ZooKeys 550: 153–171 (2016)
doi: 10.3897/zookeys.550.9673
http://zookeys.pensoft.net

RESEARCH ARTICLE

A peer-reviewed open-access journal
ZooKeys
Launched to accelerate biodiversity research

Unlocking *Index Animalium*: From paper slips to bytes and bits

Suzanne C. Pilsk[1], Martin R. Kalfatovic[1], Joel M. Richard[1]

1 *Smithsonian Libraries, Washington DC, USA*

Corresponding author: *Suzanne C. Pilsk* (PilskS@si.edu)

Academic editor: *E.Michel* | Received 23 March 2015 | Accepted 25 March 2015 | Published 7 January 2016

http://zoobank.org/F8D9D471-E11A-4EE9-AEE8-BDEB26F0DAB2

Citation: Pilsk SC, Kalfatovic MR, Richard JM (2016) Unlocking *Index Animalium*: From paper slips to bytes and bits. In: Michel E (Ed.) Anchoring Biodiversity Information: From Sherborn to the 21st century and beyond. ZooKeys 550: 153–171. doi: 10.3897/zookeys.550.9673

Abstract

In 1996 Smithsonian Libraries (SIL) embarked on the digitization of its collections. By 1999, a full-scale digitization center was in place and rare volumes from the natural history collections, often of high illustrative value, were the focus for the first years of the program. The resulting beautiful books made available for online display were successful to a certain extent, but it soon became clear that the data locked within the texts needed to be converted to more usable and re-purposable form via digitization methods that went beyond simple page imaging and included text conversion elements. Library staff met with researchers from the taxonomic community to understand their path to the literature and identified tools (indexes and bibliographies) used to connect to the library holdings. The traditional library metadata describing the titles, which made them easily retrievable from the shelves of libraries, was not meeting the needs of the researcher looking for more detailed and granular data within the texts. The result was to identify proper print tools that could potential assist researchers in digital form. This paper outlines the project undertaken to convert Charles Davies Sherborn's *Index Animalium* into a tool to connect researchers to the library holdings: from a print index to a database to eventually a dataset.

Sherborn's microcitation of a species name and his bibliographies help bridge the gap between taxonomist and literature holdings of libraries. In 2004, SIL received funding from the Smithsonian's Atherton Seidell Endowment to create an online version of Sherborn's *Index Animalium*. The initial project was to digitize the page images and re-key the data into a simple data structure. As the project evolved, a more complex database was developed which enabled quality field searching to retrieve species names and to search the bibliography. Problems with inconsistent abbreviations and styling of his bibliographies made the parsing of the data difficult. Coinciding with the development of the Biodiversity Heritage Library

(BHL) in 2005, it became obvious there was a need to integrate the database converted *Index Animalium*, BHL's scanned taxonomic literature, and taxonomic intelligence (the algorithmic identification of binomial, Latinate name-strings). The challenges of working with legacy taxonomic citation, computer matching algorithms, and making connections have brought us to today's goal of making Sherborn available and linked to other datasets. Partnering with others to allow machine-to-machine communications the data is being examined for possible transformation into RDF markup and meeting the standards of Linked Open Data. SIL staff have partnered with Thomson Reuters and the Global Names Initiative to further enhance the *Index Animalium* data set. Thomson Reuters' staff is now working on integrating the species microcitation and species name in the ION: Index to Organism Names project; Richard Pyle (The Bishop Museum) is also working on further parsing of the text. The *Index Animalium* collaborative project's ultimate goal is to successful have researchers go seamlessly from the species name in either ION or the scanned pages of *Index Animalium* to the digitized original description in BHL - connecting taxonomic researchers to original authored species descriptions with just a click.

Keywords
Metadata, Digitization, Linked Open Data

Background

The Smithsonian Libraries' collections support the varied museums and research centers that support the mandate for the "increase and diffusion of knowledge" established by the benefactor James Smithson. The diversity of the subject matter in the Libraries collection reflects the range of topics, disciplines and activities undertaken by Smithsonian researchers. The Libraries has developed along the lines of the Institution to support the vast array of topics that has become the largest complex of museums and research centers in the world with 20 libraries supporting 19 museums and 9 researcher centers. The Institution's natural history collections date back to the 18th century and have been collected to assist in the study and stewardship of the extensive specimen collections. The United States National Museum was established within the Institution in 1858 and moved into a separate, individual museum in 1910. The Natural History Library collections of Smithsonian Libraries have grown in conjunction with the National Museum to help researchers identify and document specimen collections. With a substantial amount of focus since its founding on classic collections-based research of systematics and taxonomy, the collection in the library supports the discovery of species and naming. The reliance on historical literature to perform the work has made for a strong library collection. The Smithsonian Libraries has grown to take on a role of providing authoritative information and creates innovate services for the curators, scientists, and researchers. (Smithsonian Institution Libraries. *Rare Books and Special Collections in the Smithsonian Institution Libraries*. 1995) This includes the move towards providing the necessary information in digital form alongside the traditional print collections.

To build and preserve along with the supporting of present day research needs, the Libraries looks for ways to have the information within the collection reach the

needed patron whenever and wherever they may be. Acquiring digital data, electronic journals and resources and current research database subscriptions is one aspect of the Libraries reach. Scanning the holdings of collections to provide better access was another step towards the delivery of critical information to the researchers. The Libraries' first started digitization projects in 1996. By 1999, a full-scale digitization center was in place and rare volumes from the natural history collections, often of high aesthetic value, were the focus for the first years of the program. The resulting beautiful books made available for online display were successful to a certain extent; but it soon became clear that the data locked within the texts needed to be converted to more usable and re-purposable form via digitization methods that went beyond simple page imaging. There was a critical need to include text conversion elements. This "freeing the data locked on the page" began to be the goal of the Smithsonian Libraries' digitization program. The online version of reference sources began to be scanned with the text created into datasets and was the natural progression from the initial tomes with pleasing plates. Sherborn's *Index Animalium* was one of the Libraries first attempts at digitization for database conversion.

I.

Smithsonian Libraries is a traditional library with books on shelves with librarians and staff ready to assist the patron with their information requests. Traditional library description of monographs and serials is based on standards within the library and information sciences. The inventories of the holdings of large academic libraries require standardized practices and efficiencies of scale to accomplish sophisticated catalogues. The Libraries cover a wide range of topics from art to zoology and is geographically located across the United States and Panama. The data that is captured for each title assists in the physical allocation of material. Yet, the granularity of the descriptive data is effectively only at the title level and does not delve into the contents, chapter, article or page level with in each title, and does not index the specific details within the texts. Most libraries' tools dating back to the card catalogues to the current online integrated library catalogues, have found that the metadata describing titles has worked with limited success. The discovery aspect of this overarching or "high" level of metadata limits the results of inquiries but sufficed for physical discovery of the titles.

Taxonomic research requires the specific citation of descriptions of species. The International regulatory codes for identifying and naming of species require in-depth research on species and genus. The rules are quite clear that when naming species, the name is considered fully formed once the description is published and available (International Commission on Zoological Nomenclature's International Code, http://www.iczn.org/iczn/index.jsp; See specifically Article 8 and Article 11). Major natural history libraries and the Smithsonian's National Museum of Natural History Library, specifically, have served the function of ensuring that publications of species names are stored and the publications are available. Yet the librarian standards of description of

splendens Podontia, Guérin-M. in Duperry, Voy. "Coquille," Zool. II (2) 1 (1838), 144.
splendens Pseudoophonus (Gebl.), V. v. Mochulski, Kaefer Russl., *post* May 1850, 31.—
splendens Pupa, Menke in L. Pfeiffer, Symb. Helic. I. 1841, 45. [Ophonus
splendens [Pycnodus], H. v. Meyer, N. Jahrb. f. Min. 1847, 574 [*n. n.*].
splendens Pycnodus, H. v. Meyer, Palaeontographica, I (5) Dec. 1849, 239.
splendens Rhizomys (Ruepp.), J. A. Wagner in Schreber, Säugth., Suppl. III. 1842, 368.—Bathyergus.
splendens Rissoa, O. G. Costa, Atti R. Accad. Sci. [Napoli] IV (Zool.) 1839, 192 [*n. n.*].
splendens Rissoa, O. G. Costa, Corrisp. Zool. I. 1839, 71 [*n. n.*].
splendens Saprinus (Payk.), W. F. Erichson, Jahrb. Insectenk. I. 1834, 178.—Hister.
splendens Sargus, J. W. Meigen, Klassif. Zweifl. Insekt. I (1) 1804, 144.
splendens Saxicola, Quoy & Gaimard in d'Urville, Voy. "Astrolabe," Zool. I. 1830, 197.
splendens Scarabaeus, Palisot de B., Ins. Afr. Amér. (—) 1809, 89.
splendens Sertularia, J. V. F. Lamouroux, Hist. Polyp. 1816, 191.
splendens Solen, J. C. Chenu, Illustr. Conch. I (18 & 19) 1843, pl. 8 [*n. et f.*].
splendens Sphenoptera, Laporte & Gory, H. N. Coléopt. II. 1839, 29.
splendens Sponsor, F. E. Guérin-M., Rev. Zool. (Soc. Cuv.) III. Dec. 1840, 357.
splendens Staphylinus, T. Marsham, Coleopt. Brit., Sept. 1802, 524.
splendens Sterculia, E. Blanchard in d'Orbigny, Amér. Mérid. (Ins.) 1842, 83.
splendens Tellina, J. De C. Sowerby, Tr. Geol. Soc. London [2] V. 1837, 136.
splendens Temnoscheila, G. R. Gray in Griffith's Cuvier, Anim. Kingd., Ins. II. 1832, 93.
splendens Tetragonus, Gory & Percheron, Mon. Cétoines (2) 1834, 67.
splendens Tityra (Wied), G. R. Gray, Gen. Birds, I. June 1846, 254.—Muscicapa.
splendens Toxotus, Laich.; Dejean, Catal. Coléopt. 1821, 112 [*n. n.*].
splendens Trochus, C. G. Giebel, De Geogn. Hercyn. 1848, 28 [*n. n.*].
splendens Turbo, O. G. Costa, Cat. Test. Sicilie, 1829, pp. ci & civ.
splendens Turdus, W. E. Leach, Zool. Miscell. II. 1815, 30. [1800.
splendens Turdus (Daud.), L. P. Vieillot, N. Dict. H. N., ed. 2, XX. 1818, 260.—Sturnus,
splendens Udorpes, V. v. Motschulsky, Bull. Soc. Imp. Nat. Moscou, XVIII. 1845 (1) 1845, [108.
splendens Unio, G. A. Goldfuss, Petref. German. II (6) 1837, 183.
splendens Venus, D. Solander, Catal. Portland Mus. 1786, 102 & 149 [*n. n.*].
splendens Volatinia (Vieill.), C. L. Bonaparte, Consp. gen. Avium, I. 1850, 474.—Fringilla.
splendescens Pagurus, R. Owen, Zool. "Blossom," 1839, 81.
splendicans Buprestis, Norw.; J. Sturm, Catal. Ins. Samml. 1826, 105.
splendidalis Epipagis (Cr.), J. Huebner, Verz. bekannt. Schmett. 1826, 358.—Phalaena,
splendidella Coleophora, F. Lienig, Isis (Oken), 1846, 296. [1781.
splendidissima Lampyris, F. Green, Doughty's Cab. N. H. II (3) 1832, 55 [*teste* C. W. R.].
splendidissima Laphria, C. R. W. Wiedemann, Aussereurop. Zweifl. Insekt. II. 1830, 645.
splendidiventris Buprestis, Laporte & Gory, H. N. Coléopt. I. 1838, 69.
splendidulana Coccyx, A. Guénée, Ann. Soc. Ent. France [2] III (2) Oct. 1845, 179.
splendidula Baridius, Cristofori & Jan, Cat. Mus. (III a, Coleopt.) 1832, 64 [*n. n.*].

Figure 1. Page from *Index Animalium* showing examples of microcitations.

these materials has fallen short of the needs of the taxonomic researcher in identifying exactly where descriptions are located within these publications – the page level metadata and the data within the page is lacking.

Smithsonian Libraries staff met with researchers from the taxonomic community to understand their path to the literature and identified resources used to connect to the library holdings. The traditional library metadata describing the titles on the shelves of libraries was not meeting the needs of the researcher looking for more detailed and granular data within the texts. Their own bibliographies and indexes were required to pinpoint the data needed. From those sources, data points had to be mapped to the library search interface (the online catalogue) with different terminology and assump-

tions. Each individual researcher had to interpret and translate access points to locate the desired material.

As seen in the other essays within this compilation, Charles Davies Sherborn stepped in to fill the data needs that the traditional library catalogues were not and could not meet. (Neal Evenhuis. "Sherborn: Work history and impact of bibliography, dating and zoological informatics.") The beauty of his *Index Animalium* are the microcitations for a species giving the genus, species, author, abbreviated title of publication, and the critical date and page specifics. This level of access within the texts of monographs and serials is the data that the libraries were failing to deliver. Smithsonian Libraries saw that Sherborn's *Index* was actually a data set that needed to be liberated off the printed page and made available digitally. The microcitations were needed in the electronic world to interact with taxonomists working in the digital world – writing, citing, and interacting with their research. The first task at hand was to scan and make available a fully searchable *Index Animalium*.

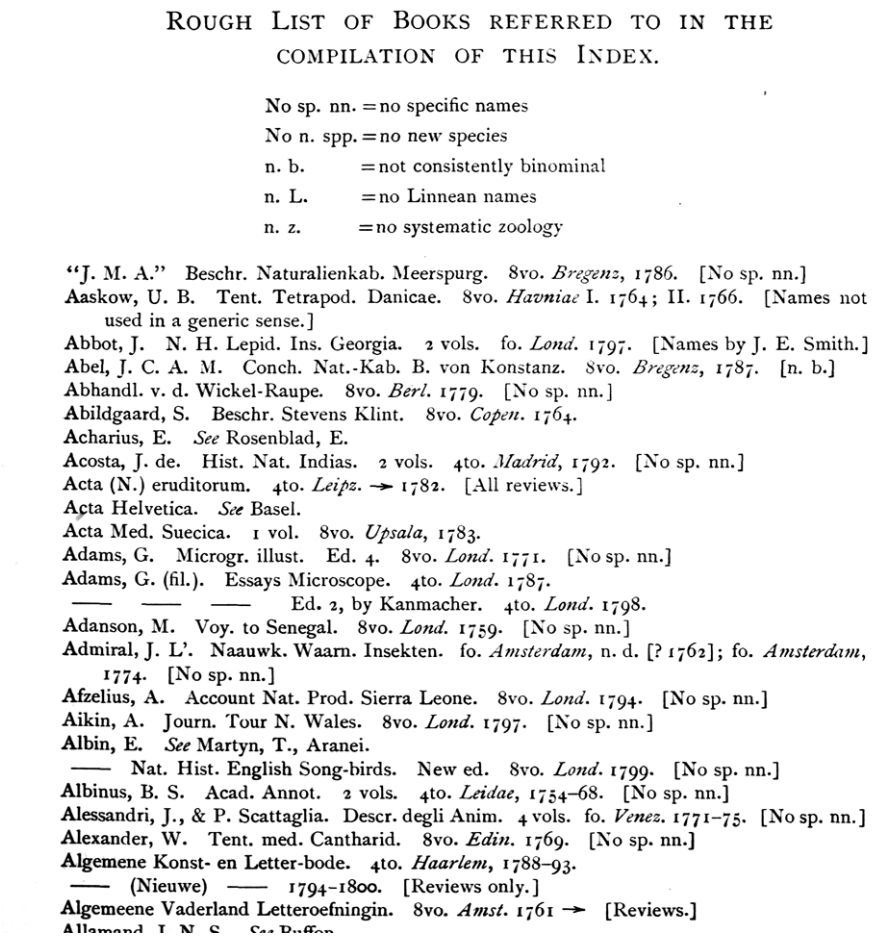
Funded by the Atherton Seidell Endowment Fund, SIL contracted to have the entire set of 30+ volumes scanned: cover to cover, over 9,000 page images. Subsequent to the imaging, the entire *Index* was re-keyed into a database. Spot checked and refined with the vendor, the final database has an accuracy rate of 99.995% and consists of over 430,000 lines of useful data. The *Index Animalium* electronic version is available at http://www.sil.si.edu/digitalcollections/indexanimalium/.

II.

The first goal of the digital e-version of Sherborn's Index, was to provide to the world a searchable version of the full tex of the index and the accompanying bibliographies. As the project continued, it became a mission to identify every volume that Sherborn examined in creating the Index. Once identified, the volumes could then be physically located with first preference being our own Smithsonian Libraries' collection. If the title was not in Smithsonian's holdings, a location would be sought within the realm of natural history libraries. This layer of access was to assist anyone using the online *Index Animalium*'s microcitations to be able to locate the book that Sherborn references.

Sherborn states in the Epilogue of *Index Animalium*, March 1922: "In any well-appointed Natural History Library there should be found every book and every edition of every book dealing in the remotest way with the subjects concerned." With over 7,700 titles listed, Sherborn gives the most comprehensive list of all important works in the study of zoology. The four bibliographies scattered throughout the multi-volume Index records every title that Sherborn examined. He included indications if the work had no systematic zoological name, no Linnaean names, inconsistent binomial names, no specific names mentioned, or if no new species were found in the texts.

Researching all the potentially related species is required for the study of species naming. The Smithsonian Libraries' online version of Sherborn's *Index* is aimed to facilitate the researcher locating all the texts that are referenced. Sherborn's bibliog-



ROUGH LIST OF BOOKS REFERRED TO IN THE COMPILATION OF THIS INDEX.

No sp. nn. = no specific names
No n. spp. = no new species
n. b. = not consistently binominal
n. L. = no Linnean names
n. z. = no systematic zoology

"J. M. A." Beschr. Naturalienkab. Meerspurg. 8vo. *Bregenz*, 1786. [No sp. nn.]
Aaskow, U. B. Tent. Tetrapod. Danicae. 8vo. *Havniae* I. 1764; II. 1766. [Names not used in a generic sense.]
Abbot, J. N. H. Lepid. Ins. Georgia. 2 vols. fo. *Lond.* 1797. [Names by J. E. Smith.]
Abel, J. C. A. M. Conch. Nat.-Kab. B. von Konstanz. 8vo. *Bregenz*, 1787. [n. b.]
Abhandl. v. d. Wickel-Raupe. 8vo. *Berl.* 1779. [No sp. nn.]
Abildgaard, S. Beschr. Stevens Klint. 8vo. *Copen.* 1764.
Acharius, E. *See* Rosenblad, E.
Acosta, J. de. Hist. Nat. Indias. 2 vols. 4to. *Madrid*, 1792. [No sp. nn.]
Acta (N.) eruditorum. 4to. *Leipz.* → 1782. [All reviews.]
Acta Helvetica. *See* Basel.
Acta Med. Suecica. 1 vol. 8vo. *Upsala*, 1783.
Adams, G. Microgr. illust. Ed. 4. 8vo. *Lond.* 1771. [No sp. nn.]
Adams, G. (fil.). Essays Microscope. 4to. *Lond.* 1787.
—— —— —— Ed. 2, by Kanmacher. 4to. *Lond.* 1798.
Adanson, M. Voy. to Senegal. 8vo. *Lond.* 1759. [No sp. nn.]
Admiral, J. L'. Naauwk. Waarn. Insekten. fo. *Amsterdam*, n. d. [? 1762]; fo. *Amsterdam*, 1774. [No sp. nn.]
Afzelius, A. Account Nat. Prod. Sierra Leone. 8vo. *Lond.* 1794. [No sp. nn.]
Aikin, A. Journ. Tour N. Wales. 8vo. *Lond.* 1797. [No sp. nn.]
Albin, E. *See* Martyn, T., Aranei.
—— Nat. Hist. English Song-birds. New ed. 8vo. *Lond.* 1799. [No sp. nn.]
Albinus, B. S. Acad. Annot. 2 vols. 4to. *Leidae*, 1754–68. [No sp. nn.]
Alessandri, J., & P. Scattaglia. Descr. degli Anim. 4 vols. fo. *Venez.* 1771–75. [No sp. nn.]
Alexander, W. Tent. med. Cantharid. 8vo. *Edin.* 1769. [No sp. nn.]
Algemene Konst- en Letter-bode. 4to. *Haarlem*, 1788–93.
—— (Nieuwe) —— 1794–1800. [Reviews only.]
Algemeene Vaderland Letteroefningin. 8vo. *Amst.* 1761 → [Reviews.]
Allamand, J. N. S. *See* Buffon.

Figure 2. Example of the first page of the first bibliography (Sherborn. *Index Animalium*, Vol. 1, p. xi).

raphy, though a very comprehensive list of important titles, is not complete in the descriptions of these titles. His use of inconsistent abbreviation, "ibid" indications, use of shortened titles, and other idiosyncrasies has made identifying the exact titles challenging.

III.

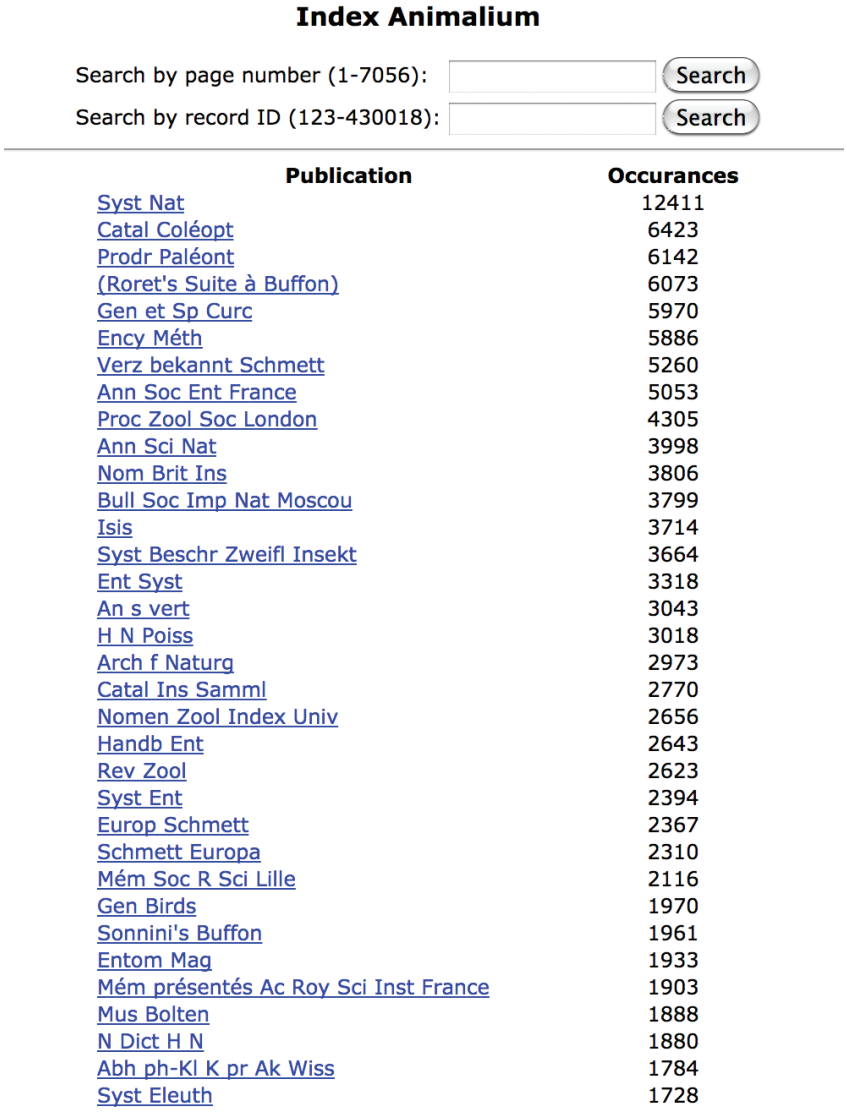
Smithsonian Libraries first foray into moving beyond pretty books to creating datasets faced many challenges including re-assessing the actual needs and deliverables of the project. The online project morphed from the initial basic scanning of the text – to a searchable database – to a goal of connecting each microcitation to the proper line in the bibliography – to the goal of having the microcitation connected to bibliography connected to physical location of the text. Difficulties emerge when computer-to-

—— Ad orat. aud. De primis anim. vert. 4to. *Jenae*, 1836.
—— Hist.-topogr. Taschenb. Jena. 8vo. *Jena*, 1836.
Zetterstedt, J. W. Orthopt. Suec. 8vo. *Lund*. 1821.
—— Resa Sveriges och Norriges Lappmarkes. 2 vols. 8vo. *Lund*, 1822.
—— Fauna Ins. Lappon. 8vo. *Hann*. 1828.
—— Resa genom Umea Lappmarker. 8vo. *Oreb*. 1833.
—— Ins. Lappon. descr. 6 pts. 4to. *Lips*. 1838–40. [Pp. 1–867, 1838; –1013, 1839; –1140, 1840.]
—— Dipt. Scand. 9 vols. 8vo. *Lund*. 1842–50. [The remainder of this work does not come within the scope of this Index.]
Zhivonishuii Sbornik. 1 vol. 4to. *St Petersb*. 1850. [B. M., popular.]
"Ziegler." [In molluscan literature his n. spp. were recorded by Menke and by Rossmaessler.]
Zieten, C. H. de. Les Pétrif. de Wurtemberg. 12 pts. fo. *Stoutg*. 1830(–33). [(1 & 2) pp. 1–16, 1830; (3 & 4) 17–32, 1831; (5–8) 33–64, 1832; (9–12) 65–96, 1833.]
Zigno, A. de. Sopra corpi organ. infusioni. Ed. 2. 8vo. *Padova*, 1842.
—— Sopra due foss. calc. Monte Padovani. 4to. *Pad*. 1845. [G. S.]
—— Sul cret. Ital. sept. 4to. *Pad*. 1846. [N. Saggi R. Acc. Sci. Padova, VI; no n. spp.]
—— Nota nov. prom. foss. il Bianconi. 8vo. *Ven*. 1846. [Atti Adun. I. R. Ist. Sci. Venezia; G. S.]
Zimmermann, C. Monogr. Carabiden. 8vo. *Berl*. 1831.
—— Das Harzgebirge. 8vo. *Darmst*. 1834. [G. S., lists.]
Zimmermann, E. A. W. Taschenb. d. Reisen. 8vo. *Leipz*. 1819. [*Teste* M. J. Rathbun.]
Zodiac. 12 nos. 4to. *Albany*, 1835–6. [B. M., no n. spp.]
Zoologe (Der). 8 pts. 8vo. *Eisenach*, 1795–97. [This is the XXI Part of "Compendiose Bibliothek d. gemeinnuetz. Kenntn. f. alle Staende," lent by the Univ. of Kiel to the B. M. (N. H.), 1907; I take the "B.—n" to be Beckstein, but Gerritt Miller and Knud Andersen think it Borkhausen.]
Zoological Journ. (The). 20 pts. in 5 vols. 8vo. *Lond*. 1824–35; 5 pts. of supplementary plates.
Zoological Magazine or Journ. of Nat. Hist. 6 pts. 8vo. *Lond*. 1833–34.
Zoological Miscellany (Gray's). 6 pts. 8vo. *Lond*. 1821–44. [Pt. 1, pp. 1–40, Feb. 1831; the other parts are dated at sign. F, G, H, I, and K.]
Zoological Miscellany (Leach's). 3 vols. 8vo. *Lond*. 1814–17. [Dates uncertain; I adopt on my evidence I, 1814; II, 1815; III, 1817; it is almost certain that vol. I began Jan. 1814.]
Zoologisches Magazin. 8vo. *Kiel*, 1817–23.
Zoologist (The). I—VIII. 8vo. *Lond*. 1843–50.
Zuerich. Nat. Gesellsch. An die zuerisch. Jugend. 4to. *Zuer*. 1799–1850; Berichte, 1826–38; Mittheil. I, II, 1847–50.

Figure 3. One example page from *Index Animalium* bibliography.

computer resolving of microcitations and bibliography entries were attempted. Most problematic was the use of computer scripts against Sherborn's inconsistent notation made it impossible for clear connections of species citations to title citation to be made in a systematic way.

Simple Regular Expressions were used to break apart the re-keyed text of Sherborn based on the lessons learned by MBL WHOI Library's project for Neave's *Nomenclator Zoologicus*. (*Nomenclator Zoologicus* online version from uBio, Marine Biological Laboratory, Woods Hole Oceanographic Institution http://uio.mbl.edu/NomenclatorZoologicus/) Regular Expressions are a simple syntax particularly suited for identifying and dividing up textual data by looking at patterns, punctuation, and even character strings or sequences. David Remsen and Patrick Leary (formerly at MBL/WHOI) used these parsing techniques to isolate titles within the *Index's* species citations. Using those strings they used comparisons of strings to match against the bibliography citations.



Index Animalium

Search by page number (1-7056): Search

Search by record ID (123-430018): Search

Publication	Occurances
Syst Nat	12411
Catal Coléopt	6423
Prodr Paléont	6142
(Roret's Suite à Buffon)	6073
Gen et Sp Curc	5970
Ency Méth	5886
Verz bekannt Schmett	5260
Ann Soc Ent France	5053
Proc Zool Soc London	4305
Ann Sci Nat	3998
Nom Brit Ins	3806
Bull Soc Imp Nat Moscou	3799
Isis	3714
Syst Beschr Zweifl Insekt	3664
Ent Syst	3318
An s vert	3043
H N Poiss	3018
Arch f Naturg	2973
Catal Ins Samml	2770
Nomen Zool Index Univ	2656
Handb Ent	2643
Rev Zool	2623
Syst Ent	2394
Europ Schmett	2367
Schmett Europa	2310
Mém Soc R Sci Lille	2116
Gen Birds	1970
Sonnini's Buffon	1961
Entom Mag	1933
Mém présentés Ac Roy Sci Inst France	1903
Mus Bolten	1888
N Dict H N	1880
Abh ph-Kl K pr Ak Wiss	1784
Syst Eleuth	1728

Figure 4. uBio parsing showing count of abbreviations found in publication area of microcitation.

Documentation regarding the parsing of *Index Animalium* data is found at http://uio.mbl.edu/Sherborne/index.html.

As seen in some of the examples below, there were some high accuracy results at times and mixed results in others. Problems with titles that are extremely common in the field of taxonomy, that Sherborn abbreviated in a way that the researcher could recognize when reading the citations, fall short when attempting to use computerized matching. Sherborn was not consistent in his abbreviation within the microcitations. Within the bibliography, he was not consistent with title, author, editor, edition, vol-

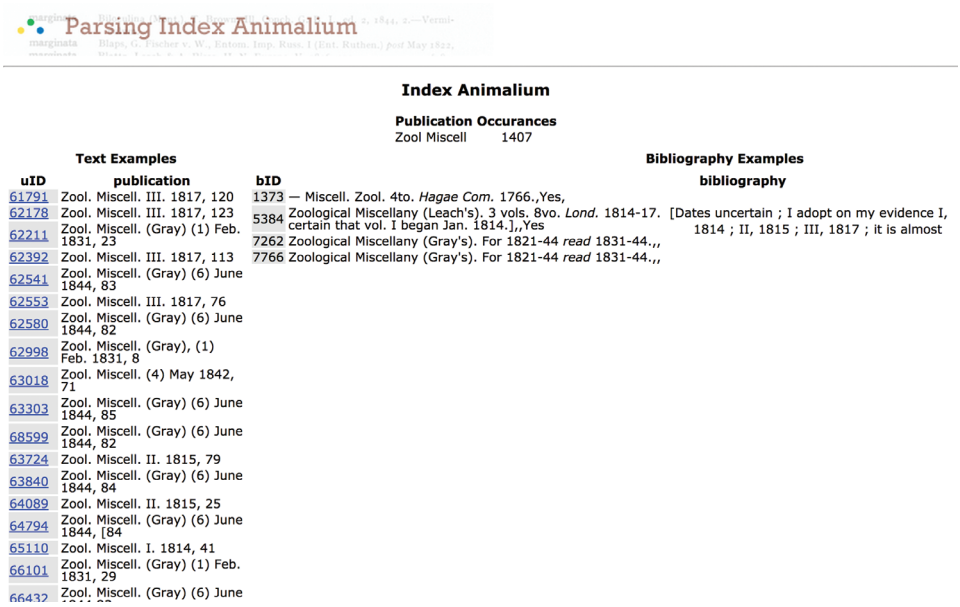


Figure 5. uBio parsing. Microcitation "Zool Miscell" found 1407 times in the *Index Animalium* and potentially matches four entries in Sherborn's bibliographies.

ume, publisher, or places of publication abbreviations. He was not always consistent on what he decided to abbreviate or how he formed the abbreviations. Having a systematic string matching between the citations to the bibliography did not prove to provide the clean matching that was needed for unambiguous one to one matching. The metadata fields do not line up for easy comparison and matching; for example: author to author versus author to editor. Century-old systems of notation translated into library standard database structure have been a road-block in speedily unlocking and connecting the data. (Pilsk, S.C., et al. "The Biodiversity Heritage Library: Advancing Metadata Practices in a Collaborative Digital Library" Journal of Library Metadata 10:136-155, 2010 doi: 10.1080/19386389.2010.506400) Connecting the index to the bibliography and the bibliography to library holdings has required many more hands and eyes than lines of script. Staff, interns and volunteers began to attempt to locate standard library records for each title in the bibliography.

IV.

A Microsoft Access database was constructed that contained only the bibliography from *Index Animalium*. Sherborn's bibliography entries were sorted by the greatest number of associated microcitations. These were searched and the full title, author, date and related identifiers were added to a database. Each line of data from the *Index's* bibliographies was matched against standard library data using Smithsonian's online

ID	21258
Bid	5384
Orig	Zoological Miscellany (Leach's). 3 vols. 8vo. Lond. 1814-17. [Dates uncertain ; I adopt on my evidence I, 1814 ; II, 1815; III, 1817 ; it is almost certain that vol. I began Jan. 1814.],,Yes
uBio Counts	1407
Has Species	Yes
Title	The zoological miscellany : being descriptions of new, or interesting animals
Author	Leach, William Elford
WorldCat ID	4915037
SIL SIRIS ID	120341
BHL Link	41372

Figure 6. Data from the Smithsonian Libraries database for one line in the *Index Animalium* Bibliography.

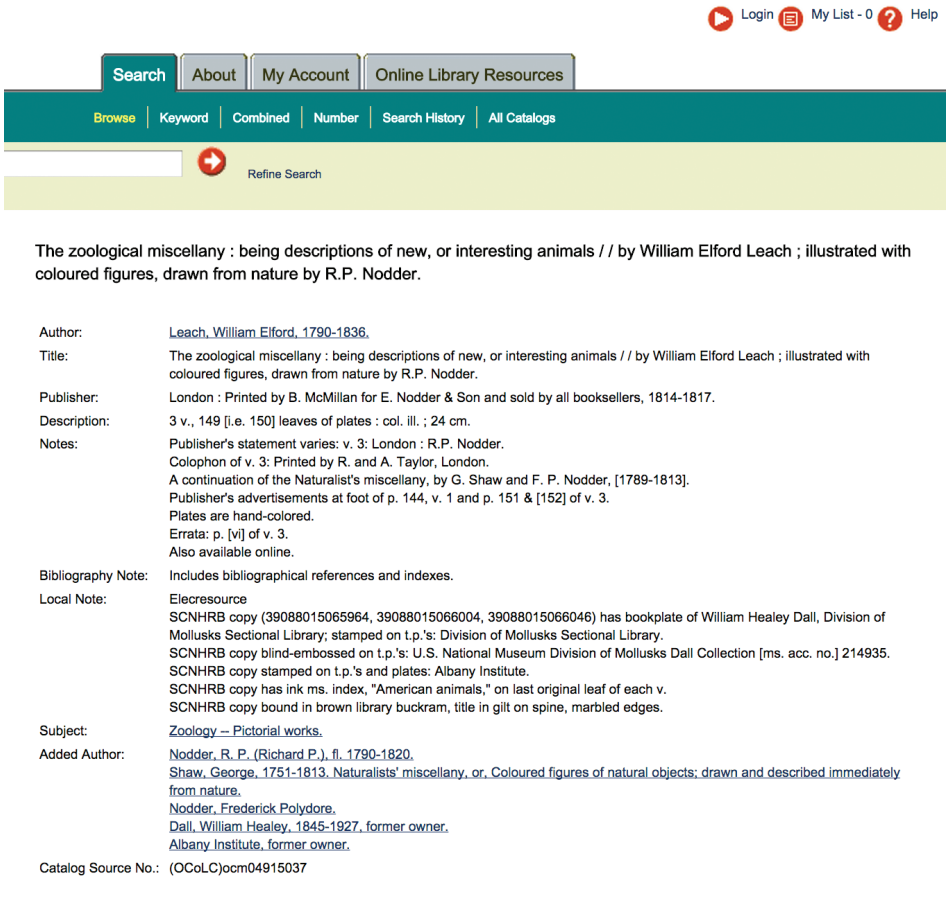


Figure 7. Screen capture of the Smithsonian Institution Online Catalog SIRIS for William Elford Leach's Zoological Miscellany.

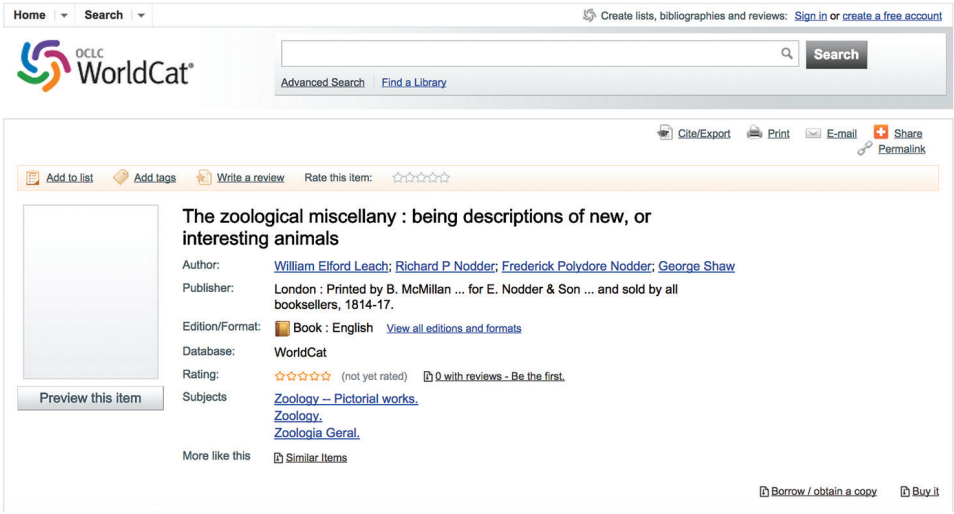


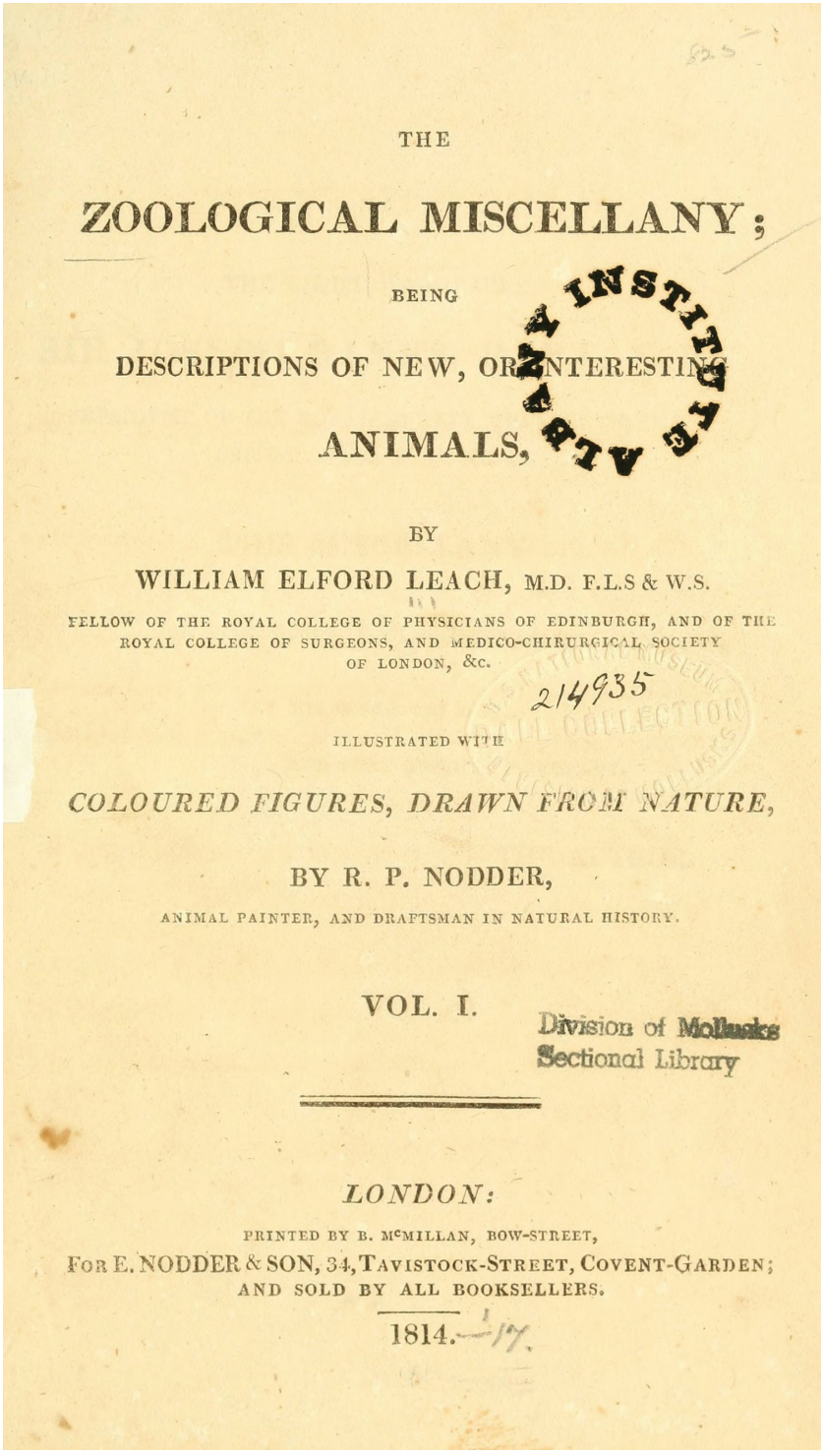
Figure 8. Screen capture from OCLC's WorldCat for William Elford Leach's Zoological Miscellany.

catalogue (SIRIS) and against the OCLC WorldCat catalogue. (Smithsonian Libraries' online catalog SIRIS is searchable via http://siris-libraries.si.edu.WorldCat is considered to be the largest network of library data. Listing library holdings from around the world, it contains metadata describing these titles following international standards. OCLC's WorldCat is available for searching http://www.worldcat.org/).

V.

The Smithsonian Libraries' *Index Animalium* project took a new direction as the Biodiversity Heritage Library (BHL) project began production. A large scale scanning project, BHL's mission is to digitize legacy natural history literature that is significant in the study and research of biodiversity. BHL is made up of a consortium of international natural history and botanical libraries. Specific funding of the BHL supports the scanning of the literature published before 1923 – titles that Sherborn referenced in his bibliographies. BHL ramped up fairly quickly and began to have full text scans online in 2005. Libraries participating in identifying and scanning the literature stretch across the globe and continue to produce millions of pages of online text ever year. Biodiversity Heritage Library information can be found at http://biodivlib.wikispaces.com/ and the collection is searchable http://www.biodiversitylibrary.org/.

As more and more of the literature becomes available online via the BHL, the *Smithsonian Libraries Index Animalium* project has, once again, shifted in the goal of service to the taxonomic researcher. Instead of getting the researcher to the library shelf for the text, it is becoming more desirable to deliver the fully scanned text to the researcher. Currently the matching of scanned titles is underway with identified titles in



THE

ZOOLOGICAL MISCELLANY;

BEING

DESCRIPTIONS OF NEW, OR INTERESTING

ANIMALS,

BY

WILLIAM ELFORD LEACH, M.D. F.L.S & W.S.

FELLOW OF THE ROYAL COLLEGE OF PHYSICIANS OF EDINBURGH, AND OF THE ROYAL COLLEGE OF SURGEONS, AND MEDICO-CHIRURGICAL SOCIETY OF LONDON, &c.

ILLUSTRATED WITH

COLOURED FIGURES, DRAWN FROM NATURE,

BY R. P. NODDER,

ANIMAL PAINTER, AND DRAFTSMAN IN NATURAL HISTORY.

VOL. I.

LONDON:

PRINTED BY B. McMILLAN, BOW-STREET,
FOR E. NODDER & SON, 34, TAVISTOCK-STREET, COVENT-GARDEN;
AND SOLD BY ALL BOOKSELLERS.

1814.

Figure 9. Title page from Biodiversity Heritage Library for William Elford Leach's Zoological Miscellany.

the *Index's* bibliographies. The anticipated result will have the researcher "click and go" from Sherborn's *Index* online to the BHL scanned text online. Apparently seamless, the connections being made behind the scenes match the online *Index's* microcitation to the full title record and resolve to the proper title in BHL.

VI.

Partnerships forming over the use of Sherborn's *Index Animalium* distribute the work into more functional pieces to achieve the seamless online research tool. Richard Pyle of the Bishop Museum is applying some reviews to the re-keyed text and providing complete citations for items that have partial data in the Smithsonian database. Another partnership is with Thomson Reuter's staff working on ION: Index to Organism Names. Thomson Reuters Index to Organism Names (ION) http://www.organism-names.com/ is a free online service to search the names included in *Zoological Record*, a continuously updated database of biological taxonomic research. As Nigel Robinson's presentation "Sherborn's *Index Animalium* Integration into ION: Access to All" demonstrated, the parsing of the microcitations and identifying the full text is underway increasing the data in ION and providing the connection needed for the taxonomic

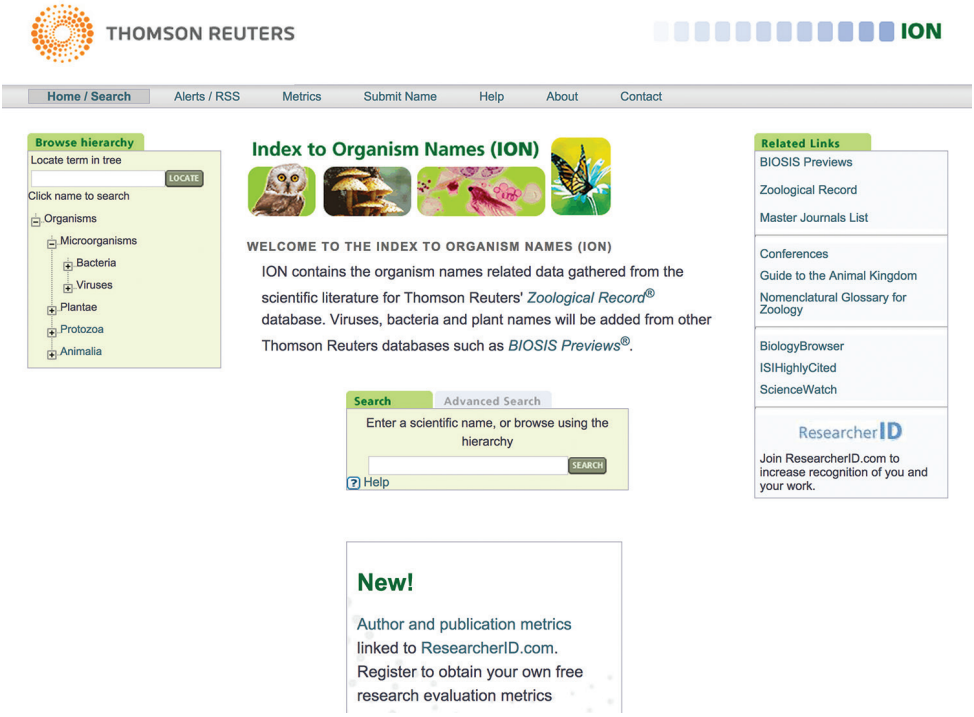
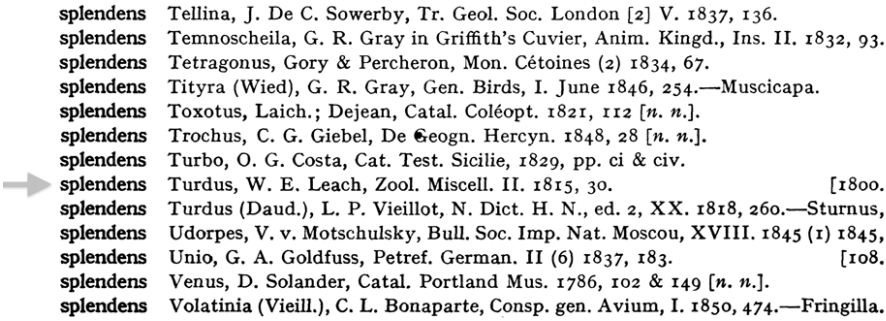


Figure 10. ION home page screen capture.



splendens Tellina, J. De C. Sowerby, Tr. Geol. Soc. London [2] V. 1837, 136.
splendens Temnoscheila, G. R. Gray in Griffith's Cuvier, Anim. Kingd., Ins. II. 1832, 93.
splendens Tetragonus, Gory & Percheron, Mon. Cétoines (2) 1834, 67.
splendens Tityra (Wied), G. R. Gray, Gen. Birds, I. June 1846, 254.—Muscicapa.
splendens Toxotus, Laich.; Dejean, Catal. Coléopt. 1821, 112 [*n. n.*].
splendens Trochus, C. G. Giebel, De Geogn. Hercyn. 1848, 28 [*n. n.*].
splendens Turbo, O. G. Costa, Cat. Test. Sicilie, 1829, pp. ci & civ.
splendens Turdus, W. E. Leach, Zool. Miscell. II. 1815, 30. [1800.
splendens Turdus (Daud.), L. P. Vieillot, N. Dict. H. N., ed. 2, XX. 1818, 260.—Sturnus,
splendens Udorpes, V. v. Motschulsky, Bull. Soc. Imp. Nat. Moscou, XVIII. 1845 (1) 1845,
splendens Unio, G. A. Goldfuss, Petref. German. II (6) 1837, 183. [108.
splendens Venus, D. Solander, Catal. Portland Mus. 1786, 102 & 149 [*n. n.*].
splendens Volatinia (Vieill.), C. L. Bonaparte, Consp. gen. Avium, I. 1850, 474.—Fringilla.

Figure 11. Sherborn citation from page 6101 of the Index Animalium: *Turdus splendens*, W.E. Leach, Zool. Miscell. II. 1815, 30.

Smithsonian ID	362382
Smithsonian IA Page Image ID	SIL34_02_24_0193
IA Page	6101
Species	splendens
Genus	Turdus
Author Forename	W.E.
Author Last name	Leach
Abbreviated Publication Name	Zool. Miscell.
Volume	II.
Year	1815
Page	30

Figure 12. Parsed Data from Smithsonian Libraries Sherborn Database.

researcher. Slides from Nigel Robinson's presentation are available at http://www.slideshare.net/iczn/4-sherborns-index-animalium-integration-into-ion.

The ION team working on *Index Animalium* at Thomson Reuters is looking at supplementing the *Zoological Record* dataset. Sherborn's data back fills ION with taxonomic names for 1758 to 1864. To achieve the data extraction from *Index Animalium*, Robinson reports that there are challenges in parsing and properly identifying the data elements. The review of the data is needed since Sherborn's use of commas, brackets and notations all have meanings that need to be carefully interpreted so as to not lose the intention. As the project has progressed, inconsistencies are coming to light that can now be documented. With this detailed look, the ION team is finding re-keying errors, as well as errors made by Sherborn, and the typesetting done based on Sherborn's initial transcriptions.

ION's management classification protocol is also being added to the microcitation so that the data can be processed and incorporated into the systems already in place at Thomson Reuter. The species and genus identified in Sherborn are being folded into the overall

Genus	Turdus
Species	splendens
Name string	Turdus splendens
Author last name	Leach
Author forename	W.E.
Abbreviated publication name	Zool. Miscell.
Full Publication name	The Zoological Miscellany
Volume	II
Year	1815
Page	30

Figure 13. Focusing on the species and genus names, ION is examining the publication abbreviation against the internal Zoological Record data and resolving the microcitations.

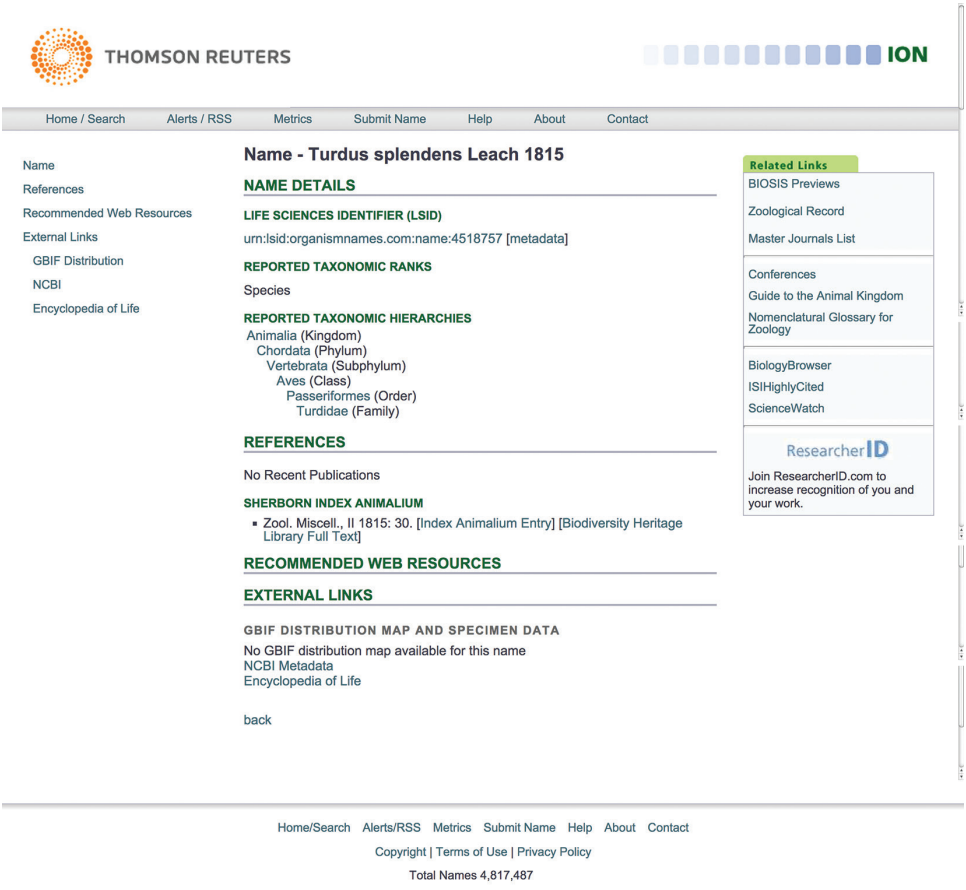


Figure 14. Results of searching in ION for *Turdus splendens*.

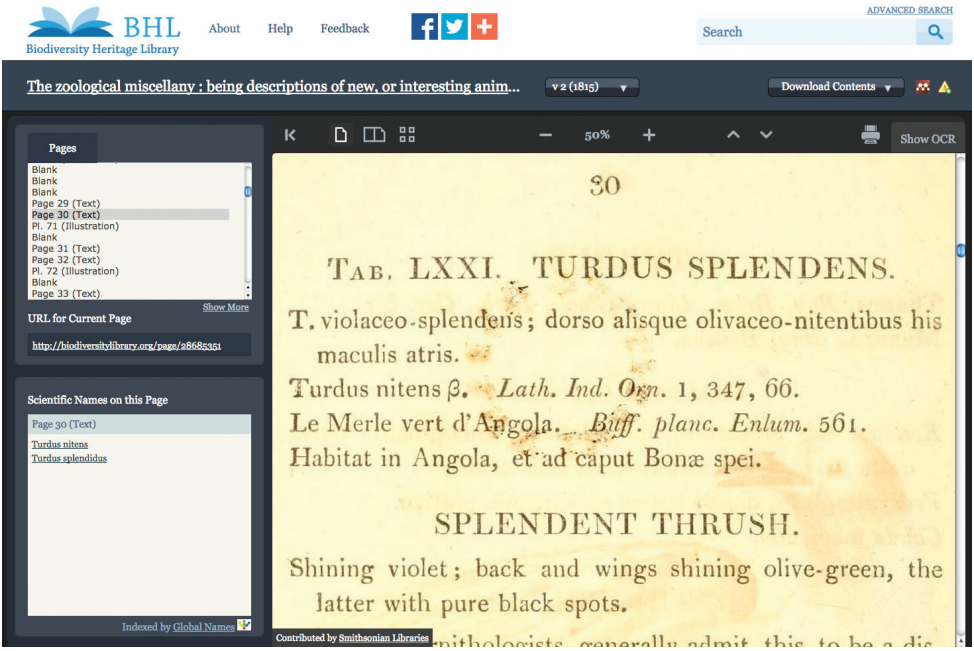


Figure 15. Results of a direct link from ION *Turdus splendens* to the page in the Biodiversity Heritage Library.

delivery of data via the ION search. Robinson's presentation illustrated the parsing with an example of one line of data from Smithsonian Libraries Sherborn Database that teases out the identification of the citation in the Smithsonian database and the various elements.

(View the page of *Index Animalium* for this reference http://www.sil.si.edu/digitalcollections/indexanimalium/volumes/pagedisplaypage.cfm?filename=SIL34_02_24_0193)

From this breakdown and reconfigurations, ION is able to map data into ION and integrate with the existing ION content to form a nomenclator of names for the literature published from 1758 onwards.

BHL provides stable consistent page identifiers for all titles scanned. In this example *Turdus splendens* page identifier http://biodiversitylibrary.org/page/28685351 is a persistent identifier allowing ION to create a direct link into the Biodiversity Heritage Library. The results are the "click and go" for the user to reach the page of the text Sherborn cites.

VII.

The challenges of working with legacy taxonomic citations, computer matching algorithms, and making connections have not stopped the attempts to continually improve the reach of Sherborn's unique and critical data to the researcher. New developments

and constant revisiting of the goals has brought us to yet another shift to today's goal of making Sherborn's *Index* available and linked to other datasets. The Smithsonian Libraries is exploring a different data structure than a relational database currently in use. Partnering with others in the world of metadata development and information sharing has led to an attempt to allow machine-to-machine communications. The *Index* is being looked at as the data set of the elements it contains. These data points are being examined for possible transformation into RDF mark up and meeting the standards of Linked Open Data. This will allow for broader discovery and access than a stand-alone database. Linked Open Data is primarily aimed at consumption by computer software, but the availability of such data allows the offering of an online research tool geared towards the general population of natural history researchers.

Linked Data is based on the concept of triples or a sentence made up of three parts: subject, predicate, and object. The subject is an identifiable "thing" that can be assigned a unique identifier. The predicate can be considered the "verb" with a controlled vocabulary that has a term defined and assigned a unique identifier. The object is the last "thing" in the triple that subject is connected. A possible triple that would be created from Sherborn's *Index Animalium* is diagrammed below. In this scheme, each species is presented as an identifier with related microcitation data pointing to the

Subject	<genus>Turdus <species>splendens
Predicate	Authored by
Object	<lastname>Leach<forename>William Elford
Subject	<genus>Turdus <species>splendens
Predicate	Published in
Object	The zoological miscellany : being descriptions of new, or interesting animals

Figure 16. Example of potential triples from Index Animalium's citation for "splendens Turdus".

scanned title and page at BHL. The goal of providing the Index as an open data set in the RDF would allow others to reuse, repurpose, and mine the data.

The details of creating a complete open linked data set out of *Index Animalium* are still being discussed and explored. Smithsonian Libraries, dedicated to providing data in an open platform, is already beginning to work on providing some Open Linked Data in a new project based off the *Taxonomic Literature: A selective guide to botanical publications and collections with dates, commentaries and types*, 2nd edition. Known by most as TL2, the entire 15 volume set has been scanned and OCRed. The data is currently available for searching and the break down into triples has begun. Smithsonian Libraries TL2 online (http://www.sil.si.edu/digitalcollections/tl-2) allows for reading or searching the entire text of the literature of systematic botany published between 1753 and 1940. Incorporating *Index Animalium*, Smithsonian's goal is to create a TL3: an online resource containing both botanical and zoological linked open data resource for taxonomic research.

VIII.

A project that began to simply provide a URL for anyone in the world to read Sherborn's *Index Animalium* has grown and changed as the fast paced world of knowledge sharing has adapted to the technology available. The *Index* has matured from the pieces of paper of Charles Davies Sherborn's carefully indicated notes of species citations to a linked data structure. The overarching goal of providing access has been achieved but there is room for it to improve by making the information usable, repurpose-able, and integrated into the researcher's workflow.

References

Alonso-Zarazaga MA, Fautin DG, Michel E (2016) The List of Available Names (LAN): A new generation for stable taxonomic names in zoology? In: Michel E (Ed.) Anchoring Biodiversity Information: From Sherborn to the 21st century and beyond. ZooKeys 550: 225–232. doi: 10.3897/zookeys.550.10043

Alonso-Zarazaga MA, Bouchet P, Pyle RL, Kluge N, Fautin DG (2016) Manual for proposing a Part of the List of Available Names (LAN) in Zoology. In: Michel E (Ed.) Anchoring Biodiversity Information: From Sherborn to the 21st century and beyond. ZooKeys 550: 283–298. doi: 10.3897/zookeys.550.10042

Dickinson EC (2016) Reinforcing the foundations of ornithological nomenclature: Filling the gaps in Sherborn's and Richmond's historical legacy of bibliographic exploration. In: Michel E (Ed.) Anchoring Biodiversity Information: From Sherborn to the 21st century and beyond. ZooKeys 550: 107–134. doi: 10.3897/zookeys.550.10170

Evenhuis NL (2016) Charles Davies Sherborn and the "Indexer's Club". In: Michel E (Ed.) Anchoring Biodiversity Information: From Sherborn to the 21st century and beyond. ZooKeys 550: 13–32. doi: 10.3897/zookeys.550.9697

Kalfatovic MR (1998) Re-Defining the Library Meme: Memory and Imagination. In: Wolf MT, Ensor P, Thomas MA (Eds) Information Imagineering: Meeting at the Interface. American Library Association, Chicago, 155–65.

Licklider JCR (1965) Libraries of the Future. MIT Press, Cambridge, 219 pp.

Lyal CHC (2016) Digitising legacy zoological taxonomic literature: Processes, products and using the output. In: Michel E (Ed.) Anchoring Biodiversity Information: From Sherborn to the 21st century and beyond. ZooKeys 550: 189–206. doi: 10.3897/zookeys.550.9702

Marcum D (2013) The Biodiversity Heritage Library: Ithaka S+R Case Study. http://sr.ithaka.org/sites/default/files/reports/SR_BHL_20140129.pdf

Page RDM (2006) Taxonomic names, metadata, and the Semantic Web. Biodiversity Informatics 3: 1–15. doi: 10.17161/bi.v3i0.25

Page RDM (2008) Biodiversity informatics: the challenge of linking data and the role of shared identifiers. Briefings in bioinformatics 9(5): 345–54. doi: 10.1093/bib/bbn022.

Page RDM (2016) Surfacing the deep data of taxonomy. In: Michel E (Ed.) Anchoring Biodiversity Information: From Sherborn to the 21st century and beyond. ZooKeys 550: 247–260. doi: 10.3897/zookeys.550.9293

Pilsk SC, Person MA, deVeer JM, Furfey JF, Kalfatovic MR (2010) The Biodiversity Heritage Library: Advancing Metadata Practices in a Collaborative Digital Library. Journal of Library Metadata 10: 136–155. doi: 10.1080/19386389.2010.506400

Remsen D (2016) The use and limits of scientific names in biological informatics. In: Michel E (Ed.) Anchoring Biodiversity Information: From Sherborn to the 21st century and beyond. ZooKeys 550: 207–223. doi: 10.3897/zookeys.550.9546

Remsen DP, Norton C, Patterson DJ (2006) Taxonomic informatics tools for the electronic Nomenclator Zoologicus. The Biological Bulletin 210(1): 18–24. doi: 10.2307/4134533.

Robinson N (2011) Sherborn's *Index Animalium* Integration into ION: Access to All. Presentation at "Anchoring Biodiversity Information: From Sherborn to the 21st Century and Beyond." Symposium, 28 October 2011. http://www.slideshare.net/iczn/4-sherborns-index-animalium-integration-into-ion

Seitchek C (2003) Century-old biology text gets an e-life. The Torch 3(8): 6–7.

Sherborn CD (1903–33) Index animalium : sive, Index nominum quae ab A.D. MDCCLVIII generibus et speciebus animalium imposita sunt, societatibus eruditorum adiuvantibus. British Museum, London. http://www.sil.si.edu/digitalcollections/indexanimalium/ and http://biodiversitylibrary.org/bibliography/14658

Shindler K (2016) A magpie with a card-index mind – Charles Davies Sherborn 1861–1942. In: Michel E (Ed.) Anchoring Biodiversity Information: From Sherborn to the 21st century and beyond. ZooKeys 550: 33–56. doi: 10.3897/zookeys.550.9975

Smithsonian Libraries (1995) Rare Books and Special Collections in the Smithsonian Institution Libraries. Smithsonian, Washington, DC, 106 pp.

Stafleu A, Cowan RS (1976-1988) Taxonomic Literature: a Selective Guide to Botanical Publications and Collections with Dates, Commentaries and Types (7 volumes). Bohn, Scheltema and Holkema, Utrecht. Online at http://www.biodiversitylibrary.org/item/103414 and http://www.sil.si.edu/digitalcollections/tl-2 [database format]

Thompson CF, Pape T (2016) Sherborn's influence on Systema Dipterorum. In: Michel E (Ed.) Anchoring Biodiversity Information: From Sherborn to the 21st century and beyond. ZooKeys 550: 135–152. doi: 10.3897/zookeys.550.9447

Weitzman AL, Lyal CHC, Garnett T (2004) The Biologia Centrali-Americana Centennial: A vision for electronic access to taxonomic resources, the information interface between libraries and systematic biology. A talk presented at The National Museum of Natural History, Smithsonian Institution, June 2004.

Weitzman AL, Lyal CHC (2004) An XML schema for taxonomic literature – taXMLit. http://www.sil.si.edu/digitalcollections/bca/documentation/taXMLitv1-3Intro.pdf

Weitzman AL, Lyal CHC (2006) INOTAXA – INtegrated Open TAXonomic Access and the "Biologia Centrali-Americana". Proceedings of the Contributed Papers Sessions Biomedical and Life Sciences Division, Special Libraries Association, 8 pp.

Welter-Schultes F, Görlich A, Lutze A (2016) Sherborn's Index Animalium: New names, systematic errors and availability of names in the light of modern nomenclature. In: Michel E (Ed.) Anchoring Biodiversity Information: From Sherborn to the 21st century and beyond. ZooKeys 550: 173–183. doi: 10.3897/zookeys.550.10041