HANS'2000: Steuerzeichen und Datenmodell

Antrag an die HANS-Nutzergemeinschaft

Thomas Berger

<ThB@gymel.com>

23.2.2003

Inhaltsverzeichnis

1. Antrag

1.1. Vorschlag

HANS'2000 als Kategorienschema bzw. Datenformat ist eine Anwendung von allegro-C und daher gewissen Rahmenbedingungen unterworfen. Darüberhinaus gibt es jedoch zusätzlich definierte Steuerzeichen (für Erfassung und programminternen Gebrauch) und sonstige Konventionen, die im Zusammenhang mit der *Bedeutung* von HANS'2000-Daten zu beachten sind.

Im Folgenden werden die Steuerzeichen und Eingabekonventionen aufgezählt und beschrieben, sofern sie als „Designkriterien" für das HANS-Datenformat von Belang sind.

1.2. Konsequenzen

1. Zukünftige Erweiterungen des HANS-Datenformats sollten konform zu den in diesem Dokument formulierten Kriterien definiert werden.

2. Für die in der aktuellen Version des HANS-Datenformats von den in diesem Dokument formulierten Kriterien abweichenden Erfassungsvorschriften („problematische" Kategorien, vgl. Anhang), sollte mittelfristig eine Umänderung auf eine konforme Erfassungsvariante beantragt und beschlossen werden.

2. Beschreibung der Struktur und Steuerzeichen

2.1. HANS als PC-MAB

Das HANS-Datenformat [1] [8] und seine Vorgängerdokumente *HANSEATICS Nr. 2* vom September 1995 und *HANSEACTICS* vom Winter 1992/93 zeigen, dass HANS'2000 in der Tradition der sogenannten PC-MAB's steht, wie sie um 1990 vielerorts für verschiedene Zwecke entwickelt oder adaptiert wurden. Der Begriff „PC-MAB" steht hier für den Versuch, durch Zitat einiger im MAB-Format definierter Feldnummern folgende Ziele für eine eigene Katalogisierungsanwendung zu erreichen:

1. Nachnutzung des in der Definition des MAB-Formats steckenden, großen intellektuellen Aufwands, ein Katalogregelwerk zu kategorisieren (d.h. Sachverhalte auf Datenfelder ‚abzubilden'). Dies bedeutet zunächst, dass Datenfelder, die für MAB als notwendig erachtet wurden, für die eigene Anwendung nicht verkehrt sein können.

2. Im – evtl. nur scheinbaren – Umkehrschluß auf ein eigenes Katalogisierungsregelwerk zu verzichten, da man Datenfeldnamen benutzt, die Anspielungen auf MAB-Datenfeldnamen sind, wobei die MAB-Datenfelder unter Bezugnahme auf vorhandene Regelwerke auszufüllen wären.

3. Eine – vage – Option auf Interoperabilität bzw. Austauschbarkeit von Daten soll gewährleistet sein.

Umgekehrt sind die PC-MABs nie eine vollständige Umsetzung und meist auch eine extrem starke Abwandlung von MAB, u.a. aus folgenden Gründen:

1. Die Struktur von MAB ist auf Geschäftsgangsparadigmen von Bibliotheksverbünden gegründet (etwa Auskopplung von Signaturen und der meisten Sacherschliessungsdaten in separate Lokal- bzw. Exemplarsätze), die dem beabsichtigten Einsatzzweck (etwa überwiegend Unikate im Nachlassbereich), dem Ausbildungsstand der KatalogisiererInnen und auch den technischen Möglichkeiten (‚nahtlose' Recherche nach ausgelagerten Daten ist ungleich schwieriger) widerspricht.

2. Aufgrund der Entscheidung beim Design von MAB, zusammengehörende, jedoch differenzierte Information nicht in Untefeldern eines Datenfeldes, sondern in aufeinanderfolgenden Feldern mit ‚arithmetischem' Bezug zu erfassen (den sogenannten ‚Periodengruppen'), erfordert Katalogisieren in ‚reinem MAB' unbillige Kopfrechenkünste von den KatalogisiererInnen (und entsprechende Klimmzüge der Software, die ja solcherart verstreute Information wieder zusammensammeln muss).

3. Gewisse, für die konkrete Anwendung benötigte Datenelemente sind bzw. waren in MAB nicht definiert.

4. MAB-Datenfelder sind des öfteren (bei gleicher Feldnummer) wiederholbar und besitzen ‚Indikatoren', die die Ausprägung des konkreten Datenfeldes codieren. Je nach eingesetzter Software oder Zielgruppe liessen sich diese Gestaltungselemente nicht umsetzen.

Für die Ordnung der Datenfelder in einem MAB-Datensatz gilt, dass diese theoretisch beliebig ist, also keine Bedeutung trägt. Die meisten real existierenden Anwendungen (besonders der Bibliotheksverbünde) erwarten jedoch, dass die Felder nach aufsteigenden Feldnummern sortiert sind. Dies ist auch eine recht sinnvolle Ordnung, weil diese Ordnung (sehr grob) die Reihenfolge des Auftretens der jeweiligen Datenelemente in einem klassischen Katalogeintrag widerspiegelt. Für die Ordnung von Feldern mit gleicher Feldnummer gilt, dass die vorhandene Ordnung (demnach die Reihenfolge bei der Erfassung) stets zwingend beizubehalten ist, insbesondere weil es Indikatoren mit der Bedeutung „Das aktuelle Feld ist eine Verweisungsform zum vorangehenden Feld" gibt.

2.2. HANS als Allegro-Anwendung

Für das HANS-Datenformat als ein mit Mitteln von allegro-C realisiertes PC-MAB gelten folgende Eigenschaften:

1. Das *Kategorienschema* ist ein dreistelliges, d.h. Datenfeldelemente (*Kategorien* im Jargon von allegro-C) sind gekennzeichnet durch drei (normalerweise numerische) Zeichen (*Kategorienummer* genannt), gefolgt von Spatium (‚Grundkategorie') oder einem beliebigen Zeichen (‚Fortsetzungs-' oder ‚Wiederholungskategorie').

2. Die Ordnung der Felder ist in der Konfigurationsdatei festgelegt, aufsteigend nach Feldnummer, jedoch ohne tieferen Sinn. Wiederholungskategorien mit gleicher Feldnummer ordnen stets nach dem (üblichen, ASCII-) Sortierwert des Wiederholungsbuchstabens, also stets die Grundkategorie zuerst, dann Wiederholungen mit Ziffern als Folgebuchstaben, dann solche mit Grossbuchstaben, dann solche mit Kleinbuchstaben. Bei der Erfassung schlägt das System üblicherweise den nächsten geeigneten Folgebuchstaben von sich aus vor.

3. Indikatoren gibt es im HANS-Datenformat nicht. Es gibt jedoch viele Situationen, in denen einem bestimmten Folgebuchstaben einer Kategorie eine spezielle Bedeutung zugeordnet ist, dies entspricht jedoch technisch eher einem eigenständigen, »zwischengeschobenen« Datenfeld als einem Indikator.

4. MAB kennt ein *Teilfeldzeichen* (‡), das zur Abteilung etwa von Identnummern oder einleitenden Wendungen innerhalb von Feldern genutzt wird. Außerdem kennt es ein *Unterfeldzeichen* 0x1F (▼, **Strg**--), das stets von einem Buchstaben oder einer Ziffer gefolgt werden muss: Dies entspricht den *Subfields* von MARC, die zur Substrukturierung von Datenfeldern genutzt werden, der hinter dem Unterfeldzeichen folgende Code kennzeichnet die Art / Bedeutung des folgenden Unterfeldes.

   Im allegro-Jargon, dessen sich auch HANS'2000 befleissigt, spricht man von *Teilfeldern* für Subfields, meint damit also *Unterfelder* im MAB-Sprachgebrauch.

   Das HANS-Datenformat sieht Teilfelder vor, das entsprechende Steuerzeichen ist ebenfalls 0x1F (▼, **Strg**--) (gefolgt vom obligatorischen Codezeichen). HANS erlaubt (bzw. erfordert) im Gegensatz zu USMARC *mixed content*: In einer Kategorie wird zuerst der »reguläre« Inhalt erfasst, daran werden Teilfelder angehängt.

5. *Normdatenverknüpfungen* sind ein Konzept von allegro-C, wobei ein in den Daten hinterlegtes Steuerzeichen gefolgt von einer Identnummer dazu führt, dass in allen Situationen anstelle dieser Identnummer eine vom dadurch referenzierten Zielsatz bereitgestellte schematische »Ansetzung« benutzt (etwa angezeigt) wird. Die für HANS'2000 gewählte Ausprägung dieses Mechanismus benutzt einen Unterstrich (_) vor und hinter der Identnummer.

6. *Nichtsortierzeichen* 0xAA (¬, **Alt**-**170**) sind stets paarig einzusetzen und können mehrere Wörter umfassen, jedoch nicht über Normdatenverknüpfungen einkopierte Inhalte (da diese wiederum Text in Nichtsortierzeichen enthalten können und eine Schachtelung von Nichtsortierzeichen in allegro-C nicht möglich ist).

7. Das *Entstoppungszeichen* (**@**) ist unmittelbar vor einem Wort zu erfassen, welches bei Verstichwortung aufgrund der Stoppwortliste nicht indexiert werden würde. Wegen der zunehmenden Wichtigkeit des Zeichens @ und der abnehmenden Wichtigkeit von Stoppwortlisten ist die Nutzung des Entstoppungszeichens für allegro-C-Anwendungen generell nicht empfehlenswert.

8. Die aus manchen MAB-Anwendungen bekannte *Stichwortkennung* (in ansonsten nicht zu indexierenden Datenfeldern werden zu indexierende Teilstrings durch geschweifte Klammern {...} als dennoch zu indexieren ausgezeichnet) hat keine Entsprechung in HANS'2000.

9. Die in allegro-C vorhandene Möglichkeit, bis zu 6 Ebenen hierarchisch strukturierter Untersätze innerhalb eines Datensatzes zu erfassen, wird von HANS'2000 nicht genutzt.

2.3. Genuine HANS-Festlegungen

Über die von allegro-C angebotenen Strukturierungsmethoden und Steuerzeichen hinaus benutzt HANS'2000 noch folgende Elemente:

1. Das Zeichen 0x10 (▶, **Strg**-**P**) dient als *internes Wiederholungszeichen*, um innerhalb einer (Wiederholungs-)kategorie gleichwertige Inhalte aneinanderzureihen.

2. Über Normdatenverknüpfung eingeblendete Inhalte enthalten am Anfang des Ersetzungstextes oft noch einmal die eigene Identnummer, wie bei einer Verknüpfung durch Unterstriche (_) umschlossen, danach enthalten sie in geschweifte Klammern {...} eingeschlossen die »Ansetzung«, die wiederum Teilfelder enthalten kann. Diese sind jedoch in ein »aufrechtes Dreieck« 0x1E (▲) umgewandelt, um die Teilfeldstruktur der aufnehmenden Kategorie nicht zu beeinflussen. In der Aufnehmenden Kategorie kann man wiederum unmittelbar an die Notation _..._ der Verknüpfung einen eigenen Text in {...} anhängen, dieser überblendet dann wiederum die eingeblendete Ansetzung.

3. Das Teilfeld ▼b hat oftmals die stereotype Bedeutung einer *Einleitenden Wendung* zur gegebenen Kategorie, ist als Text also als vor dem »regulären« Kategorieinhalt aufzufassen, unabhängig von seinem konkreten Auftreten in der Kategorie (stets nach dem »regulären« Kategorieinhalt).

4. Die Zeichenfolge Spatium, Gleichheitszeichen, Spatium (␣=␣, manchmal auch nur Gleichheitszeichen ohne Spatien) dient zur Abteilung einer (sortierfähigen) Indexform und der (anzuzeigenden) Vorlageform innerhalb einer Kategorie. Eine der beiden Formen darf im Allgemeinen auch fehlen, aus technischen Gründen entfallen dann auch die Spatien an dieser Seite des Gleichheitszeichens.

5. Gewisse Kategorien bzw. Teilfelder werden mit *codierten Inhalten* belegt. Oftmals gibt es dabei dann einen speziellen Code z mit der Bedeutung ‚uncodiert, Bedeutung folgt im Klartext'.

Folgende Zeichen dienen zur rudimentären Formatierung und tragen keine weitere Bedeutung (im Sinne einer Anwendungssteuerung). Von ihrem Einsatz wird abgeraten, da bibliothekarische Daten nach allgemeiner Auffassung keine Formatierungsinformationen enthalten sollen und die Exportierbarkeit in andere Anwendungen daher auch langfristig nicht gegeben sein wird:

1. Das *Nichtbrechende Leerzeichen* 0x0F (⍟, **Alt**-**15**) verhindert einen Zeilenumbruch an dieser Stelle, auch lassen sich davon mehrere aneinanderreihen. Die Bedeutung ist ansonsten die eines Leerzeichens. (Gewisse Kategorien, z.B. Signaturen, werden von HANS'2000 automatisch so behandelt, als enthielten sie nur nichtbrechende Leerzeichen).

2. Das (englische) Paragraphenzeichen 0x14 (¶, **Strg**-**T**) dient zur Erfassung / Erzwingung von Zeilenumbrüchen innerhalb der Darstellung einer Kategorie.

2.4. Akzidenzien

Folgende Festlegungen des HANS-Datenformats scheinen uneinheitliche adhoc-Notationen zu benutzen:

Kategorie `#370z`	Zeilenwechsel der Vorlage werden als 0x10 (▶, **Strg**-**P**) erfasst, bei der Ausgabe entsprechend der diplomatischen Konvention als `|` wiedergegeben.
Kategorie `#403r`	Die Kategorie enthält Freitext oder den festen Code `r`, optional hinter einem (mißbräuchlich genutzten) Teilfeldzeichen 0x1F (▼, **Strg**--) weitere Angaben.
Kategorie `#524`	Die Kategorie enthält Freitext mit der Bedeutung »Darin:«, optional kann als erstes Zeichen des Kategorieinhalts ein Code angegeben werden, der vom (mißbräuchlich genutzten) Teilfeldzeichen 0x1F (▼, **Strg**--) abgeschlossen wird, um ihn als Code zu kennzeichnen.
Kategorie `#806c`	Die Kategorie enthält als erstes Zeichen einen Codebuchstaben, gefolgt von Gleichheitszeitszeichen `=`. Danach dann eine oder durch 0x10 (▶, **Strg**-**P**) gegliedert mehrere Namensformen.

3. Neue Festlegungen

3.1. Interne und externe Wiederholungen

Gewisse Kateorien von HANS'2000 sind sowohl extern (durch Wiederholungszeichen an der Kategorienummer) als auch intern (durch Erfassung des Trennzeichens 0x10 (▶, **Strg**-**P**) wiederholbar. In diesem Fall kann die interne Wiederholung nicht stärker gliedern als die externe. Zu beachten ist jedoch, dass die externe Wiederholung nicht zwingend stärker gliedert als eine interne. Konkret hängt dies von der Parametrierung im jeweiligen Fall ab, die von mindestens den folgenden Faktoren beeinflusst wird:

1. ob eine externe Wiederholung in der Ausgabe ein anderes, »stärkeres« Trennzeichen bewirkt als die interne.
2. ob eine externe Wiederholung die Ausgabe einer ggfls. für die Kategorie definierten automatischen einleitenden Wendung erzwingt.
3. ob das Auftreten einer (automatischen oder im Teilfeld ▼`b` explizit angegebenen) einleitenden Wendung in der Ausgabe ein anderes, »stärkeres« Trennzeichen bewirkt.

Generell sind für bedeutungswandelnde, indikatorartige Folgezeichen nur Kleinbuchstaben zulässig (Bsp. `#540h` für die ISBN.

In dem Fall, wo eine Kategorie sowohl (durch Wiederholungszeichen an der Kategorienummer) extern wiederholbar ist als auch Folgezeichen für eine Bestimmte Bedeutungsabwandlung der Kategorie definiert sind, sind die Wiederholungsfolgezeichen auf die Großbuchstaben `A` bis `Z` beschränkt.

3.2. Reihenfolge der Untergliederung

Klärungsbedarf besteht, wenn mehrere der beschriebenen Steuerzeichen innerhalb *einer* Kategorie eingesetzt werden. Hierfür wird nun folgende Reihenfolge festgelegt, die am wenigsten Konflikte mit der aktuellen Version des HANS-Datenformats erwarten lässt:

1. Die interne Kategoriewiederholung durch 0x10 (►, **Strg**-**P**) gliedert am stärksten.

   Dies bedeutet beispielsweise, dass bei der Arbeit mit Teilfeldern innerhalb einer Kategorie das Zeichen 0x10 (►, **Strg**-**P**) *nicht* zur Aufzählung gleichwertiger Inhalte *innerhalb* eines gegebenen Teilfelds genutzt werden kann, weil sein Auftreten die erste interne Kategoriewiederholung (mit dem Teilfeld) beendet und eine weitere interne Kategoriewiederholung (ohne Teilfeld) startet.

2. Interne Kategoriewiederholungen (bzw. Kategorien, falls diese nicht intern wiederholt sind) werden durch Teilfelder untergliedert.

   Dies bedeutet etwa, dass ein Gleichheitszeichen in einem Teilfeld *dieses* konkrete Teilfeld in Ansetzungs- und Vorlageform aufteilt. Oder etwa, dass die auf einen »regulären« Kategorieinhalt mit Gleichheitszeichen folgenden Teilfelder sich auf die gesamte Kategorie (bzw. aktuelle interne Wiederholung) beziehen, nicht nur auf die »Vorlageform« als Textbestandteil unmittelbar vor den Teilfeldern.

   Eine scheinbare *Ausnahme* stellen einleitende Wendungen aus Teilfeldern ▼`b` dar, die strenggenommen nur für die aktuelle interne Wiederholung gültig sind. Aufgrund der Art der Darstellung bei Anzeige und Druck suggerieren diese Wendungen jedoch, dass sie auch für darauffolgende Wiederholungen (bis etwa zum Auftreten der nächsten Wendung) zuständig sind. (HANS'2000 verstärkt die Suggestion, indem es nicht für jede interne oder externe Wiederholung stets erneut die Standardwendung für die jeweilige Kategorie einblendet). Dieser »Vererbungseffekt« von Wendungen auf nachfolgende Kategoriewiederholungen ohne Wendung kann bei Exporten nicht zuverlässig abgebildet werden, der resultierende Bedeutungsverlust kann evtl. durch konsequente Erfassung abgemildert werden, etwa indem man Teilfelder ▼`b` fallweise entweder in allen internen Wiederholungen einer Kategorie einsetzt oder nur in der ersten (vorausgesetzt, es handelt sich um eine Kategorie, die auch externe Wiederholungen zulässt).

3. Teilfelder bzw. der »reguläre« Inhalt von internen Kategoriewiederholungen bzw. Kategorien werden manchmal durch die Gleichheitszeichen-Notation in Ansetzungs- und Vorlageform aufgeteilt.

4. Normdatenverknüpfungen sind frei darin, über die »Ansetzung« eine Substrukturierung durch mit `=` gegliederte Ersetzungstexte in die beherbergende Kategorie einzublenden. Das Einblenden von internen Kategoriewiederholungszeichen ist jedoch verboten, Teilfelder werden automatisch in das »Einblendungs-Teilfeldzeichen« 0x1E (▲) umgewandelt, solche Teilfelder gelten nur innerhalb eingeblendeten Ersetzungstexts.

Anhang A. Mit den skizzierten Regeln in Konflikt stehende Kategorien

Die folgenden Festlegungen des HANS-Datenformats enstprechen nicht der oben entwickelten Struktur:

Kategorie `#004`, Kategorie `#007` Die von allegro-C automatisch an die Kategorien angehängte Bearbeiterkennzeichnung ist durch ein einzelnes Teilfeldzeichen 0x1F (▼, **Strg**--) an den Datumsstempel angehängt.

Kategorie `#025s` Das interne Wiederholungszeichen 0x10 (▶, **Strg**-**P**) dient dazu, innerhalb der einzelnen Teilfelder die Codes wiederholbar zu halten.

Kategorie `#054l` Ein Datum wird an den Freitext angehängt, getrennt durch ein einzelnes Teilfeldzeichen 0x1F (▼, **Strg**--).

Kategorie `#054w` Datum und weitere Informationen werden durch ein einzelnes Teilfeldzeichen 0x1F (▼, **Strg**--) getrennt an die Wertangabe angehängt.

Kategorie `#080ff` Das Teilfeldzeichen 0x1F (▼, **Strg**--) ohne Folgebuchstaben dient zur Gliederung des Kategorieinhalts in Hierarchiestufen

Kategorie `#403r` Die Kategorie enthält Freitext oder den festen Code `r`, optional hinter einem Teilfeldzeichen 0x1F (▼, **Strg**--) weitere Angaben.

Kategorie `#410`, Kategorie `#410d` Hinter der Angabe der Vorlage- bzw. Druckform für den Ort ist im »regulären« Kategorieinhalt hinter `␣:␣` noch weitere Information (Verlag bzw. Drucker) erfasst.

Die Jahresangabe wird durch ein einzelnes Teilfeldzeichen 0x1F (▼, **Strg**--) getrennt an den Ort etc. angehängt.

Kategorie `#410f`, Kategorie `#410g` Die Jahresangabe wird durch ein einzelnes Teilfeldzeichen 0x1F (▼, **Strg**--) getrennt an den Ort angehängt.

Kategorie `#516ff` Die Kategorien als Ganzes werden um ein (leeres) Teilfeld ▼`p` (polyglott) ergänzt, wenn mehr als drei interne Wiederholungen hätten erfasst werden müssen.

Kategorie `#523` (Satzart `qq`) Die Kategorie enthält einen Code, danach optional und durch ein einzelnes Teilfeldzeichen 0x1F (▼, **Strg**--) abgetrennt Freitextinformation.

Kategorie `#524` Die Kategorie enthält Freitext mit der Bedeutung »Darin:«, optional kann als erstes Zeichen des Kategorieinhalts ein Code angegeben werden, der vom einem Teilfeldzeichen 0x1F (▼, **Strg**--) abgeschlossen wird, um ihn als Code zu kennzeichnen.

Kategorie `#70C` (Satzart `qc`) Die Kategorie enthält einen Code, danach durch ein einzelnes Teilfeldzeichen 0x1F (▼, **Strg**--) abgetrennt Freitextinformation.

Kategorie `#800z` (Satzarten `p` und `k`) An den »regulären« Kategorieinhalt anschliessend wird durch ein einzelnes Teilfeldzeichen 0x1F (▼, **Strg**--) abgetrennt Freitextinformation angegeben, die den »Code« `z` im Kategoriefolgebuchstaben erläutert.

Kategorie `#806c` Die Kategorie enthält als erstes Zeichen einen Codebuchstaben, gefolgt von Gleichheitszeitszeichen `=`. Danach dann eine oder durch 0x10 (▶, **Strg**-**P**) gegliedert mehrere Namensformen.

Kategorie `#808`, Kategorie `#809` Die Kategorie enthalten ein Datum, dann optional ein ein einzelnes Teilfeldzeichen 0x1F (▼, **Strg**-**-**) gefolgt von einem Codebuchstaben, dann erst (optional) Gleichheitszeitszeichen = und die Druckform des Datums. Danach dann optional das »echte« Teilfeld ▼a.

Kategorie `#860ff` In der Zeittafel werden Zeitangaben und zugehörige Texte durch Teilfeldzeichen 0x1F (▼, **Strg**-**-**) getrennt.

Quellen

[1] Harald Weigel. Thomas Berger. *HANS : Datenformat. HANS-Datenformat*. 1996-.

Versionsgeschichte

Version [1]	Stand: 26.8.1996
Version [. . .]	1998/99
Version [5]	19.1.2000

Letzte Überarbeitung

Online unter ftp://ftp.sub.uni-hamburg.de/pub/hans/misc/doku/hnsdform.pdf.