



Terminologie- austausch

Deutscher Terminologietag 2008

Angelika Zerfass, zerfass@zaac.de

Was bisher geschah...

- Seit Mitte der 60er Jahre des 20. Jahrhunderts wird immer wieder an Standardformaten für den Austausch terminologischer Daten gearbeitet
 - MATER (Magnetic Tape Exchange for Terminolgical Records - Großrechnersysteme)
 - MicroMATER (mit dem Aufkommen von PCs)
 - 1987, TEI (Text Encoding Initiative) für die Auszeichnung literatur- und geisteswissenschaftlicher Texte
 - Designer von MicroMATER und TEI tun sich zusammen, um einen neuen Standard zu schaffen, der zu ISO Norm werden sollte (ISO 12200:1999 MARTIF – Machine-readable Terminology Interchange Format)

Was jetzt geschieht...

- MARTIF
 - bietet den „negotiated interchange“, Daten können also nur dann ausgetauscht werden, wenn sich die beiden Beteiligten verständigen
 - Z.B. über den Inhalt bestimmter Datenkategorien
 - DB 1: Genus = mask./fem./neutr. DB 2: Genus = m/f/n
 - MARTIF basiert auf SGML und kann nicht alle Sprachen verarbeiten (vor allem lateinisch basierte und weitere Sprachen nur mit zusätzlichem Aufwand)
- Erhofft wird der „blind interchange“, also ein Austausch von Daten aus unterschiedlich strukturierten Datenbanken, bei dem keine Absprache stattfinden muss.
- Umstellung auf XML (um einen Austausch im Internet zu ermöglichen)
- Definition fester Inhalte für Datenkategorien
- Unterstützung für mehr Zeichensätze
- TBX - TermBase eXchange (ISO 30042 – in Arbeit)

OLIF

- Open Lexicon Interchange Format (OLIF) wurde von Nutzern und Herstellern von MÜ Systemen entwickelt (im Rahmen eines Projektes der Europäischen Kommission, OTELO)
- Die aktuelle Version wird vom OLIF Consortium weiterentwickelt
- Austausch von lexikographischen und terminologischen Daten zwischen NLP (natural language processing) Systemen
- Konvertierungsmechanismen für OLIF/TBX sind angedacht

TBX (TermBase Exchange)

- TBX entstand in der OSCAR Gruppe der LISA. Wurde im SALT Projekt weiterentwickelt
- TBX
 - Terminological Markup Language (TML)
 - Basiert auf den ISO Standards für Datenkategorien (12620), Terminological Markup Framework (16642) und Martif 12200

Aus der Spezifikation...

- TBX is an open XML-based standard format for terminological data
- TBX is designed to support the analysis, representation, dissemination, and exchange of information from human-oriented terminological databases (term bases)

Von der LISA website

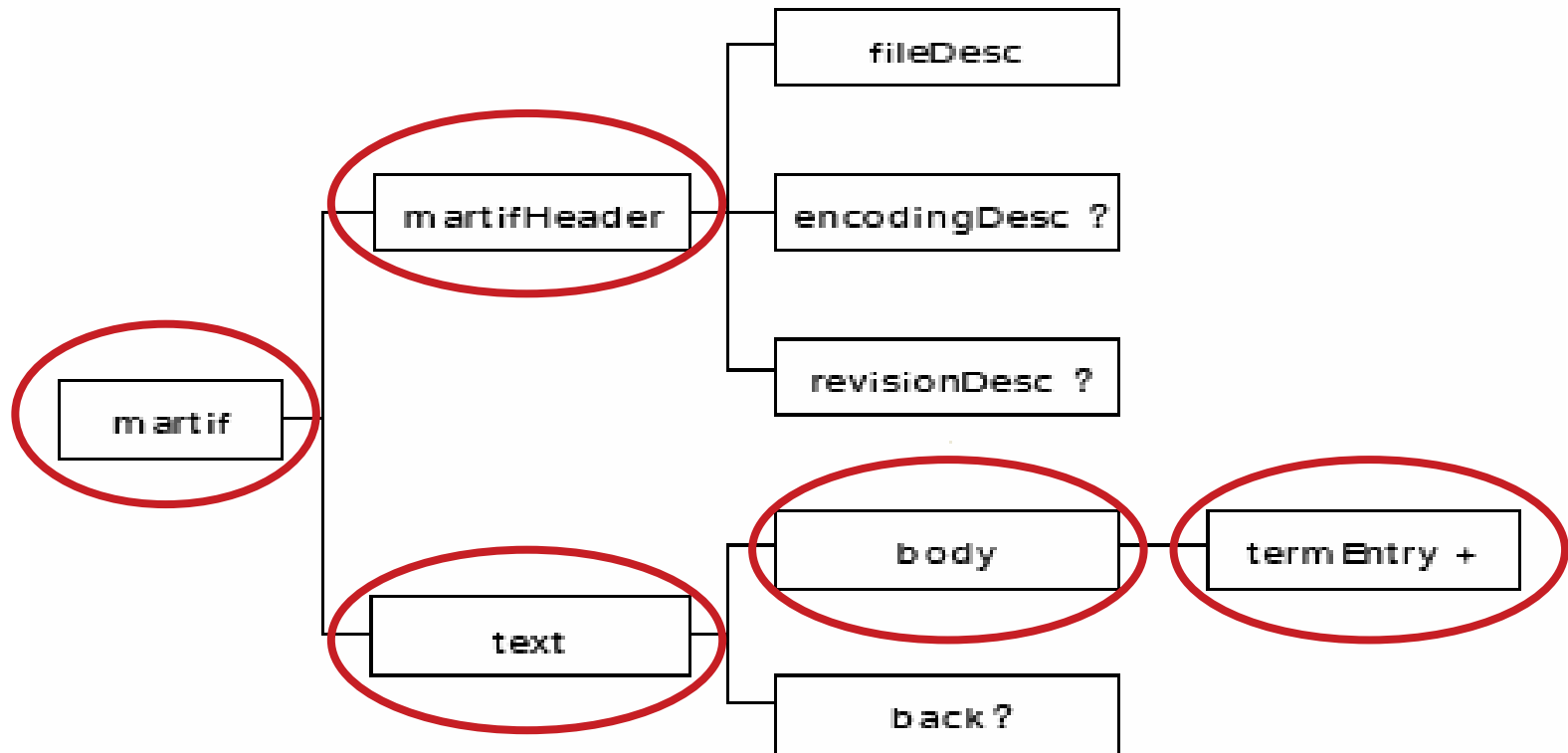
TBX offers users the following advantages:

- **Better control**, both over the language that represents your company in and over your brand representation.
- **Improved quality**. Controlling terminology means that localized text is more likely to represent what you want it to and helps localizers maintain consistency and quality.
- **Reduced localization cost and faster time to market**. Sharing your terminology data with service providers helps them improve accuracy, reducing time spent in revisions and in terminology research.
- **Freedom to use the right linguistic tools**. By using open standards like TBX you are not limited to any one tool that may not be suitable for a specific project.

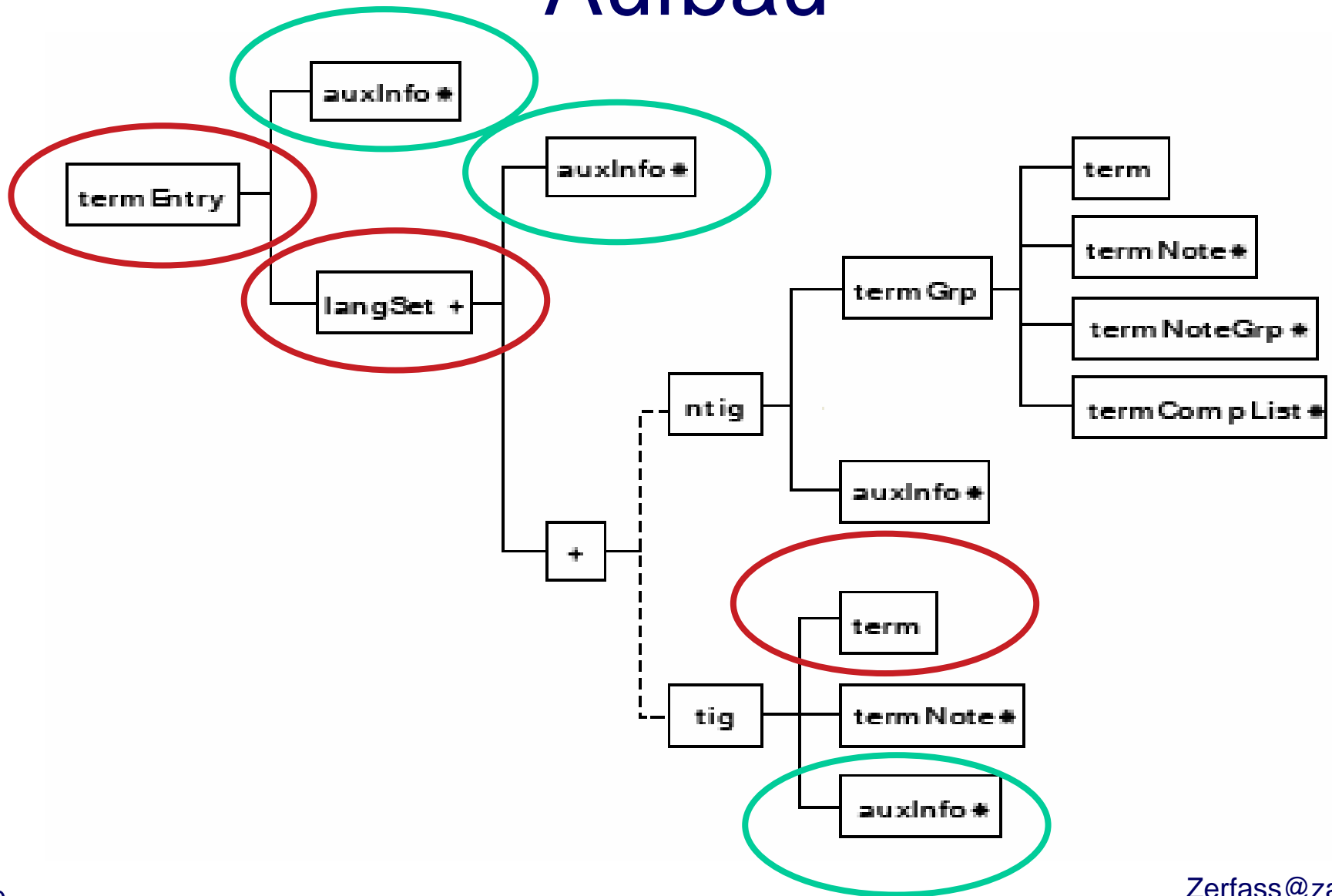
TBX Basic

- TBX-Basic is a standard under development by the OSCAR Steering Committee of LISA. TBX-Basic is intended to be a lighter version of TBX, particularly suited to small or medium sized language industries. While the primary audience is localization service providers (LSPs), the format is also suited for any language application that requires a lightweight approach to terminology management, such as controlled authoring and content management.

Aufbau



Aufbau



Multilinguale TMX Datei

```
<tu  tuid="1" datatype="Text" srclang="en-us">  
  <tuv xml:lang="en-us">  
    <seg>This is a test.</seg>  
  </tuv>  
  <tuv xml:lang="de">  
    <seg>Dies ist ein Test.</seg>  
  </tuv>  
  <tuv xml:lang="ja">  
    <seg>テストです。 </seg>  
  </tuv>  
  <tuv xml:lang="zh-cn">  
    <seg>这是试验。 </seg>  
  </tuv>  
</tu>
```

```

<tu tuid="1" datatype="Text" srclang="en-us">
  <tuv xml:lang="en-us">
    <seg>This is a test.</seg>
  </tuv>
  <tuv xml:lang="de">
    <seg>Dies ist ein Test.</seg>
  </tuv>
  <tuv xml:lang="ja">
    <seg>テストです。</seg>
  </tuv>
  <tuv xml:lang="zh-cn">
    <seg>这是试验。</seg>
  </tuv>
</tu>

```

TMX

```

- <termEntry id="c2">
  - <descrip type="subjectField">
    Hardware \ Other Processing Units and Specialized Devices
  </descrip>
  <descrip type="relatedConceptBroader">acceptor</descrip>
  - <langSet xml:lang="en">
    <admin type="productSubset">Retail Store Solutions</admin>
    - <adminGrp>
      <admin type="sourceIdentifier">Translation Services Center</admin>
    </adminGrp>
    - <ntig>
      - <termGrp>
        <term>bill acceptor</term>
        <termNote type="partOfSpeech">noun</termNote>
      </termGrp>
      - <descrip type="context">
        Accepts bill denominations of $1, $2, $5, $10, $20, $50 and $100
        holds 600 bills. It detects and rejects counterfeit bills.
      </descrip>
    </ntig>
  </langSet>
  - <langSet xml:lang="fr">
    <admin type="productSubset">Retail Store Solutions</admin>
    - <ntig>
      - <termGrp>
        <term>accepteur de billets</term>
        <termNote type="partOfSpeech">nom</termNote>
      </termGrp>
    </ntig>
  </langSet>
</termEntry>

```

TBX

Globale Informationen
im Eintragskopf

Sprachzuordnung

Administrative Daten
einer Sprache

Benennung Englisch

Informationen
auf Benennungsebene

Sprachzuordnung

Benennung Französisch

```
- <termEntry id="c2">
```

```
- <descrip type="subjectField">
```

```
Hardware \ Other Processing Units and Specialized Devices
```

```
</descrip>
```

```
<descrip type="relatedConceptBroader">acceptor</descrip>
```

```
- <langSet xml:lang="en">
```

```
<admin type="productSubset">Retail Store Solutions</admin>
```

```
- <adminGrp>
```

```
<admin type="sourceIdentifier">Translation Services Center</admin>
```

```
</adminGrp>
```

```
- <ntig>
```

```
- <termGrp>
```

```
<term>bill acceptor</term>
```

```
<termNote type="partOfSpeech">noun</termNote>
```

```
</termGrp>
```

```
- <descrip type="context">
```

```
Accepts bill denominations of $1, $2, $5, $10, $20, $50 and $100. The bill acceptor  
holds 600 bills. It detects and rejects counterfeit bills.
```

```
</descrip>
```

```
</ntig>
```

```
</langSet>
```

```
- <langSet xml:lang="fr">
```

```
<admin type="productSubset">Retail Store Solutions</admin>
```

```
- <ntig>
```

```
- <termGrp>
```

```
<term>accepteur de billets</term>
```

```
<termNote type="partOfSpeech">nom</termNote>
```

```
</termGrp>
```

```
</ntig>
```

```
</langSet>
```

```
</termEntry>
```

Konvertierung aus MultiTerm nach TBX (Beispiel Medtronic)

```
- <conceptGrp>
  <concept>7331</concept>
  <system type="entryClass">0</system>
- <transacGrp>
  <transac type="origination">local</transac>
  <date>2005-03-10T14:45:27</date>
</transacGrp>
- <transacGrp>
  <transac type="modification">UMuegge</transac>
  <date>2005-04-22T14:06:32</date>
</transacGrp>
- <descripGrp>
  <descrip type="EntryStatus">approved</descrip>
</descripGrp>
- <descripGrp>
  <descrip type="BusUnit">CRM</descrip>
</descripGrp>
- <descripGrp>
  <descrip type="Product">Concerto/Virtuoso</descrip>
</descripGrp>
- <descripGrp>
  <descrip type="Project">Concerto/Virtuoso</descrip>
</descripGrp>
- <descripGrp>
  <descrip type="Definition">the capability of a distance telemetry enabled programmer
  suitable implanted device who are within range of the programmer</descrip>
</descripGrp>
- <languageGrp>
  <language type="English" lang="EN-US" />
- <termGrp>
  <term>patient identifier</term>
  - <descripGrp>
    <descrip type="PartOfSpeech">noun</descrip>
  </descripGrp>
```

```
- <body>
  - <termEntry id="c7331">
    - <transacGrp>
      <transac
        type="terminologyManagementTransactions">modification</transac>
      <transacNote type="responsibility">UMuegge</transacNote>
      <date>2005-04-22T14:06:32</date>
    </transacGrp>
    - <transacGrp>
      <transac
        type="terminologyManagementTransactions">origination</transac>
      <transacNote type="responsibility">local</transacNote>
      <date>2005-03-10T14:45:27</date>
    </transacGrp>
    <admin type="businessUnitSubset">CRM</admin>
    <admin type="productSubset">Concerto/Virtuoso</admin>
    <admin type="projectSubset">Concerto/Virtuoso</admin>
    <descrip type="definition">the capability of a distance telemetry
    programmer to remotely identify one or more patients with a
    implanted device who are within range of the programmer</descrip>
    - <langSet xml:lang="EN-US">
      - <ntig>
```

Medtronic Mapping Tabel



Medtronic Data Mapping

Medtronic	TBX
<code><conceptGrp></code>	<code><termEntry></code>
<code><transacGrp></code> <code><transac type="origination" >/transac></code> <code><date></date></code> <code></transacGrp></code>	<code><transacGrp></code> <code><transac type="origination"></transac></code> <code><date></date></code> <code></transacGrp></code>
<code><transacGrp></code> <code><transac type="modification"></transac></code> <code><date></date></code> <code></transacGrp></code>	<code><transacGrp></code> <code><transac type="modification" >/transac></code> <code><date></date></code> <code></transacGrp></code>
<code><descrip type="EntryStatus">approved</descrip></code>	<code><termNote type=" administrativeStatus">approved</termNote></code>
<code><descripGrp></code> <code><descrip type="Security">Public</descrip></code> <code></descripGrp></code>	<code><adminGrp></code> <code><admin type="securitySubset">Public</admin></code> <code></adminGrp></code>
<code><descripGrp></code> <code><descrip type="BusUnit">CRM</descrip></code> <code></descripGrp></code>	<code><adminGrp></code> <code><admin type="businessUnitSubset">CRM</admin></code> <code></adminGrp></code>

Anwendungsmöglichkeiten

- Prüfungen in der Ausgangssprache
- Verwendung (TM Systeme) und Prüfung der Verwendung in der Zielsprache
- Stoppwortliste in Terminologieextraktionsprogrammen
- Lexikon für die Analyse in maschinellen Übersetzungssystemen und linguistischen Extraktionsprogrammen
- Indizierung von Dokumentenbeständen / Inhalte von CMS für Wissensdatenbanken
- Austausch von terminologischen Daten zwischen verschiedenen Terminologiedatenbanksystemen
- Darstellung der Terminologie im Internet
- Optimierung von Suchmaschinen und Text Mining (z.B. als Schlüsselwörter, vor allem bei Synonymen)

Case study – IBM

- Authoring in XML (DITA)
- Granular identification of strings in the source
- Term checker
- IBM terminology in spell checker
- Term extraction and pre-translation
- Centralized terminology services
- Multilingual, multipurpose database
- Terminology plugged into search engine

Informationen zu TBX

- ◉ www.lisa.org/tbx
- ◉ <http://www.mith2.umd.edu/thes/ytbx.html>



z a a c

Angelika Zerfaß

Fragen?