	B	C
和名	属名	種小名
ビロードハマキ	Cerace	xanthocosma
アトキハマキ	Archips	audax



	A	B	C
1	和名	属名	種小名
2	ビロードハマキ	Cerace	xanthocosma
3	ビロードハマキ	Cerace	xanthocosma
4	アトキハマキ	Archips	audax
5	アトキハマキ	Archips	audax
6	アトキハマキ	Archips	audax
7	アトキハマキ	Archips	audax
8	ビロードハマキ	Cerace	xanthocosma

# 対訳表による入力：VLOOKUP関数

シート「Sheet1」

シート「Species」

	A
1	和名
2	ピロードハマキ

+

A	B	C
和名	属名	種小名
ピロードハマキ	Cerace	xanthocosma
アトキハマキ	Archips	audax

シート「Species」のA列からC列を対訳表にする  
シート「Sheet1」のA2の値を、対訳表の一番左の列で検索し、  
見つかったら、左から2番目の列（=属名）の値を返す

```
=VLOOKUP(A2,Species!A:C,2,FALSE)
```

	A	B	C
1	和名	属名	種小名
2	ピロードハマキ	Cerace	xanthocosma

# 入力のゆらぎの解消

入力を統一する

データ(修正前)
東京
東京都
神奈川県
神奈川
東京都
東京
神奈川県
東京
東京都
東京都
神奈川
神奈川



データ(修正後)
東京都
東京都
神奈川県
神奈川県
東京都
東京都
神奈川県
東京都
東京都
東京都
神奈川県
神奈川県

# 入力のゆらぎの解消

- データを別シートにコピー
- 重複の削除
- 並び替え

データ(修正前)
東京
東京都
神奈川県
神奈川
東京都
東京
神奈川県
東京
東京都
東京都
神奈川
神奈川



データ(修正前)
神奈川
神奈川県
東京
東京都

データ(修正前)	データ(修正後)
神奈川	神奈川県
神奈川県	神奈川県
東京	東京都
東京都	東京都

データ(修正前)	データ(修正後)
東京	東京都
東京都	東京都
神奈川県	神奈川県
神奈川	神奈川県
東京都	東京都
東京	東京都
神奈川県	神奈川県
東京	東京都
東京都	東京都
東京都	東京都
神奈川	神奈川県
神奈川	神奈川県

- ゆらぎの対訳表を作成

- 対訳表を反映

# 応用：和名入力エラーの解消

和名(修正前)
ビロウドハマキ
ビロードハマキ
アトキハマキ
アトキハマキ
アトハマキ
アトキハマキ
ビロードハマキ



和名(修正前)	和名チェック	和名(修正後)	和名(修正後)チェック
ビロードハマキ	ビロードハマキ	ビロードハマキ	ビロードハマキ
ビロウドハマキ	#N/A	ビロードハマキ	ビロードハマキ
アトハマキ	#N/A	アトキハマキ	アトキハマキ
アトキハマキ	アトキハマキ	アトキハマキ	アトキハマキ

- 別シートにコピー
- 重複削除、並び替え
- 対訳表存在チェック
- エラー修正表

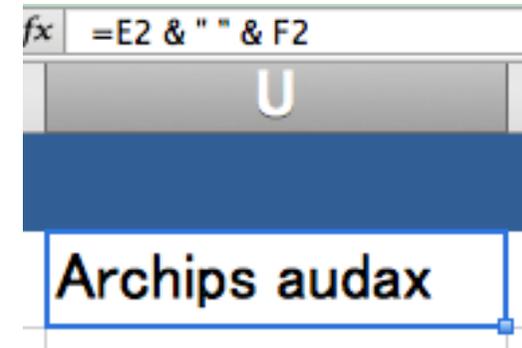


- 和名が無い時
- 修正表で和名修正
  - 対訳表で学名追加

和名(修正前)	和名(修正後)	属名	種小名
ビロウドハマキ	ビロードハマキ	Cerace	xanthocosma
ビロードハマキ	ビロードハマキ	Cerace	xanthocosma
アトキハマキ	アトキハマキ	Archips	audax
アトキハマキ	アトキハマキ	Archips	audax
アトハマキ	アトキハマキ	Archips	audax
アトキハマキ	アトキハマキ	Archips	audax
ビロードハマキ	ビロードハマキ	Cerace	xanthocosma

# 余談：まとめる・わかる

- 学名を一つのセルにまとめたい
  - 文字列の足し算記号 (&) を使う
  - IF関数を使う



例：亜種が無い時とある時との場合分け  
属=E2, 種=F2, 亜種=G2 に入っているとき

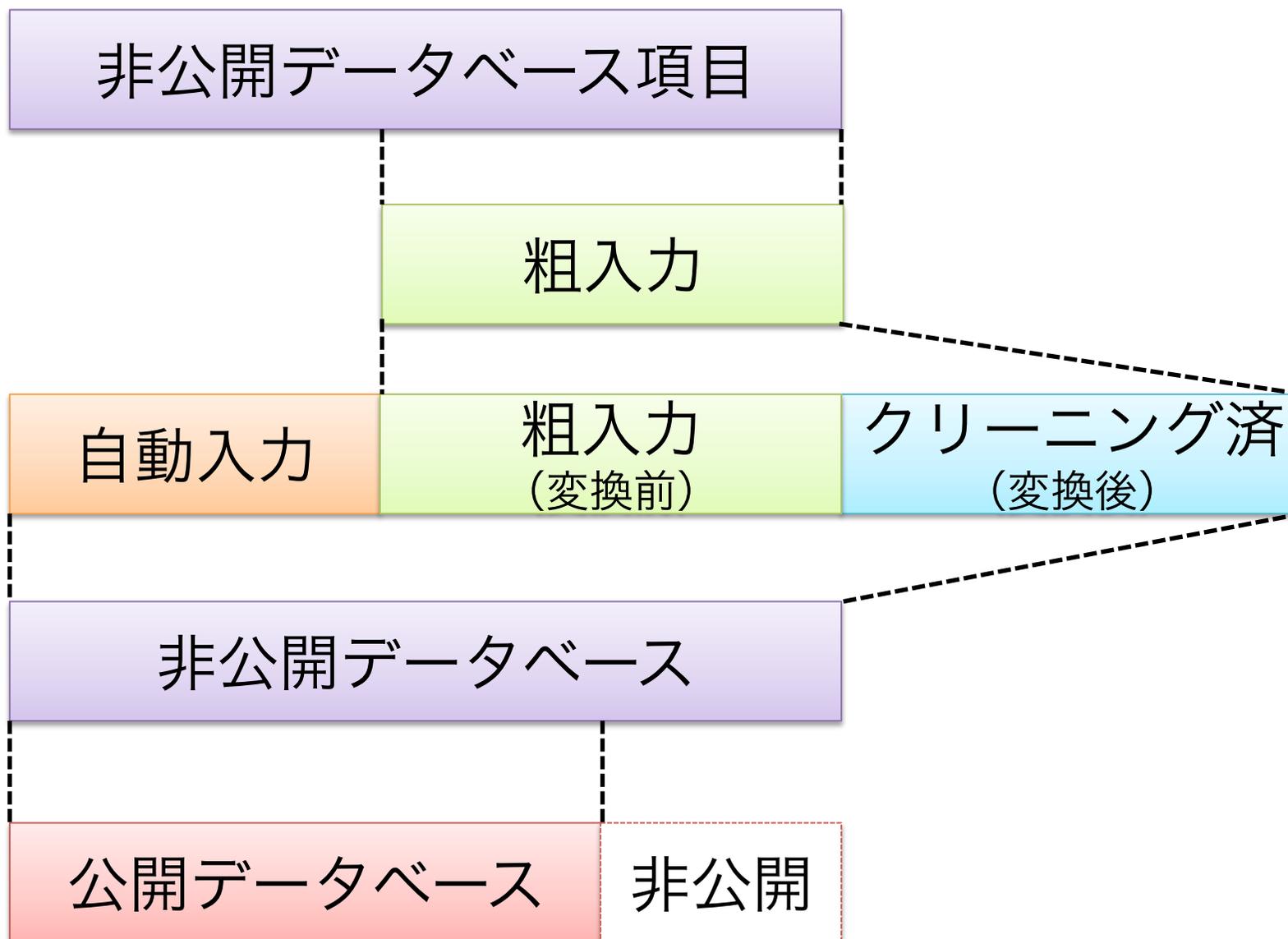
亜種が無い時 `E2 & "" & F2`

亜種がある時 `E2 & "" & F2 & "" & G2`

➡ `E2 & F2 & IF(G2 = "", "", "" & G2)`

- 学名を要素ごとに分割したいとき
  - 「事前整形ツール」 「変換ツール」 を使いましょ  
う

# まとめ：入力から最終出力まで



# 電子化の流れ

## 設計

- ・ソフトウェアの選択
- ・データ項目の決定

## 作成

- ・データの入力
- ・クリーニング・整形

## 使用

- ・実務での使用
- ・S-Netなどからの公開

# 実務での利用

- 所蔵品の検索
- 研究・業務での利用
- 報告書の作成
  - エクセルからワードへの変換
- 独自のデータベース構築

アトキハマキ *Archips audax* Razowski  
吹上御苑, 31. VIII. 2000, 1♂; 3. VI. 2004, 2♂3♀;  
25. V. 2005, 1♂. em. 12. IV. 2001, 1♂, em. 11. V.  
2002, 1♀, ex *Camellia sinensis* チャノキ.

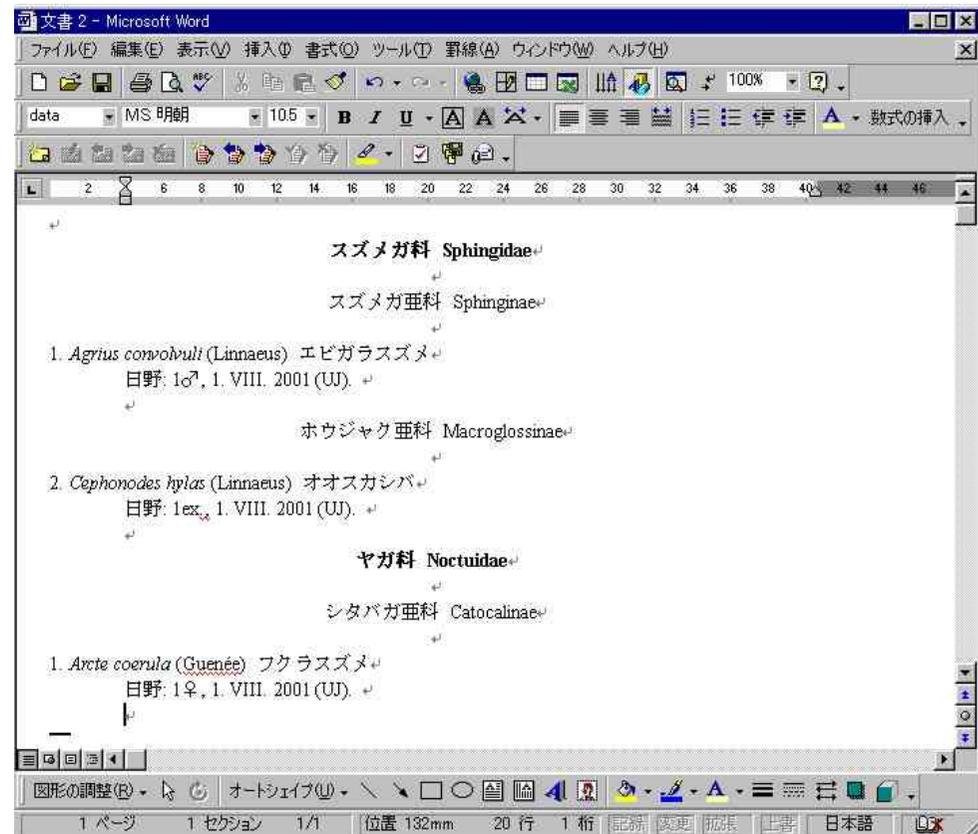
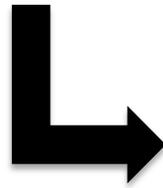
# MakeList

## ファウナ調査リスト作成支援マクロ

- エクセルで作成 (Excel VBA)
- 学名付加・並び替え・ワードで出力
- 関係者へ配布  
→ 想像以上の反響



	地名	雄個体数	雌個体数	個体数	年	月	日	採集者
3	日野	1♂			2001	8	1	UU
4	日野		1♀		2001	8	1	UU
5	日野			1ex	2001	8	1	UU



スズメガ科 Spingidae

スズメガ亜科 Spinginae

1. *Agrius convolvuli* (Linnaeus) エビガラスズメ  
日野: 1♂, 1. VIII. 2001 (UU).

ホウジャク亜科 Macroglossinae

2. *Cephonodes hylas* (Linnaeus) オオスカシバ  
日野: 1ex, 1. VIII. 2001 (UU).

ヤガ科 Noctuidae

シタバガ亜科 Catocalinae

1. *Arcte coerulea* (Guenée) フクラスズメ  
日野: 1♀, 1. VIII. 2001 (UU).

# 種名データベースシステム Ruga-name

- エクセルシート 2 枚で公開できる (データベース情報・種名)
- シート差し替えで即時更新
- 無料で使える (オープンソース)
  - α版を公開中 (GitHub)
- 実際の例
  - 日本産蝶類和名学名便覧



## BINRAN

The Current Checklist of Japanese Butterflies  
日本産蝶類和名学名便覧

分類体系

[戻る](#)

- Family Hesperidae セセリチョウ科 [属一覧](#) [種一覧](#)
  - Subfamily Coeliadinae アオバセセリ亜科 [属一覧](#) [種一覧](#)
  - Subfamily Pyrginae チャマダラセセリ亜科 [属一覧](#) [種一覧](#)
  - Subfamily Heteropterae チョウセンキボシセセリ亜科 [属一覧](#) [種一覧](#)
  - Subfamily Hesperinae セセリチョウ亜科 [属一覧](#) [種一覧](#)
- Family Papilionidae アゲハチョウ科 [属一覧](#) [種一覧](#)
  - Subfamily Parnassiinae ウスバアゲハ亜科 [属一覧](#) [種一覧](#)
    - Tribe Zerynthiini ギフチョウ族 [属一覧](#) [種一覧](#)
    - Tribe Parnassiini ウスバアゲハ族 [属一覧](#) [種一覧](#)
  - Subfamily Papilioninae アゲハチョウ亜科 [属一覧](#) [種一覧](#)
    - Tribe Troidini キンタアゲハ族 [属一覧](#) [種一覧](#)
    - Tribe Papilionini アゲハチョウ族 [属一覧](#) [種一覧](#)
    - Tribe Graphiini アオスジアゲハ族 [属一覧](#) [種一覧](#)

学名	<i>Burara aquilina aquilina</i>
科 (学名)	Hesperidae
科 (和名)	セセリチョウ科
亜科 (学名)	Coeliadinae
亜科 (和名)	アオバセセリ亜科
族 (学名)	
族 (和名)	
属 (学名)	Burara
属 (和名)	キバネセセリ属
亜属 (学名)	
種名	aquilina
種内分類群名	aquilina
著者	(Speyer, 1879)
和名	キバネセセリ
別名	
亜種範囲	

<http://binran.lepimages.jp/>

学名	<i>Burara aquilina aquilina</i>
科 (学名)	Hesperidae
科 (和名)	セセリチョウ科
亜科 (学名)	Coeliadinae
亜科 (和名)	アオバセセリ亜科
族 (学名)	
族 (和名)	
属 (学名)	Burara
属 (和名)	キバネセセリ属
亜属 (学名)	
種名	aquilina
種内分類群名	aquilina
著者	(Speyer, 1879)
和名	キバネセセリ
別名	
亜種範囲	

# 「オープンデータ」という考え方

- 誰でも自由に使える形でデータを公開
  - 生データをそのまま公開する
  - オープンライセンスを適用する
    - 日本産蝶類和名学名便覧の場合  
クリエイティブ・コモンズ・ライセンス (CC-BY)
    - Ruga-nameの場合  
MITライセンスを適用予定
- 今後の政策の動きにも注意
  - 「オープンガバメント」の急激な動き

# まとめ

- どういうソフトを使えばいいの？
  - 管理データベースとしてのエクセル中心に紹介
- どういう項目を用意すればいいの？
  - 項目の決め方のガイドライン
- どういう所に気をつければいいの？
  - データ入力とクリーニングの2つの段階
- どういう使い道があるの？
  - 報告書作成や独自データベース