Digitale Entomologische Information aus dem Deutschen Entomologischen Institut

Eckhard K. Groll

Leibniz-Zentrum für Agrarlandschaftsforschung e. V.
Deutsches Entomologisches Institut

Abstract: Digital Entomological Information from the Deutsches Entomologisches Institut

The Deutsches Entomologisches Institut (DEI) in Müncheberg / Germany is constructing an open archive of entomological literature. This paper describes 5 steps in the workflow: selection of literature, digitizing articles or books, generation of man and machine readable documents, classification of documents according to source (syntactic level), and classification of documents according to content (semantic level). The selection concentrates on rare, legal to copy, not yet digitized, taxonomic literature owned by the DEI. A book scanner is used to avoid damage to the sensitive and valuable material. The extracted images are attached to PDF files with text under image by OCR software. To classify the documents a controlled vocabulary of taxa names and entomological terms is used.

Key words: open archive, digitized documents, historical entomological papers

Dr. E. K. Groll, Leibniz-Zentrum für Agrarlandschaftsforschung e. V. - Deutsches Entomologisches Institut, Eberswalder Straße 84, 15374 Müncheberg, E-Mail: groll@zalf.de

Traditionell wurden und werden im Deutschen Entomologischen Institut (DEI) bzw. in Kooperationsprojekten unter Mitwirkung der Wissenschaftler des DEI zahlreiche Informationen mit entomologischem Bezug gesammelt und herausgegeben, z. B. Bibliographien der entomologischen Weltliteratur ((Index I) Horn & Schenkling 1928-1929; Derksen & Scheiding 1963-1975; Gaedike & Smetana 1978, 1984), Verbleib entomologischer Sammlungen (Horn, Kahle & al. 1935, 1990), Typenkataloge sowie aktuell die Zeitschriften „Beiträge zur Entomologie“ und „Nova Supplementa Entomologica“. Bereits digital verfügbar sind weiterhin die „Biographien der Entomologen der Welt“ (Groll 2006), „ECatSym: Elektronischer Katalog der Symphyta der Welt“ (Taeger & Blank 2006) und eine Bildersammlung (Groll & Schubert 2006). Den Forderungen der Informationsgesellschaft gehorchend werden sowohl Daten als auch Metadaten verstärkt in institutseigenen Archiven frei zugänglich digital bereitgestellt und in Verbünden vernetzt.

Material und Methoden

Im folgenden wird ein im Aufbau befindliches Volltextarchiv historischer Zeitschriftenartikel mit entomologischem Inhalt dargestellt. Die Gewinnung der Dokumente erfolgt in folgenden Schritten:

- Auswahl der Objekte
- Scannen der Objekte
- Generierung von durch Menschen und Maschinen lesbaren Dokumenten
- Erschließung der Dokumente auf syntaktischem Niveau
- Erschließung der Dokumente auf semantischem Niveau

Als Masterliste zur Auswahl und Beschreibung der Dokumente für das Archiv wird eine möglichst vollständige Bibliographie der entomologischen Weltliteratur benötigt. Hier bietet sich die gegenwärtig im DEI erarbeitete Neuauflage des Index I an (Taeger, Groll & Gaedike 2007, im vorliegenden Band).

Im ersten Schritt werden, ausgehend von dieser Liste, lohnenswerte Bücher und Zeitschriften ausgewählt. Kriterien sind hierbei u. a.: Ist das Dokument im Bestand der Bibliothek (B15) vorhanden? Ist es frei von Verwertungsrechten Dritter? Ist es selten oder schwer zugänglich? Ist es anderweitig digital verfügbar? Und schließlich, besteht Bedarf an diesem Dokument?

Der nahezu vollständige Bestand der B15 an historischer entomologischer Literatur bietet den Vorteil, schnell an die meisten Originalquellen heran zu kommen.

Bei Werken, die dem Urheberrecht unterliegen, ist die aufwändige individuelle Rechteklärung mit jedem Autor oft eine kaum unüberwindbare Hürde. Da die Autorenrechte erst 70 Jahre nach dem Tod des Autors auslaufen, werden von vielen anderen Projekten insbesondere ältere Werke im Internet zugänglich gemacht. Das vorgestellte Volltextarchiv wird deshalb zunächst Publikationen bis ca. 1900 enthalten.

Bei der Beantwortung der Frage nach dem Bedarf gehen wir davon aus, dass taxonomische Literatur als der Ausgangspunkt der wissenschaftlichen Auseinandersetzung mit den Taxa niemals veraltet. Vielmehr muss von jedem Systematiker und Taxonom jederzeit auf diese Literatur zurückgegriffen werden können, um den Vergleich der Beschreibung mit ihm vorliegenden Insektentypen zu ermöglichen. Originalbeschreibungen wurden mitunter in nur regional vertriebenen Zeitschriften, in Reisebeschreibungen und selbst in fachfremden Periodika abgedruckt und sind dadurch sehr schwer zugänglich.

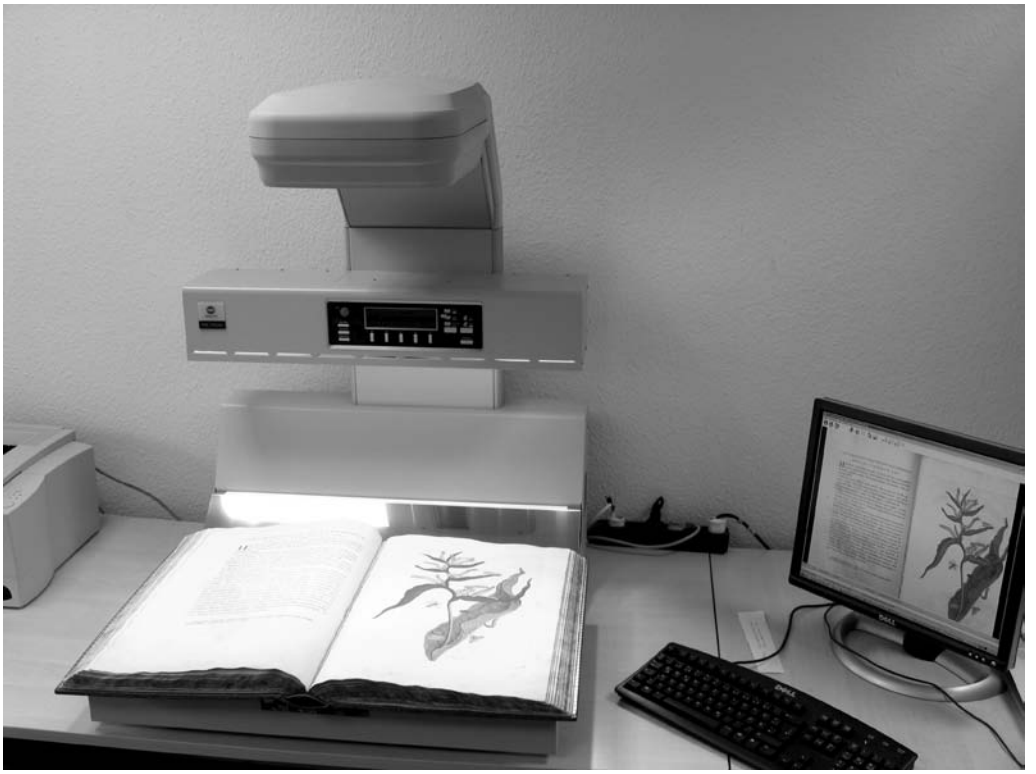


Abb. 1: Buchscanner PS 7000

Nach der Prüfung auf Vollständigkeit und Lesbarkeit werden alle Seiten des Buches oder Artikels entsprechend den Richtlinien der DFG (DFG 2007) gescannt und als einzelne Bilddateien abgespeichert. Die technischen Parameter sind: Auflösung 300 x 300 dpi, Farbtiefe 1 oder 8 Bit (Texte), 24 Bit (Farbbilder), Dateiformat TIF. Bei dem in Abb. 1 dargestellten Gerät handelt es sich um einen Buchscanner PS 7000 der Firma Minolta. Beim Digitalisieren erfolgt das Abtasten der Buchseiten in Aufsicht von oben. Die Bücher liegen normal aufgeschlagen auf einer Buchwippe, die sich der Krümmung der Bindung und dem Höhenunterschied der Buchteile schonend anpasst. Optik und Elektronik des Gerätes entzerren dabei die Seitenwölbung und korrigieren Schatten. Die Belastung des Papiers durch Licht ist gegenüber herkömmlichen Kopierern auf ein Minimum beschränkt. Auch das Umblättern und Festhalten der Buchseiten kann mit geringster mechanischer Belastung geschehen.

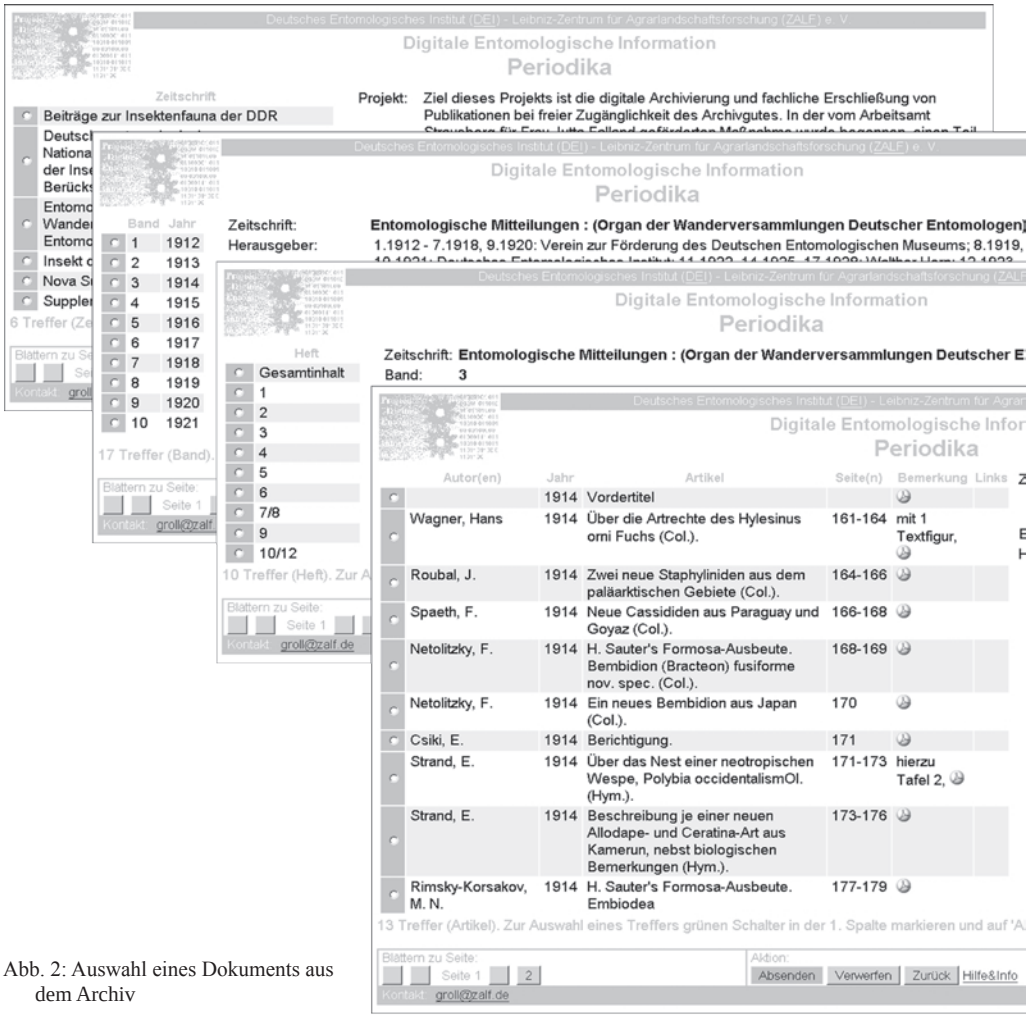


Abb. 2: Auswahl eines Dokuments aus dem Archiv

Im Schritt 3 werden die einzelnen Seiten wieder zu Dokumenten kombiniert, mittels OCR-Software verarbeitet und mit den zugehörigen Metadaten aus der Masterliste versehen. Als Metadaten werden Autor, Jahr, Titel, Zeitschriftentitel, Band, Heft, Seiten, Abbildungen, Verknüpfungen (zugehörige Teile, Fortsetzungen etc.), Bemerkungen, Rechte und Angaben zum Digitalisierungsprozess eingegeben. Die Namen der Autoren stammen hierbei aus der DEI-eigenen Datenbank der Entomologen der Welt. Das erlaubt eine Verbindung der Namen mit den Biographien und Porträts dieser Personen. Auch die Zeitschriftentitel werden in einer separaten Datenbank gepflegt, die gleichzeitig eine Ressource für die Neuauflage des Index I (Taeger, Groll & Gaedike 2007) und die Bestandsmeldungen an die Zeitschriftendatenbank (ZDB) ist. Eine derartige Atomisierung aller Daten und Speicherung in einer relationalen Datenbank lässt schließlich die spätere Generierung von standardisierten XML-Dokumenten, z.B. im Metadaten Encoding and Transmission Standard (METS 2007) zum Austausch und zur Archivierung der Daten zu.

Resultate sind eine Datenbank der Dokumente und zugehörige PDF-Dateien mit Abbild der Seiten und unterliegendem Text (text under image). Mittels der verborgenen Texte kann innerhalb des Dokumentes nach Wörtern gesucht werden. Bereits in diesem Zustand wird das Volltextarchiv im Internet bereitgestellt (Groll 2006).

Weiterhin bilden die unterliegenden Texte den Rohstoff für die fachliche Erschließung der Dokumente. Im 4. Schritt werden daraus einerseits ein Fachwörterbuch generiert und andererseits jedes Dokument mit diesem Fachwortschatz verschlagwortet. Somit ist eine Recherche über alle Dokumente des Volltextarchivs möglich. Anders als in Internet-Suchmaschinen kann jedoch nicht jedes beliebige Wort gesucht werden. Dafür unterstützt das hierarchisch strukturierte Wörterbuch den Nutzer bei der Auswahl der Schlagwörter. Schließlich kann das Wörterbuch die Texterkennung durch die OCR-Software in Schritt 3 optimieren. Während die Digitalisierung derzeit von vielen Bibliotheken und Einrichtungen vorangetrieben wird, kann der entomologische Fachwortschatz nur von Entomologen und Fachinformatoren angelegt und gepflegt werden und passt damit gut in das Forschungsprofil des DEI.

Mehr noch, ein zukünftiges Projekt soll die automatische semantische Erschließung der Dokumente ermöglichen. So wie der Mensch beim Lesen eines Dokuments auf gelerntes Wissen und ggf. auf Nachschlagewerke zurückgreift, muss der Computer bei Such-, Kommunikations- und Entscheidungsaufgaben über eine maschinenlesbare Wissensrepräsentation verfügen können. Dazu ist eine entomologische Wissensbasis mittels Wörterbüchern, Thesauri und Ontologien aufzubauen.

Ergebnisse

Der Aufbau und die Pflege eines Volltextarchivs entomologischer Fachliteratur sind eine dauerhafte Aufgabe und wird sich viel flexibler als ein Archiv gedruckter Materialien an neue technologische Möglichkeiten anpassen lassen. Insofern kann eine solche Datenbank niemals ein fertiges Ergebnis sein. Die bereits digitalisierten Dokumente wurden deshalb von Anfang an in einer frei zugänglichen Applikation im Internet verfügbar gemacht (Groll 2006). Abb. 2 zeigt die schrittweise Auswahl eines Dokuments aus dem Archiv.

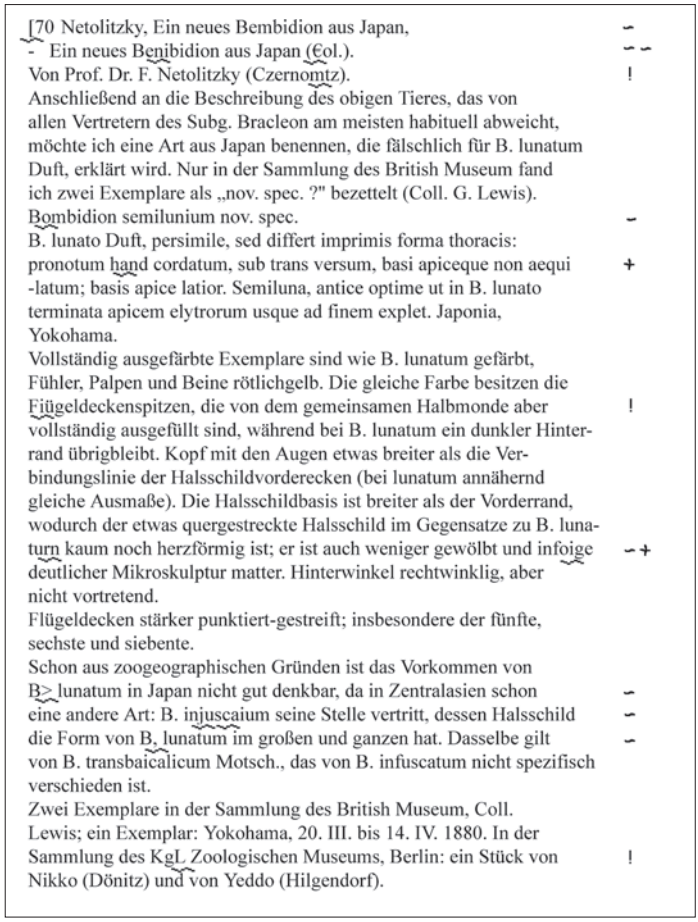
Diskussion

Im Arbeitsablauf treten verschiedene Probleme, wie Erkennungsfehler der OCR-Software oder mangelnde Komponenten, z. B. Wörterbücher auf. Bei genauer Analyse sind sie jedoch nicht technischer Natur, sondern hängen von der Verfügbarkeit digitaler Informationen ab.

So enthält der aus einem gescannten Seitenbild gewonnene Text (Abb. 3) 13 Fehler, was auf die 300 Wörter eine Rate von ca. 4 % ergibt. Unkritisch sind dabei acht mit ~ gekennzeichnete Fehler: [70 = 70 (Seitennummer ist in den Metadaten korrekt), Benbidion, Bombidion = Bembidion (kommt im Text korrekt vor), €ol. = Col. (kann aus Bembidion abgeleitet werden), B. lunaturn, B> lunatum, B, lunatum = B. lunatum (kommt im Text korrekt vor) und injuscatum = infuscatum (kommt im Text korrekt vor). Zwei weitere werden als indifferent eingeschätzt: hand = haud (lat.) und infoige = infolge. Sie könnten bei einer semantischen Bearbeitung der Texte von Bedeutung sein. Drei Fehler (1%) hingegen sind kritisch, weil sie bisher nur durch Nachbearbeitung korrigiert werden können: Czernomtz = Czernowitz (Ort), Fiügeldeckenspitzen = Flügeldeckenspitzen und KgL = Kgl. (Museum).

Einige der kritischen Fehler können durch Integration eines nutzerspezifischen Wörterbuchs in die OCR-Software erkannt und zum Teil automatisch korrigiert werden.

Für den Aufbau eines solchen Wörterbuches findet man schnell Daten auf zahlreichen Webseiten. Aufbau, Syntax und Semantik der Seiten unterscheiden sich jedoch so grundlegend, dass man doch wieder mühsam editieren muss. An Bemühungen, bereits für kommerzielle Bereiche entwickelte Komponenten, z. B. objektorientierte Programmierung, RDF (Herman, Swick & Brickley 2007), XML (eXtended Markup Language), Kommunikationsprotokolle und Datenstandards zum Aufbau einer entomologischen Wissensbasis mangelt es nicht, z. B. ABCD Schema, Darwin Core, DELTA (DEscription Language for Taxonomy) oder Species 2000 Common Data Model (Johnson 2007). Nun gilt es, Anleitungen zur Anwendung solcher Werkzeuge zu verfassen und konsequent maschinenlesbare Ressourcen z. B. Kataloge (Verzeichnis der Taxa), Klassifikationen (System der Insekten) und Ontologien (entomologische Konzepte) zu erstellen.



[70 Netolitzky, Ein neues Bembidion aus Japan, ~
- Ein neues Benibidion aus Japan (Col.). ~ ~
Von Prof. Dr. F. Netolitzky (Czernomtz). !
Anschließend an die Beschreibung des obigen Tieres, das von allen Vertretern des Subg. Bracleon am meisten habituell abweicht, möchte ich eine Art aus Japan benennen, die fälschlich für B. lunatum Duft, erklärt wird. Nur in der Sammlung des British Museum fand ich zwei Exemplare als „nov. spec. ?" bezettelt (Coll. G. Lewis).
Bombidion semilunium nov. spec. ~
B. lunato Duft, persimile, sed differt imprimis forma thoracis: pronotum hand cordatum, sub trans versum, basi apiceque non aequi + -latum; basis apice latior. Semiluna, antice optime ut in B. lunato terminata apicem elytrorum usque ad finem explet. Japonia, Yokohama.
Vollständig ausgefärbte Exemplare sind wie B. lunatum gefärbt, Fühler, Palpen und Beine rötlichgelb. Die gleiche Farbe besitzen die Flügeldeckenspitzen, die von dem gemeinsamen Halbmonde aber ! vollständig ausgefüllt sind, während bei B. lunatum ein dunkler Hinterrand übrigbleibt. Kopf mit den Augen etwas breiter als die Verbindungslinie der Halsschildvorderecken (bei lunatum annähernd gleiche Ausmaße). Die Halsschildbasis ist breiter als der Vorderrand, wodurch der etwas quergestreckte Halsschild im Gegensatze zu B. lunaturn kaum noch herzförmig ist; er ist auch weniger gewölbt und infoige ~ + deutlicher Mikroskulptur matter. Hinterwinkel rechtwinklig, aber nicht vortretend.
Flügeldecken stärker punktiert-gestreift; insbesondere der fünfte, sechste und siebente.
Schon aus zoogeographischen Gründen ist das Vorkommen von B> lunatum in Japan nicht gut denkbar, da in Zentralasien schon ~ eine andere Art: B. injuscaium seine Stelle vertritt, dessen Halsschild ~ die Form von B. lunatum im großen und ganzen hat. Dasselbe gilt ~ von B. transbaicalicum Motsch., das von B. infuscatum nicht spezifisch verschieden ist.
Zwei Exemplare in der Sammlung des British Museum, Coll. Lewis; ein Exemplar: Yokohama, 20. III. bis 14. IV. 1880. In der Sammlung des KgL Zoologischen Museums, Berlin: ein Stück von ! Nikko (Dönitz) und von Yeddo (Hilgendorf).

Abb. 3: Mittels OCR-Software gewonnener Text
Fehlerklassen: ~ unkritisch
+ indif ferent
! kritisch

Literatur

Derksen, W. & Scheiding, U. (1963-1975): Index Litteraturae Entomologicae. Serie II: Die Welt-Literatur über die gesamte Entomologie von 1864 bis 1900. – **I** (1963) A-E, **II** (1965) F-L, **III** (1968) M-R, **IV** (1975) S-Z: I-Xii+697; 678; 528; 482

DFG (2007): Praxisregeln im Förderprogramm „Kulturelle Überlieferung" – DFG-Vordruck 12.151- 3/07 - II 21

Gaedike, R. & Smetana, O. (1978): Ergänzungen und Berichtigungen zu Walter Horn und Sigmund Schenkling: Index Litteraturae Entomologicae, Serie I, die Welt-Literatur über die gesamte Entomologie bis inklusive 1863. Teil I A-K. – Beiträge zur Entomologie, Berlin **28** (2): 329-436

Gaedike, R. & Smetana, O. (1984): Ergänzungen und Berichtigungen zu Walter Horn und Sigmund Schenkling: Index Litteraturae Entomologicae, Serie I, die Weltliteratur über die gesamte Entomologie bis inklusive 1863. Teil II: L-Z. – Beiträge zur Entomologie, Berlin **34** (1): 167-291

Groll, E. K. (2006): Digitale Entomologische Information – Zeitschriften. Datenbank 1. Version, DEI im ZALF e.V. – http://www.zalf.de/home_zalf/institute/dei/php/archiv/journal.php

Groll, E. K. [ed.] (2006): Entomologen der Welt (Biographien, Sammlungsverbleib). Datenbank 2. Version, DEI im ZALF e.V. – http://www.zalf.de/home_zalf/institute/dei/php/biograph/biograph.php

Groll, E. K. & Schubert, E. (2006): Digitale Entomologische Information – Bilder. Datenbank 1. Version, DEI im ZALF e.V. – http://www.zalf.de/home_zalf/institute/dei/php/archiv/bilder.php

Herman, I.; Swick, R. & Brickley, D. (2007): Resource Description Framework (RDF) - http://www.w3.org/RDF/

Horn, W. H. R. & Kahle, I. (1935-1937): Über entomologische Sammlungen, Entomologen und Entomo-Museologie (Ein Beitrag zur Geschichte der Entomologie). Teile I-III. – Entomologische Beihefte aus Berlin-Dahlem, Berlin-Dahlem **2**, **3**, **4**: VI+1-160;161-296; 297-536, Taf. I-XVI; XVII-XXVI; XXVII-XXXVIII

Horn, W. H. R.; Kahle, I.; Friese, G. & Gaedike, R. (1990): Collectiones entomologicae. Ein Kompendium über den Verbleib entomologischer Sammlungen der Welt bis 1960. – Akademie der Landwirtschaftswissenschaften der DDR, Berlin **1**; **2** : 1-220; 221-573, 38 Taf., 125 Photos

Horn, W. H. R. & Schenkling, S. (1928-1929): Index Litteraturae Entomologicae, Serie I: die Welt-Literatur über die gesamte Entomologie bis inklusive 1863. – Berlin-Dahlem, Selbstverlag W. Horn **1-4**: XXI p., 1426 p., 4 Taf.

Johnson, N. F. (2007) Biodiversity Informatics – Annual Review of Entomology **52**: 421-438

METS (2007): Metadaten Encoding and Transmission Standard – http://www.loc.gov/standards/mets/mets-schemadocs.html

Taeger, A. & Blank, S. M. (2006): ECatSym - Elektronischer Katalog der Symphyta (Insecta, Hymenoptera) der Welt. Daten Version 2 (11. August 2006). – Digitale Entomologische Information, Müncheberg, http://www.zalf.de/home_zalf/institute/dei/php_e/ecatsym/ecatsym.php

Taeger, A.; Groll, E. K. & Gaedike, R. (2007): Bibliographie der entomologischen Literatur Serie I: von ihren Anfängen bis 1863 (Index novus litteraturae entomologicae, Serie I: usque ad 1863) – Mitt. Dtsch. Ges. allg. angew. Ent. **16**