An evaluation of taxonomic name finding & next steps in Biodiversity Heritage Library (BHL) developments

Chris Freeland

Technical Director, BHL

Director of Bioinformatics,
Missouri Botanical Garden

Goals of BHL

- Scan public domain biodiversity literature.
- Negotiate rights to copyrighted materials.
- Ingest content digitized by others.
- Provide interfaces & APIs for repository.
  - GUIs
  - Services for data mining & citation resolution

http://www.biodiversitylibrary.org

BHL Institutions

Botanical Gardens

- Missouri Botanical Garden
- New York Botanical Garden
- Royal Botanic Garden, Kew

University Libraries

- Botany Libraries, Harvard University
- Ernst Meyer Library of the Museum of Comparative Zoology, Harvard University
- University of Illinois

Museums

- American Museum of Natural History (New York)
- Natural History Museum (London)
- Smithsonian Institution (Washington)
- The Field Museum (Chicago)

Bioinformatics Institutes

- MBL/WHOI
- uBio.org

Now Online

- More than:
  - 22,000 volumes
  - 9.2 million pages

Only 290 million to go!

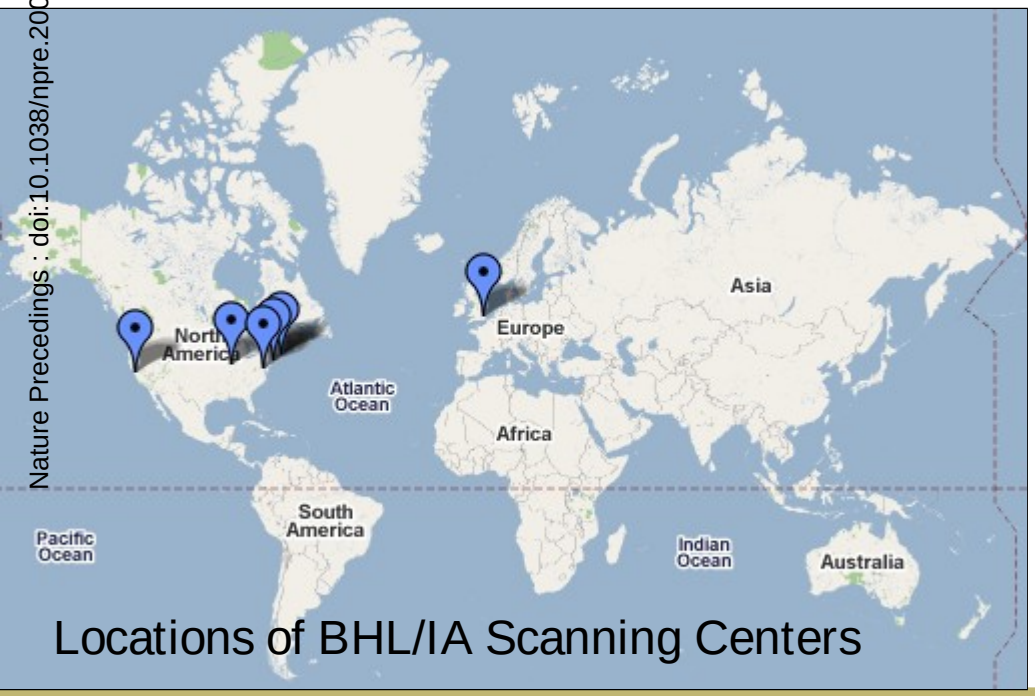
- Avg. monthly growth rate
  - 1,500 volumes
  - 600,000 pages

See you in 2048!

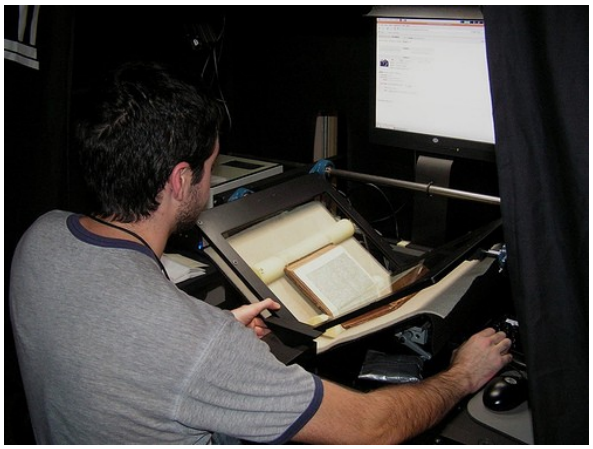
Scanning Operations

BHL uses scanning centers established by Internet Archive for mass scanning.

Some partner libraries also scan in-house.



Locations of BHL/IA Scanning Centers



Want to expand international footprint:

- mirrored content
- ingest from global data providers

Complexities of distributed, mass scanning

Flora of Colorado,
by Per Axel Rydberg

from NYBG

☆☆☆☆☆
(not yet rated)

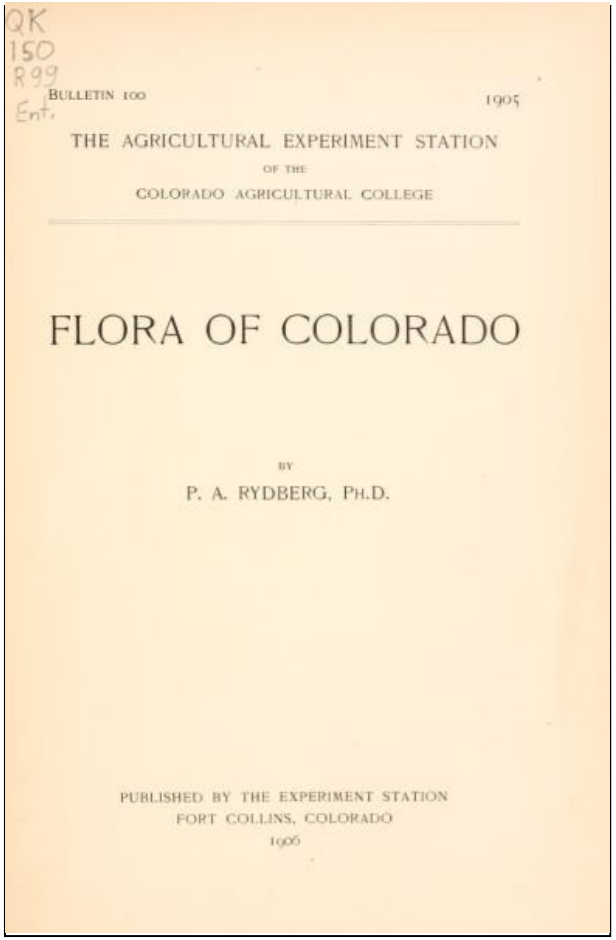
Type:	Book; English
Publisher:	Fort Collins, Experiment Station, 1906.
Editions:	3 Editions
OCLC:	13927851
Related Subjects:	Plants -- Colorado.

Flora of Colorado
by Per Axel Rydberg

from Smithsonian

☆☆☆☆☆
(not yet rated)

Type:	Book : State or province government publication; English
Publisher:	Fort Collins, Colo. : Agricultural Experiment Station of the Colorado Agricultural College, 1906.
Editions:	3 Editions
OCLC:	1577518
Related Subjects:	Botany -- Colorado.



Open Access Data

The snakes of Australia; an illustrated and descriptive catalogue of all the known species. By Gerard Krefft...

Publisher: Sydney,T. Richards, Government Printer,1869.

PDF

OCR

JP2

XML

Name	Last Modified	Size	Type
Parent Directory/		-	Directory
snakesofaustrali00kref.djvu	2008-Mar-25 07:21:43	3.8M	image/vnd.djvu
snakesofaustrali00kref.gif	2008-Mar-25 03:28:19	287.3K	image/gif
snakesofaustrali00kref.pdf	2008-Mar-25 07:41:56	7.9M	application/pdf
snakesofaustrali00kref_abbyy.gz	2008-Mar-25 06:55:05	3.5M	application/octet-stream
snakesofaustrali00kref_bw.pdf	2008-Mar-25 08:13:20	6.4M	application/pdf
snakesofaustrali00kref_dc.xml	2008-Mar-10 15:58:11	0.4K	application/xml
snakesofaustrali00kref_djvu.txt	2008-Mar-25 08:13:37	238.4K	text/plain
snakesofaustrali00kref_djvu.xml	2008-Mar-25 06:56:50	1.9M	application/xml
snakesofaustrali00kref_files.xml	2008-Mar-25 08:13:43	4.1K	application/xml
snakesofaustrali00kref_flippy.zip	2008-Mar-25 03:52:06	4.8M	application/zip
snakesofaustrali00kref_jp2.zip	2008-Mar-25 03:27:28	43.7M	application/zip
snakesofaustrali00kref_marc.xml	2008-Mar-10 15:58:11	1.8K	application/xml
snakesofaustrali00kref_meta.mrc	2008-Mar-10 15:58:11	0.5K	application/octet-stream
snakesofaustrali00kref_meta.xml	2008-Jun-06 21:14:15	1.5K	application/xml
snakesofaustrali00kref_metasource.xml	2008-Mar-10 15:58:11	0.4K	application/xml
snakesofaustrali00kref_orig_jp2.tar	2008-Mar-25 01:14:46	87.3M	application/x-tar
snakesofaustrali00kref_scandata.xml	2008-Mar-25 01:14:40	90.2K	application/xml

Name Finding via TaxonFinder

uBio

Discussion.—Pigmy antechinuses are the smallest of the dasyurids and are among the smallest-sized mammals. They are not well-represented in collections and are poorly known as to their dental morphology, their ecology, and their taxonomy.

A comparison of the Madura Cave specimens with those of Recent species shows many similarities, such as size and the tendency to crowd the premolars. Comparison also reveals a number of differences that indicate that the Madura Cave form is different from the described Recent species. In general, the cheek teeth of the Madura Cave form are slightly larger than their counterparts in Recent *Planigale ingrami*, especially $P_{\overline{4}}$ and $M_{\overline{2}-\overline{4}}$ (figs. 12, 13; tables 1, 2). Comparison with $M_{\overline{1}-\overline{3}}$ of *Antechinus maculatus* is very close (fig. 12). However, $P_{\overline{4}}$ is much longer in the Madura Cave form than it is in either *Planigale ingrami* or *Antechinus maculatus*. The $M_{\overline{4}}$

SOAP response

Name finding via TaxonFinder

Submit to NameBank

Extract names

```
<?xml version="1.0" encoding="UTF-8"?>
<allNames>
    <entity>
        <nameString>SOBRALIA amplexicaulis</nameString>
        <parsedName canonical="SOBRALIA amplexicaulis">
            <component type="name" rank="genus">SOBRALIA</component>
            <component type="name" rank="species">amplexicaulis</component>
        </parsedName>
        <score>1</score>
        <namebankID>8652923</namebankID>
    </entity>
    <entity>
        <nameString>SoBRALIA liliastrum</nameString>
        <parsedName canonical="SoBRALIA liliastrum">
            <component type="name" rank="genus">SoBRALIA</component>
            <component type="name" rank="species">liliastrum</component>
        </parsedName>
        <score>1</score>
        <namebankID>3493662</namebankID>
    </entity>
    <entity>
        <nameString>Epidendrum liliastrum</nameString>
        <parsedName canonical="Epidendrum liliastrum">
            <component type="name" rank="genus">Epidendrum</component>
            <component type="name" rank="species">liliastrum</component>
        </parsedName>
        <score>1</score>
        <namebankID>8764188</namebankID>
    </entity>
```

SoBRALIA liliastrum
Epidendrum liliastrum
SOBRALIA
Serapias
Cymbidium hirsutum
SOBRALIA amplexicaulis

Name Finding in action with Taxonomic Intelligence…

(hab. s. sp.)
exemplini ulteriori sub-

icatis terminalibus.

233.
libus, racemo terminali."
is et locis meridionalibus,

aspersum. Germen ca-
que evaginata. R. et P.
Presl. Reliq. Hoenk p.

inferioribus foliaceis, vaginatus. Flores 2 poll
Hab. in Peruvia.

Name Finding Stats to date*

- Have mined more than ***30 million*** name string occurrences
  - 4.3 million unique
- More than ***23.3 million*** name strings verified by NameBank
  - 1.1 million unique

*19 October 2008

3. Sobralia liliastrum.

S. foliis lanceolatis acuminatis, racemo brevi terminali flexuoso, bracteis acuminatis ovario longioribus, alis columnæ maximis falcatis.

Epidendrum liliastrum. *Salzm. Herb. Bah.*

Hab. ad *Bahiam*, in fruticetis sabulosis; *Salzmann.* (*hab. s. sp.*)

Partes omnes succulentas, exsiccatione contusas, examini ulteriori subjicere nequivi.

4. Sobralia? Caravata.

S. foliis lanceolatis pubescentibus, capitulis imbricatis terminalibus.

Caravata-Miri *Vernaculè.*
Serapias Caravata. *Aubl. Guian.* 2. 816. *t.* 320.
Cymbidium hirsutum. *Willd. Sp. Pl.* 4. 94.

Hab. in sylvis *Guianæ* supra arbores; Aublet.

Flores lutei.

Dubiæ.

? 5. Sobralia amplexicaulis. *Fl. Peruv. Syst.* 233.

S. "bulbis fasciculatis, foliis cordatis amplexicaulibus, racemo terminali."

Hab. affatim in *Peruvia*, in *Chinchao* runcationibus et locis meridionalibus, florens à Maio ad Augustum; Ruiz et Pavon.

Labellum squamis linearibus apice dilaceratis conspersum. Germen calyculo 3-dentato connatum. Folia nervosa venosaque evaginata. R. et P.

? 6. Huc etiam fortassè referenda est *Bletia ciliata* Presl. Reliq. Hænk p. 99. quæ *S. dichotoma* esset, si flores majores.

Hab. in *Peruvia.*

CIV. ACANTHOPHIPPIUM.

Bibliography for "Sobralia dichotoma" by Title

As of 16 Oct 2008 3:12AM (New Search)

15 pages found in **9** titles

View NameBank record

View Page in Book

Pichis Trail, Yapas, 1350–1600 meters, in dense forest, "epiphyte; buds greenish yellow," *Killip & Smith 25574*.

Sobralia crocea Reichb. f. Fl. des Serres 8: 247. 1853; Cogn. Martius Fl. Bras. 3, pt. 5: 341. 1901. *Cyathoglottis crocea* Poepp. & Endl. Nov. Gen. ac Sp. 1: 55. 1836.

Plant epiphytic or terrestrial, slender, glabrous. Stems suffruticose, numerous, about 3–6 dm. high when mature, loosely leaved except through the basal portion. Leaves 3–6, oblong-lanceolate, elliptic-lanceolate or elliptic, about 7–14 cm. long, up to 3 cm. wide, acute to short-acuminate, distichous, erect-spreading, subclasping, lightly chartaceous. Flowers small for the genus, terminal, 1–3 (perhaps more), in the axils of sheaths, fugacious. Ovary linear-cylindric, 4-angled. Sepals similar, linear-lanceolate or elliptic-lanceolate, acute, thinly membranaceous, saffron-yellow to red-orange, about 2.2–3 cm. long, up to 7 mm. wide. Petals similar to the sepals, but a little smaller. Lip somewhat shorter than the sepals and surrounding the column in natural position, up to 2.3 cm. long and 1.2 cm. wide, oblong to oblong-elliptic when expanded, subtruncate at the apex, undulate and crenate on the anterior margins; disc provided through the middle with about 4 narrow keels which are somewhat dilated and coarsely dentate above. Column shorter than the lip, about 1.3 cm. long, terminated by a pair of falcate, retrorse lobes. Capsule about 6 cm. long at maturity.

Cuzco: Prov. of Paucartambo, Sta. Isabel to Asunción, 1800 meters, epiphyte, *Vargas 5537* (flowers agglutinated).—Huánuco: Near Pampayaco (Pampayacu) and Cuchero (Cochero), rather rare, *Poeppig 1580* (type). Cierra Azul, "Vic. Estacion de Te," cut on Pucallpa Road, 1070 meters, flowers orange with paler lip, *Seibert 2250, 2251*. Tingo María (Divisoria), 1500 meters, terrestrial on open bluffs, flowers orange, *Carpenter 102*.—San Martín: On road to Divisoria, 59 km. from Tingo María, on highway to Pucallpa, on bank, 1600 meters, *Allard 21290*. Same data as last, 1250 meters, *Allard 21320*.

In the recent collections examined, which seem to be surely referable to this concept, the lip shows about 4 narrow keels which are coarsely dentate above, rather than the triangular-falcate lamellae surrounded by fleshy club-shaped warts, as described.

Sobralia dichotoma Ruíz & Pav. Syst. Veg. Fl. Peruv. et Chil. 1: 232. 1798; Lindl. Fol. Orch. Sobralia 2, no. 1. 1854; Cogn. Martius Fl. Bras. 3, pt. 5: 346. 1901. *Cattleya dichotoma* (as tichotoma) Beer, Prakt. Stud. Fam. Orch. 215. 1854. *S. Mandonii* Reichb. f. Xen. Orch. 2: 175, t. 175, I, fig. 1. 1873.

APIs & Data Sharing

- Name Service (Documentation)
  - REST: XML or JSON
- Data Export (Documentation)
  - Monthly export of BHL titles, volumes, pages, names in delimited files
- Citation Resolver *v0.1*
  - *available by end of 2008*

- NameCount
- NameCountBetweenDates
- NameGetDetail
- NameList
- NameListBetweenDates
- NameSearch

Name Finding Evaluation

See Poster in hall

- Structured and performed by Qin Wei
  - Ph.D. student at UIUC, working with Bryan Heidorn
- Methodology
  - Scholarly volunteers manually identified scientific names on random sample of **392 pages** in BHL corpus
  - Compared those against **OCR**,then two name finding algorithms (**TaxonFinder & FAT**)
- Goals
  - Spark discussion, set baseline for future work

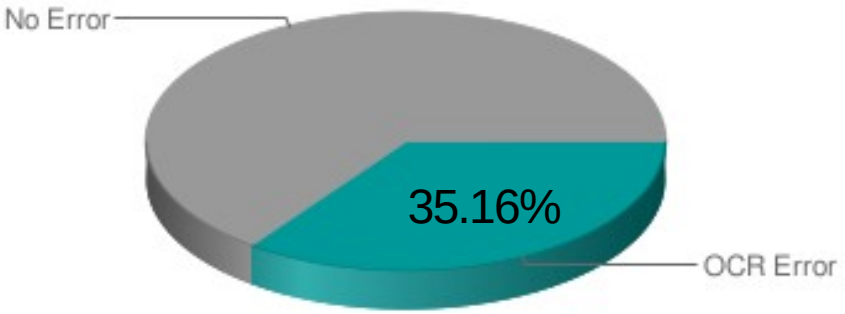
Characteristics of sample

Number of Pages	392
Average Number of Words per Page	446.8
Average Number of Names per Page	7.7
Total Number of Names	**3003**
Total Number of Unique Names	**2610**

= 86.91%

OCR error rate *for names only*

Of the 3,003 names, 1,056 were incorrectly transcribed by OCR.



Top OCR errors

1	Insert Space	8	n->v
2	Omit Space	9	l->i
3	e->c	10	r->i
4	u->l	11	u->ii
5	u->n	12	h->l
6	i->l	13	h->ii
7	c->e	14	e->o

Performances of algorithms

	TaxonFinder	FAT
Precision	40.32%	28.20%
Recall	36.62%	23.34%
F-score	**38.47%**	**25.77%**

Excluding names with OCR errors

Precision	43.77%	32.25%
Recall	25.82%	17.21%
F-score	**34.80%**	**24.73%**

Including names with OCR errors

Considerations

- Improving OCR software is out of scope
  - Google’ s Tesseract is only viable open source option
  - Flurry of activity in 2006-2007, quiet since
- Rekeying is expensive given size of corpus
  - Will not scale

Recommendations

- Enhance “ fuzzy” retrieval in algorithms
  - Exception rules to overcome OCR errors
- More work needed in this space
  - More evaluations & experiments
  - Robust training sets
    - *reCAPTCHA* for names?

Up next: BHL Article Repository

- YouTube Broadcast Yourself™ for biodiversity articles
- " Safe harbor" model
  - BHL provides platform
  - Community provides content
    - Scientists, students, libraries
- Implemented using Fedora

FedoraCommons

And if that wasn’ t enough…

- Additional services
  - Title Resolver, LSIDs
- Distributed architecture
  - data & applications
- Interface improvements
  - Internationalization
- Further evaluations & experiments
  - rich test bed for information retrieval

Contact

Chris Freeland

4344 Shaw Blvd.

St. Louis, MO 63110

chris.freeland@mobot.org

http://www.biodiversitylibrary.org