21.03.2006 Index Terminorum (ITER)

Er wird gewonnen aus Werken des Thesaurus eruditionis, die lexikalisch angelegt oder durch reichhaltige Register erschlossen sind. (Geeignete Werke aus Hist & Pol können einbezogen werden.) Er enthält Allgemeinbegriffe (nomina appellativa) und Namen (nomina propria). Er setzt sich aus domainspezifischen Listen zusammen. Die Einträge enthalten neben dem Lemma, das auch ein mehrgliedriger Ausdruck sein kann, Herkunftsangaben sowie lexikalische und semantische Daten, die in der Regel aus dem Kontext der Quelle gewonnen werden. Das Datenschema umfasst in der laufenden Arbeitsphase 24 Felder (Kategorien). Später sollen weitere Felder hinzukommen, unter anderem für morphologische Angaben und die Zuordnung zu aktuellen Klassifikationen (DDC) und Schlagwortlisten (SWD, PND).

Folgende Daten werden, soweit automatisch machbar, extrahiert:

LEM = LEMma. Das kann auch ein Ausdruck aus mehreren Wörtern sein. - 70 Z.

IDN = IDentNummer (automatisch vergebene laufende Nummer nach Folge der Erstellung) - bis zu 7 Z. (= Limit 10 Mio.)

URL = URL der Quellenseite (image), der das Lemma entnommen ist. (Obligatorisches Feld) Aus diesem Feld wird in der Anzeige der Kurztitel des betreffenden Werks generiert. (Ich gehe hier davon aus, dass wir dafür kein Datenfeld benötigen.) - 90 Z.

BIB = Ein in der Quelle angeführtes autoritatives Werk (locus classicus, Handbuch u.ä.) [BIBliographic item]: Notiert wird nur der Name des Verfassers (ggf. nur der des Kommentators), bei Anonyma der Sachtitel. - 200 Z.

VAR = Graphische VARiante – wird nur notiert, sofern sie (vermutlich) nicht in den von TERMINI benutzten Perseus-Tools reglat und Morpheus enthalten ist. Triviale Varianten (z.B. coelum und celum zu caelum) werden nicht erfasst. - 100 Z.

ABB = Abkürzung [ABBreviation]. Auch wenn eine Abkürzung als Lemma vorliegt, wird sie in das Feld ABB eingetragen, die Vollform aber als Lemma verwendet. Im Index erscheint dann unter der Abkürzung eine Verweisung auf die Vollform. - 25 Z.

SYN = SYNonym. - 100 Z.

MET = METasemie, d.h. eine Redefigur, die den durch das Lemma direkt bezeichneten Gegenstand indirekt bezeichnet bzw. umschreibt (translatio vel immutatio verborum). - 100 Z.

PHR = PHRase: Stehende Verbindung des Lemmas mit einem Attribut (v.a. Epitheton), Verbum u.ä.; Phraseologismus - 200 Z.

LAT = LATein – wird nur dann belegt, wenn das Lemma nicht die lateinische Sprachform hat. - 100 Z.

GRE = Griechisch (der Antike und des Mittelalters [GREek]) - 100 Z.

ENG = ENGlisch - 100 Z.

GER = Deutsch [GERman] - 100 Z.

FRA = FRAnzösisch - 100 Z.

ITA = ITAlienisch - 100 Z.

SPA = SPAnisch - 100 Z.

DUT = Niederländisch [DUTch] - 100 Z.

LAN = eine oben nicht aufgeführte Sprache [LANguage] - 50 Z.

DEF = DEFinition bzw. Paraphrase bzw. Beschreibung aus Quelle. - 120 Z.

HEA = Überschrift aus Quelle [HEAding] - 100 Z.

BRO = Oberbegriff [BROad term] – ein weiterer Begriff, der mit dem durch das Lemma bezeichneten Gegenstand durch (im begriffslogischen Sinn) "vererbte" Merkmale verbunden ist (z.B. 'Säugetier' zu 'Hund'). Hier können ausnahmsweise auch mehrere Begriffe, die unterschiedlichen Hierarchien entstammen, genannt werden (z.B. Säugetier – Haustier – Nutztier). Zur Vermeidung von Redundanz wird in der Regel nur der hierarchisch nächste Oberbegriff angegeben. - 70 Z.

SIT = Raum, Umgebung, Ganzes (auch Herkunftsort, Wirkungsort, Institution) [SITuated in, lat. SITus, dt. SITz] - 70 Z.

DAT = Zeitspanne (z.B. Lebensdaten) oder Zeitpunkt [DATe(s)] - 30 Z.

VID= Verweisung ['VIDe'] innerhalb der Quelle. Verweisungen können bei unterschiedlichen Relationen stehen: SYN, Äquivalent in anderer Sprache, VAR, BRO, SIT. Es ist daher bei automatischem Verfahren ratsam, eine Verweisung zunächst als solche (VID) zu registrieren. Beim Zieleintrag wird der Verweisungseintrag oft in einem spezifischeren Feld auftreten, so dass im Lauf der Zeit die unspezifische VID-Relation durch eine spezifische Kategorie aus der obigen Auswahl ersetzt werden kann. - 70 Z.

Verfahren:

Umfangreichere Korrekturen sollten vor der Einspeisung in die Datenbank im Grundtext (z.B. XML) vorgenommen werden, da die Rückführung der nachher in der Datenbank vorgenommenen Korrekturen in den Grundtext zu aufwendig ist.

In ein Feld können mehrere Einträge - abgesetzt durch das Zeichen '|' - eingebracht werden, die Felder LEM, IDN, URL und HEA ausgenommen.

Jedes Werk wird in einer eigenen Datenbank bearbeitet. Eine zeitgleiche Bearbeitung durch andere wird vermieden. Will man eine zweite Datenbank öffnen, um ein anderes Werk (z.B. bei der Bearbeitung von Jonston: Historia naturalis den Nomenclator von Junius) zu konsultieren, so darf man sie nicht bearbeiten, ohne sich vorher zu versichern, dass niemand sonst an ihr arbeitet. Hilfskräfte sollten, um mit der Datenbank arbeiten zu können, auf ihrem PC (Laptop) Open Office installieren.

Es geht zunächst darum, möglichst bald einen wesentlichen Ansatz von Termini zu demonstrieren. Daher sollten jetzt nur sichere (d.h. zu mindestens 95% korrekte) Datenextraktionen ausgeführt werden. In einer späteren Phase soll die Extraktion erweitert und mit einem gewissen Arbeitsaufwand verbessert werden. Spielt bei der Extraktion die Großschreibung eine Rolle, so können die Fälle, in denen die Großschreibung allein dem Satzbeginn zuzuschreiben ist, dadurch eliminiert werden, daß hinter dem Punkt am Satzende ein zweites Spatium eingefügt wird. Es genügt, wenn dies zu Beginn der Extraktion im Arbeitstext durchgeführt wird. Im Grundtext ist es nicht erforderlich.

Zuerst wählt WS ein Werk aus, analysiert die Gegebenheiten und macht Vorschläge zur Extraktion. LW prüft das und bringt die Anleitung in eine operative Form. Hilfskräfte führen das aus, wobei u.U. Teams aus je einer latinistischen und computerlinguistischen Kraft gebildet werden.

Format: Schwerpunkt von Termini ist nicht das Markup unserer Texte, sondern die Sammlung strukturierter Wortschatzdaten, mit denen eine rudimentäre Wissensbasis aufgebaut wird. Deren Daten sollen dann helfen, die Suche in (unmarkierten!) Texten zu verbessern. Wir bringen - auf der Basis von Texten und anderen Datenquellen - unsere Daten in die Datenbank ein und bearbeiten sie darin. Von hier können sie auch im XML-Format ausgegeben werden.

Darstellung: ITER wird zunächst die (auf ein einziges Werk bezogenen) Einträge rein additiv in mechanischer Sortierfolge aneinanderreihen. Die Nutzbarmachung von ITER für die Suche bleibt im allgemeinen vorerst dem Nutzer überlassen. Nur im Bereich der Personennamen wird derzeit an einer solchen Nutzbarmachung gearbeitet. Eine Perspektive für eine weitgehende automatische Nutzung der in ITER verfügbaren Daten für Kodierung und Recherche stellt vielleicht der statistische Vergleich von Textsegmenten auf der Basis von Frequenz- und Ähnlichkeitswerten dar.