Bundesverwaltungsamt

IsyFact-Standard

Konzept
Umgang mit Sonderzeichen

Version 1.11
27.03.2015

IsyFact Standards

Bundesverwaltungsamt
Der zentrale Dienstleister des Bundes

„“ des Bundesverwaltungsamts ist lizenziert unter einer Creative Commons Namensnennung 4.0 International Lizenz.

„Konzept Umgang mit Sonderzeichen“
des Bundesverwaltungsamts ist lizenziert unter einer
Creative Commons Namensnennung 4.0 International Lizenz.

Die Lizenzbestimmungen können unter folgender URL heruntergeladen werden: http://creativecommons.org/licenses/by/4.0

Ansprechpartner:

Referat Z II 2
Bundesverwaltungsamt
E-Mail: isyfact@bva.bund.de
Internet: www.isyfact.de

„“ des Bundesverwaltungsamts ist lizenziert unter einer Creative Commons Namensnennung 4.0 International Lizenz.

Dokumentinformationen

Dokumenten-ID:	Konzept_Umgang_mit_Sonderzeichen.docx

Java Bibliothek / IT-System

Name	Art	Version
isy-sonderzeichen	Bibliothek	siehe isyfact-bom v1.3.6

„“ des Bundesverwaltungsamts ist lizenziert unter einer Creative Commons Namensnennung 4.0 International Lizenz.

Inhaltsverzeichnis

1. Einleitung 6

2. Überblick 7

2.1. Aufbau und Zweck des Dokuments 7

2.2. Anforderungen und Randbedingungen 7

2.3. Festlegung des Zeichensatzes und der Codierung 7

3. Konfigurationseinstellungen für den Zeichensatz 8

3.1. Betriebssystem 8

3.2. Oracle Datenbank 8

3.3. Hibernate 8

3.4. Java 9

3.5. Java Property-Dateien 9

3.6. Maven 9

3.7. XML 10

3.8. HTML 10

4. Transformation von Sonderzeichen 12

4.1. Transkription 12

4.1.1 Zeichensätze und Sprachen 12

4.1.2 Anwendungsbereiche in einer IsyFact-Systemlandschaft 12

4.1.3 Transkriptionsregeln 13

4.1.4 Umsetzung im System 13

4.2. Umcodierung 15

4.3. Filtern von Zeichen 15

4.4. Spezifikation von fachlichen Datentypen 15

5. Bibliothek „isy-sonderzeichen“ 16

5.1. Funktionsweise 16

5.2. Einbindung der Bibliothek in eine Anwendung 16

5.2.1 Integration mit Maven 16

5.2.2 Instanziierung der Transformator Factory 17

5.3. Schnittstellendefinition 17

„“ des Bundesverwaltungsamts ist lizenziert unter einer Creative Commons Namensnennung 4.0 International Lizenz.

6. Zulässige Zeichen innerhalb der IsyFact.....20

6.1. Standardzeichen.....20

6.2. Zusätzliche Zeichen.....20

7. Quellenverzeichnis.....35

8. Abbildungsverzeichnis 36

9. Tabellenverzeichnis.....37

„“ des Bundesverwaltungsamts ist lizenziert unter einer Creative Commons Namensnennung 4.0 International Lizenz.

1. Einleitung

In Anwendungen nach IsyFact-Standard werden Daten in internationaler Schreibweise erfasst und gespeichert. Dies sind z.B. die Namen von Personen und Orten. Während die Eingabe von internationalen Sonderzeichen über eine deutsche Tastatur nur einen beschränkten Umfang an Sonderzeichen erlaubt, werden möglicherweise auch Daten abgelegt, die international landesspezifisch erfasst wurden. Hierin können alle landestypischen Sonderzeichen enthalten sein. Insbesondere also auch solche, die nicht über eine deutsche Tastatur eingegeben werden können.

Eine auf dem IsyFact-Standard aufsetzende Architektur umfasst mehrere technische Systeme in einer Umgebung. Für jedes dieser Systeme muss sichergestellt werden, dass die benötigten Sonderzeichen durchgängig verarbeitet werden können und beim Datenaustausch zwischen diesen Systemen einheitlich durchgereicht und korrekt interpretiert werden.

„“ des Bundesverwaltungsamts ist lizenziert unter einer Creative Commons Namensnennung 4.0 International Lizenz.

2. Überblick

2.1. Aufbau und Zweck des Dokuments

In diesem Dokument werden zunächst die Anforderungen aufgeführt, die an die Verarbeitung von Sonderzeichen innerhalb von auf IsyFact-Standards basierenden Anwendungen gestellt werden. Anschließend wird der Zeichensatz festgelegt, der innerhalb der Anwendungen verwendet werden soll. Abschließend wird für die technischen Systeme angegeben, wo die entsprechenden Konfigurationseinträge zur Verwendung des Zeichensatzes vorgenommen werden müssen.

Internationalisierung bedeutet, ein Programm so zu entwerfen und umzusetzen, dass die Anpassung an andere Sprachen möglich ist, ohne den Quellcode zu ändern.

Internationalisierung ist nicht Bestandteil dieses Dokuments!

2.2. Anforderungen und Randbedingungen

An die Verarbeitung von Sonderzeichen nach IsyFact-Standards bestehen die folgenden Anforderungen:

- Prinzipiell (technisch) muss jedes Sonderzeichen nutzbar sein.
- Jedes System nach IsyFact-Standards muss die Sonderzeichen in gleicher Weise verarbeiten können.
- Nach Erlass des BMI ist für das Personenstands,- Melde- und Ausländerwesen die Verarbeitung sämtlicher Zeichen gem. [XOEVStringLatin] vorgegeben. Weder mehr noch weniger Zeichen darf/muss das System verarbeiten können.

2.3. Festlegung des Zeichensatzes und der Codierung

Im SAGA-Standard 4.0 [SAGA40] wird der Zeichensatz Unicode v4.x (ISO 10646:2003) in der UTF-8-Codierung als obligatorisch aufgeführt. Das BMI hat als Rahmenbedingung für seinen Verantwortungsbereich auf der Basis des SAGA-Standards festgelegt, für die Zeichencodierung in neuen Systemen ausschließlich UTF-8 zu nutzen. Dieser Zeichensatz stellt ausreichend viele der weltweit existierenden Buchstaben, Ziffern und Symbole zur Verfügung, um Daten in internationalen Schreibweisen abbilden zu können.

Gemäß den Anforderungen des String-latin Zeichensatzes des BMI [XOEVStringLatin] soll die IsyFact genau diesen Zeichensatz unterstützen. Daher wird festgelegt, dass für IsyFact-Standard-basierte Anwendungen der Zeichensatz Unicode v4.x in der UTF-8-Codierung zu verwenden ist.

„“ des Bundesverwaltungsamts ist lizenziert unter einer Creative Commons Namensnennung 4.0 International Lizenz.

3. Konfigurationseinstellungen für den Zeichensatz

Im Folgenden wird die Konfiguration der technischen Systeme zur Verwendung des Zeichensatzes erläutert. Um zu erreichen, dass jedes IsyFact-Standard-konforme System Sonderzeichen in gleicher Weise verarbeitet, wird durchgängig Unicode v4.x in der UTF-8-Codierung verwendet.

3.1. Betriebssystem

Die Standard-Zeichencodierung aller in der Plattform verwendeten Betriebssysteme muss einheitlich auf die Verwendung von Unicode v4.x in der UTF-8-Codierung gesetzt werden.

Als Beispiel wird hier das Betriebssystem SUSE Linux Enterprise Server (SLES) 10 betrachtet. Hier ist die Standard-Zeichencodierung UTF-8. Diese kann über den Konfigurationsparameter

```
LC_CTYPE = UTF8
```

auch für jeden Benutzer individuell gesetzt werden.

3.2. Oracle Datenbank

Die Zeichencodierung aller in der Plattform verwendeten Datenbanken muss ebenfalls einheitlich auf die Verwendung von Unicode v4.x in der UTF-8-Codierung gesetzt werden.

Als Beispiel wird hier die Datenbank Oracle 11g betrachtet. Oracle unterstützt ab Version 10g Release 2 Unicode v4.0. Oracle empfiehlt [DGSG], neue Datenbanken als Unicode-Datenbanken anzulegen. Hierzu muss beim CREATE DATABASE die folgende Eigenschaft gesetzt werden:

```
CHARACTER SET AL32UTF8
```

3.3. Hibernate

Für Hibernate werden der Unicode-Zeichensatz und die UTF-Zeichencodierung über die beiden Parameter

```
hibernate.connection.useUnicode = true
```

und

```
hibernate.connection.characterEncoding = utf-8
```

konfiguriert. Im Kontext der IsyFact-Standards wird Hibernate nicht direkt, sondern über JPA und Spring genutzt. Hierzu sind diese Einstellungen in der entsprechenden Konfigurationsdatei `jpa.xml` unter den Properties des Entity Managers wie folgt abzulegen:

„“ des Bundesverwaltungsamts ist lizenziert unter einer Creative Commons Namensnennung 4.0 International Lizenz.

```
<property
    <bean id="entityManagerFactory"
class="org.springframework.orm.jpa.LocalContainerEntityManagerFactoryBean">
  <property name="jpaProperties">
    <props>
      …
      <prop key="hibernate.connection.useUnicode">true</prop>
      <prop key="hibernate.connection.characterEncoding">utf-8</prop>
  </property>
</bean>
```

In Eclipse ist an mehreren Stellen die Zeichencodierung zu setzen. Das erfolgt über den Preferences-Dialog von Eclipse, der über die Menüleiste aufgerufen wird („Window -> Preferences...“). Folgende Einstellungen sind zu machen:

```
General -> Workspace: Text file encoding - Other = UTF-8
Web and XML -> CSS Files: Encoding = ISO 10646/Unicode(UTF-8)
Web and XML -> HTML Files: Encoding = ISO 10646/Unicode(UTF-8)
Web and XML -> JSP Files: Encoding = ISO 10646/Unicode(UTF-8)
Web and XML -> XML Files: Encoding = ISO 10646/Unicode(UTF-8)
```

Achtung: Diese Einstellungen sind Workspace-spezifisch, d.h. sie müssen für jeden Workspace individuell eingestellt werden.

3.4. Java

Im Java-Compiler wird die Zeichencodierung der Quelldateien beim Aufruf über den Parameter

```
-encoding UTF-8
```

gesetzt. In der JVM wird die Standard-Zeichencodierung beim Aufruf über den Parameter

```
-Dfile.encoding=UTF-8
```

gesetzt.

3.5. Java Property-Dateien

Bis zur Java-Version 1.5 werden Property-Dateien grundsätzlich ISO 8859-1 codiert gelesen und geschrieben. Das ist unabhängig von den Einstellungen des Zeichensatzes in der JVM und im Betriebssystem. Das Tool native2ascii (Native-to-ASCII Converter, siehe http://docs.oracle.com/javase/1.5.0/docs/tooldocs/windows/native2ascii.html) kann für die Umcodierung von Property-Dateien verwendet werden.

Bei XML-basierten Property-Dateien, können auch andere Zeichencodierungen verwendet werden.

3.6. Maven

Der Build erfolgt mit Maven. Hier ist die Zeichencodierung wie folgt zu setzen:

```
<project>
```

„“ des Bundesverwaltungsamts ist lizenziert unter einer Creative Commons Namensnennung 4.0 International Lizenz.

```
...
  <build>
    <plugins>
      <plugin>
        <groupId>org.apache.maven.plugins</groupId>
        <artifactId>maven-resources-plugin</artifactId>
...
        <configuration>
          <encoding>UTF-8</encoding>
        </configuration>
      </plugin>
...
      <plugin>
        <artifactId>maven-compiler-plugin</artifactId>
        <configuration>
...
          <compilerArguments>
            <encoding>UTF-8</encoding>
...
          </compilerArguments>
        </configuration>
      </plugin>
...
    </plugins>
...
  </build>
...

  <reporting>
    <plugin>
      <groupId>org.apache.maven.plugins</groupId>
      <artifactId>maven-javadoc-plugin</artifactId>
...
        <configuration>
          <encoding>UTF-8</encoding>
        </configuration>
    </plugin>
...

  </reporting>
...
</project>
```

3.7. XML

UTF-8 ist die Standard-Zeichencodierung für XML. Das wird in der ersten Zeile der XML-Datei wie folgt deklariert:

```
<?xml version="1.0" encoding="UTF-8"?>
```

3.8. HTML

In HTML wird die Zeichencodierung in den Metadaten des HEAD-Tags wie folgt angegeben:

```
<meta http-equiv="Content-Type"
      content="text/html; charset=utf-8" />
```

„“ des Bundesverwaltungsamts ist lizenziert unter einer Creative Commons Namensnennung 4.0 International Lizenz.

Damit dürfen auch keine HTML-Sonderzeichen mehr verwendet werden, sondern nur noch UTF-8-codierte Zeichen.

„“ des Bundesverwaltungsamts ist lizenziert unter einer Creative Commons Namensnennung 4.0 International Lizenz.

4. Transformation von Sonderzeichen

In den Fällen, wo kein Unicode-Zeichensatz verwendet werden kann, müssen Sonderzeichen eventuell in andere Darstellungen oder Codierungen umgewandelt werden. Hierzu gibt es prinzipiell drei Möglichkeiten: die Transkription, die Umcodierung und das Filtern von Zeichen. In diesem Kapitel werden diese drei Möglichkeiten in je einem Unterkapitel beschrieben.

4.1. Transkription

Transkription (Umschreibung) ist eine aussprachebasierte Darstellung eines fremden Alphabetes mit dem eigenen Alphabet, also z.B. die Darstellung russischer Namen in kyrillischer Schreibweise mit dem deutschen Alphabet. Transkription wird eingesetzt, um ohne Kenntnisse einer fremdem Sprache und des zugehörigen Alphabets eine halbwegs richtige Aussprache von Wörtern zu ermöglichen. Eine eindeutige Rückübertragung ist in der Regel nicht möglich. Im Folgenden werden die Festlegungen zur Transkription im Rahmen der IsyFact-Standards beschrieben.

4.1.1 Zeichensätze und Sprachen

Wie in Kapitel 2.3 festgelegt, wird für die IsyFact der Zeichensatz Unicode v4.x in der UTF-8-Codierung verwendet. Die Transkription überführt die internationalen Sonderzeichen aus dem Unicode v4.x Zeichensatz in den ASCII-Zeichensatz.

Im Rahmen der IsyFact werden zur Zeit von der Transkription nur kyrillische, griechische und lateinische Zeichen unterstützt, da hiermit die im europäischen Raum gebräuchlichen Zeichen abgedeckt sind.

4.1.2 Anwendungsbereiche in einer IsyFact-Systemlandschaft

Transkription ist an den folgenden Stellen von Bedeutung:

Datenaustausch mit anderen Systemen

Für die Anwendungen nach IsyFact-Standard ist der zu verwendende Zeichensatz festgelegt. Andere Systeme, mit denen diese kommunizieren, können aber einen anderen Zeichensatz verwenden. Hier müssen die Daten zunächst in den Zeichensatz des Zielsystems umgewandelt werden. Die Umwandlung kann durch Transkription geschehen.

Beispiel: Ein Nachbarsystem arbeitet ausschließlich mit dem ASCII-Zeichensatz. Daten einer Anwendung nach IsyFact-Standard werden zunächst umgeschrieben und dann dem Nachbarsystem übergeben.

Einheitliche Repräsentation von Daten

Für Namen können verschiedene ländertypische Schreibweisen genutzt werden. Trotzdem sollen Daten aber vergleichbar sein. Hier kann die Transkription zu einer einheitlichen (normierten) Schreibweise führen. Werden dann Suchen auf den umgeschriebenen Daten durchgeführt, erhöht sich die Wahrscheinlichkeit, dass der Gesuchte in der Trefferliste ist. Dadurch verbessert sich aber nicht unbedingt die Trefferqualität.

„“ des Bundesverwaltungsamts ist lizenziert unter einer Creative Commons Namensnennung 4.0 International Lizenz.

Beispiel: „Müller“ wird im Originalschreibweise gespeichert und für die Suche zu „Mueller“ umgeschrieben. Eine Suchanfrage nach „Müller“ wird zunächst zu „Mueller“ umgeschrieben, dann gesucht und auch gefunden. Eine Suchanfrage nach „Mueller“ braucht nicht umgeschrieben werden und wird gefunden.

Transkription wird in der Regel nur für Namen verwendet, also für Vornamen, Nachnamen und Ortsbezeichnungen.

4.1.3 Transkriptionsregeln

Die Transkription basiert auf dem ICAO-Standard (ICAO-MRTD). Der ICAO-Standard wurde ursprünglich für das automatische Lesen von Dokumenten in der Luftfahrt entwickelt. Er umfasst 142 Abbildungsvorschriften (Regeln) für lateinische und kyrillische Buchstaben. Für die Abbildung von griechischen Zeichen wird der Standard ISO-843 verwendet.

Während der ISO-Standard (ISO-9) für die Transkription von kyrillischen Zeichen noch diakritische lateinische Zeichen verwendet, ist bei ICAO-MRTD das Ziel, diakritische Zeichen vollständig zu vermeiden, um eine Abbildung auf den ASCII-Zeichensatz zu ermöglichen. Eine bereits umgeschriebene Zeichenfolge wird durch eine erneute Transkription nicht mehr verändert.

Die Tabelle für die Abbildungsregeln ist im Dokument [Transskriptionsregeln] enthalten.

4.1.4 Umsetzung im System

Daten werden immer im Originalformat gespeichert. Umgeschriebene Daten können bei Bedarf zusätzlich abgelegt werden. Dabei sind die der Transkription zugrunde liegenden Parameter ebenfalls mit abzulegen. Dies führt zu folgendem Datentyp für umgeschriebene Zeichenfolgen:

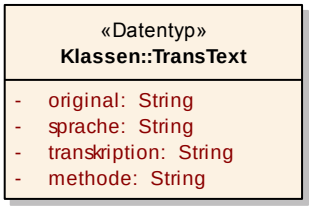

Abbildung 1: Datentyp für umgeschriebene Texte

Die Attribute für den Datentyp „TransText“ haben die folgende Bedeutung:

Attribut	optional	Beschreibung
`original`	nein	Originaltext im Unicode-Zeichenformat
`sprache`	ja	Sprachcode gemäß ISO 639 für die Sprache des Originaltextes

„“ des Bundesverwaltungsamts ist lizenziert unter einer Creative Commons Namensnennung 4.0 International Lizenz.

Attribut	optional	Beschreibung
`transkription`	nein	umgeschriebener Text
`methode`	nein	Kennzeichen für den bei der Transkription verwendeten Satz von Transkriptionsregeln, also der Methode nach der die Transkription durchgeführt wurde. Verschiedene Versionen der gleichen Transkriptionsregeln können durch eigene Kennzeichen abgebildet werden.

Tabelle 1: Attribute des Datentyps „TransText“

Die Transkription soll nicht als zentraler Dienst sondern als Komponente umgesetzt werden, die bei Bedarf in die Anwendungen eingebunden wird. Dabei sind die Transkriptionsregeln in einer oder mehreren Konfigurationsdateien hinterlegt, die von der Komponente eingelesen werden. Darüber wird auch eine einfache Erweiterbarkeit der Transkriptionsregeln gewährleistet. Es ist möglich, mehrere Sätze von Transkriptionsregeln zu hinterlegen, um so auch andere Standards für die Transkription verwenden zu können.

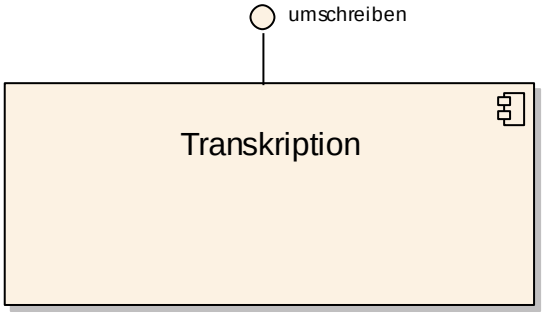

Abbildung 2: Komponente Transkription

Die Komponente Transkription bietet nach außen nur die Methode

```
TransText umschreiben(String text, String sprache,
                      String methode)
```

an. Hier ist der Parameter `text` der umzuschreibende Text, `sprache` der Sprachcode gemäß ISO 639 und `methode` das Kennzeichen des zu verwendenden Satzes von Transkriptionsregeln. Ergebnis ist die umgeschriebene Darstellung des Textes gemäß dem Datentyp `TransText`. Im Fehlerfall werden entsprechende Exceptions geworfen. Die Angabe der Sprache ist optional. Ist die Sprache unbekannt, d.h. es wird kein Sprachcode übergeben, dann wird die Sprache bei der Transkription nicht berücksichtigt.

„“ des Bundesverwaltungsamts ist lizenziert unter einer Creative Commons Namensnennung 4.0 International Lizenz.

4.2. Umcodierung

Textdaten, die von der Anwendung aus einer Datei eingelesen werden oder über eine Programm-Schnittstelle übergeben werden, können eventuell nicht in UTF-8 codiert sein.

Textdateien werden in der Standard-Zeichencodierung der JVM eingelesen und gespeichert (siehe auch Kapitel 3.4). Sollte eine andere Zeichencodierung verwendet werden, so muss dies explizit im Code umgesetzt werden.

Das kann z.B. erfolgen, indem die Dateien mit einem `InputStreamReader` gelesen werden bzw. mit einem `OutputStreamWriter` geschrieben werden. In beiden Klassen kann im Konstruktor der Zeichensatz angegeben werden. Beim Lesen werden die Daten dann automatisch decodiert bzw. beim Schreiben codiert.

Dieses Verfahren kann für beliebige Byte-Arrays verwendet werden, so dass auch Daten, die über eine Programm-Schnittstelle übergeben werden, so umcodiert werden können.

4.3. Filtern von Zeichen

Neben den druckbaren Zeichen enthält der Unicode-Zeichensatz auch nicht druckbare Steuerzeichen (Ugs. „Schmierzeichen“). Diese Zeichen können an der Oberfläche bei der Übernahme aus anderen Programmen über die Zwischenablage oder beim Import von Daten in eine IsyFact-konforme Anwendung gelangen. Diese Zeichen sind prinzipiell bei der Validierung der Daten auszufiltern. Ob der Benutzer von diesem Vorgang informiert wird oder ob Log-Einträge geschrieben werden, hängt von der Fachlichkeit der jeweiligen Anwendung ab. Je nach Anwendung kann es auch sinnvoll sein, einige Steuerzeichen, wie z.B. einen Zeilenumbruch, zuzulassen. Diese von der Anwendung abhängigen Festlegungen müssen in der Spezifikation bzw. im Systementwurf der jeweiligen Anwendung beschrieben werden.

4.4. Spezifikation von fachlichen Datentypen

Bereits in der Spezifikation ist darauf zu achten, dass für einen fachlichen Datentyp die zulässigen Zeichen genau angegeben werden. Nur so können die entsprechenden Validierungen konzipiert und umgesetzt werden. Hier ist der Datentyp String bzw. Alpha in der Regel zu grob. Hier müssen abgestufte Typen für Textinhalte definiert werden, z.B. Alpha-Latein-Basis (alle großen und kleinen lateinischen Buchstaben ohne diakritische Zeichen), Alpha-Latein-Diakrit (alle großen und kleinen lateinischen Buchstaben inklusiv diakritische Zeichen), Alpha-Europa (alle großen und kleinen lateinischen, griechischen und kyrillischen Zeichen, inklusiv diakritischer Zeichen).

„“ des Bundesverwaltungsamts ist lizenziert unter einer Creative Commons Namensnennung 4.0 International Lizenz.

5. Bibliothek „isy-sonderzeichen“

Dieses Kapitel beschreibt die Verwendung des Bausteins „isy-sonderzeichen“.

Der Baustein „isy-sonderzeichen“ ist eine Querschnittskomponente, die anderen Anwendungen Services zur Transformation von Zeichenketten zur Verfügung stellt.

Die Bibliothek stellt dabei eine feste Anzahl von Transformatoren zur Verfügung, die für eine einheitliche Transformation von Zeichenketten innerhalb der Systemumgebung sorgen.

5.1. Funktionsweise

Die Transformatoren arbeiten alle nach dem gleichen Schema. Sie unterscheiden sich nur durch unterschiedliche Tabellen, die zur Zeichentransformation herangezogen werden.

(1) Alle Zeichen werden gemäß einer Mapping-Tabelle transformiert [SLMapping].

(2) Unbekannte oder nicht abbildbare Zeichen werden durch Leerzeichen ersetzt.

(3) Leerzeichen am Anfang und am Ende der Zeichenkette werden entfernt.

(4) Zwei aufeinanderfolgende Leerzeichen werden durch ein einzelnes Leerzeichen ersetzt.

Transformatoren müssen in der Regel projektspezifisch entwickelt werden. Darüber hinaus werden folgende Transformatoren mitgeliefert:

Identischer Transformator

Dieser Transformator bildet alle gültigen String.Latin-Zeichen auf sich selber ab (Spalte C, Tabelle [SLMapping]). Der Nutzen dieses Transformators liegt darin, dass alle nicht String.Latin-Zeichen aus der übergebenen Zeichenkette entfernt werden. Dieser Transformator ermöglicht keine Vorgabe der maximalen Zeichenlänge.

5.2. Einbindung der Bibliothek in eine Anwendung

Um die Bibliothek in einer Anwendung nutzen zu können, sind zwei Schritte notwendig

- Integration mit Maven und
- Instanziierung der Transformator Factory.

5.2.1 Integration mit Maven

In der POM der Anwendung muss die Abhängigkeit hinzugefügt werden:

„“ des Bundesverwaltungsamts ist lizenziert unter einer Creative Commons Namensnennung 4.0 International Lizenz.

```
<dependency>
       <groupId>de.bund.bva.pliscommon</groupId>
       <artifactId>isy-sonderzeichen</artifactId>
       <version><aktueller Version der Bibliothek></version>
</dependency>
```

5.2.2 Instanziierung der Transformator Factory

Die Transformator-Factory und ein konkreter Transformator werden über Spring instanziiert.

```
<bean  id="sonderzeichenTransformatorFactory"
       class="de.bund.bva.pliscommon.plissonderzeichen.core.tr
       ansformation.TransformatorFactory">
       <property
          name="transformator"
          ref="sonderzeichenTransformator" />
       <property
          name=" transformationsTabelle"
          value="${Pfad_zu_einer_zusaetzlichen_Tabelle}"/>
</bean>
<bean  id="sonderzeichenTransformator"
       class="de.bund.bva.pliscommon.plissonderzeichen.core.tr
       ansformation.impl.IdentischerTransformator">
</bean>
```

In obigem Beispiel wird dabei der Transformator *IdentischerTransformator* geladen. Jeder der Transformatoren setzt bereits eine fest implementierte Transformationstabelle nach einem bestimmten Vorgehen um (siehe 5.1).

Bei der Konfiguration der *TransformatorFactory* kann die zusätzliche (optionale) Eigenschaft *transformationsTabelle* dazu genutzt werden, eine weitere Transformationstabelle anzugeben. Die Regeln in dieser Tabelle überschreiben dabei existierende alte Regeln. Es findet also eine Ergänzung der existierenden Regeln statt.

5.3. Schnittstellendefinition

Der Aufruf des Transformators erfolgt über die jeweilige Methode der Transformator Schnittstelle. Folgende Methoden stehen zur Verfügung:

Methode	Parameter
`transformiere` Transformiert eine Zeichenkette auf der Basis der zugrunde liegenden Transformationstabelle. Leerzeichen am Anfang und am Ende der Zeichenkette werden entfernt. Doppelte Leerzeichen innerhalb der Zeichenkette werden zu einem Leerzeichen umgewandelt.	`String zeichenkette` Die zu transformierende Zeichenkette
`transformiere` Transformiert eine Zeichenkette	`String zeichenkette` Die zu transformierende

„“ des Bundesverwaltungsamts ist lizenziert unter einer Creative Commons Namensnennung 4.0 International Lizenz.

Methode	Parameter
analog der zuvor beschriebenen `transformiere`-Funktion. Stellt zusätzlich sicher, dass die Zeichenkette nach der Operation die angegebene Länge hat. Es wird dabei nicht unterschieden, ob die ursprüngliche Zeichenkettenlänge bereits das Maximum überschritten hat oder erst durch eine Transformation die Zeichenkette verlängert wurde.	Zeichenkette `int maximaleLaenge` Die maximale Länge der Zeichenkette
`transformiereOhneTrim` Transformiert eine Zeichenkette analog der zuvor beschriebenen `transformiere`-Funktion. Es werden jedoch keine Leerzeichen am Anfang/Ende der übergebenen Zeichenkette entfernt.	`String zeichenkette` Die zu transformierende Zeichenkette
`getRegulaererAusdruck` Gibt den regulären Ausdruck zurück, der alle gültige Zeichenketten beschreibt, deren Zeichen in der jeweiligen Zeichenkategorie aufgeführt sind.	`String[] kategorieListe` Eine Liste mit den Zeichenkategorien. Gültige Werte sind `LETTER, NUMBER, PUNCTUATION, SEPARATOR, SYMBOL, OTHER.` Die Werte sind der Konstantenklasse `ZeichenKategorie` zu entnehmen.
`getGueltigeZeichen` Gibt alle gültigen Zeichen des Transformators zurück.	`String kategorie` Eine Zeichenkategorie aus `LETTER, NUMBER, PUNCTUATION, SEPARATOR, SYMBOL, OTHER.`

Hinweis zur Funktion `transformiere`

Die Transformationsfunktion arbeitet die Zeichenkette char für char ab. Sollte ein Unicode-Character, welcher aus mehreren char Objekten besteht definiert sein (non-BMP character, z.B. I mit angehängtem Circumflex (\\u006C\\u0302), so liefert die Transformationsfunktion das korrekte Ergebnis, kann aber nicht zwischen String.Latin- und Nicht-String.Latin-Zeichen unterscheiden. So könnten Zeichen außerhalb des

„“ des Bundesverwaltungsamts ist lizenziert unter einer Creative Commons Namensnennung 4.0 International Lizenz.

Definitionsbereichs (z.B. alle \\u####\\u0302) der Transformation transformiert werden.

Zur Überprüfung ob eine Zeichenkette innerhalb des für den Transformator gültigen Bereichs liegt, sollte daher die Funktion getRegulaererAusdruck(String[]) benutzt werden um einen regulären Ausdruck für alle gültigen Zeichen zu erstellen.

„“ des Bundesverwaltungsamts ist lizenziert unter einer Creative Commons Namensnennung 4.0 International Lizenz.

6. Zulässige Zeichen innerhalb der IsyFact

Die im Rahmen der IsyFact zugelassenen Zeichen gliedern sich in Standardzeichen und zusätzliche Zeichen. Die Standardzeichen müssen von jeder Anwendung immer unterstützt werden. Die zusätzlichen Zeichen müssen nur unterstützt werden, wenn dies entsprechend vereinbart wurde. Die Festlegungen für die zulässigen Zeichen orientieren sich an den Festlegungen, die für das Meldewesen getroffen wurden.

Die für die IsyFact zulässigen Zeichen werden im Folgenden aufgeführt. (s. Kapitel 2.2)

6.1. Standardzeichen

- Großbuchstaben: A-Z Ä Ö Ü
- Kleinbuchstaben: a-z ä ö ü ß
- Ziffern: 0-9
- **Sonderzeichen**: ' ( ) + , - . / Leerzeichen

6.2. Zusätzliche Zeichen

In der Tabelle 2 sind die Zeichen dargestellt, die zusätzlich unterstützt werden. Damit die Zeichen in der Spalte „Glyph“ korrekt dargestellt werden, muss ein Font installiert sein, der alle Zeichen unterstützt. (z.B. Code2000, erhältlich unter http://www.code2000.net).

Unicode-Wert U+….	Glyph	Unicode-Zeichenname
0009		CHARACTER TABULATION
000A		LINE FEED
000D		CARRIAGE RETURN
0021	!	EXCLAMATION MARK
0022	"	QUOTATION MARK
0023	#	NUMBER SIGN
0024	$	DOLLAR SIGN
0025	%	PERCENT SIGN
0026	&	AMPERSAND
002A	*	ASTERISK
003A	:	COLON
003B	;	SEMICOLON
003C	<	LESS-THAN SIGN
003D	=	EQUALS SIGN

„“ des Bundesverwaltungsamts ist lizenziert unter einer Creative Commons Namensnennung 4.0 International Lizenz.

Unicode-Wert U+….	Glyph	Unicode-Zeichenname
003E	>	GREATER-THAN SIGN
003F	?	QUESTION MARK
0040	@	COMMERCIAL AT
0044+0302	D̂	LATIN CAPITAL LETTER D WITH COMBINING CIRCUMFLEX ACCENT
004A+030C	J̌	LATIN CAPITAL LETTER J WITH COMBINING CARON
004C+0302	L̂	LATIN CAPITAL LETTER L WITH COMBINING CIRCUMFLEX ACCENT
004D+0302	M̂	LATIN CAPITAL LETTER M WITH COMBINING CIRCUMFLEX ACCENT
004E+0302	N̂	LATIN CAPITAL LETTER N WITH COMBINING CIRCUMFLEX ACCENT
005B	[	LEFT SQUARE BRACKET
005C	\	REVERSE SOLIDUS
005D	]	RIGHT SQUARE BRACKET
005E	^	CIRCUMFLEX ACCENT
005F	_	LOW LINE
0060	`	GRAVE ACCENT
0064+0302	d̂	LATIN SMALL LETTER D WITH COMBINING CIRCUMFLEX ACCENT
006C+0302	l̂	LATIN SMALL LETTER L WITH COMBINING CIRCUMFLEX ACCENT
006D+0302	m̂	LATIN SMALL LETTER M WITH COMBINING CIRCUMFLEX ACCENT
006E+0302	n̂	LATIN SMALL LETTER N WITH COMBINING CIRCUMFLEX ACCENT
007B	{	LEFT CURLY BRACKET
007C	|	VERTICAL LINE
007D	}	RIGHT CURLY BRACKET
007E	~	TILDE
00A1	Í	INVERTED EXCLAMATION MARK
00A2	¢	CENT SIGN
00A3	£	POUND SIGN
00A4	¤	CURRENCY SIGN

„“ des Bundesverwaltungsamts ist lizenziert unter einer Creative Commons Namensnennung 4.0 International Lizenz.

Unicode-Wert U+….	Glyph	Unicode-Zeichenname
00A5	¥	YEN SIGN
00A6	¦	BROKEN BAR
00A7	§	SECTION SIGN
00A8	¨	DIAERESIS
00A9	©	COPYRIGHT SIGN
00AA	ª	FEMININE ORDINAL INDICATOR
00AB	«	LEFT-POINTING DOUBLE ANGLE QUOTATION MARK
00AC	¬	NOT SIGN
00AE	®	REGISTERED SIGN
00AF	¯	MACRON
00B0	°	DEGREE SIGN
00B1	±	PLUS-MINUS SIGN
00B2	²	SUPERSCRIPT TWO
00B3	³	SUPERSCRIPT THREE
00B4	´	ACUTE ACCENT
00B5	µ	MICRO SIGN
00B6	¶	PILCROW SIGN
00B7	·	MIDDLE DOT
00B8	¸	CEDILLA
00B9	¹	SUPERSCRIPT ONE
00BA	º	MASCULINE ORDINAL INDICATOR
00BB	»	RIGHT-POINTING DOUBLE ANGLE QUOTATION MARK
00BC	¼	VULGAR FRACTION ONE QUARTER
00BD	½	VULGAR FRACTION ONE HALF
00BE	¾	VULGAR FRACTION THREE QUARTERS
00BF	¿	INVERTED QUESTION MARK
00C0	À	LATIN CAPITAL LETTER A WITH GRAVE
00C1	Á	LATIN CAPITAL LETTER A WITH ACUTE
00C2	Â	LATIN CAPITAL LETTER A WITH CIRCUMFLEX
00C3	Ã	LATIN CAPITAL LETTER A WITH TILDE

„“ des Bundesverwaltungsamts ist lizenziert unter einer Creative Commons Namensnennung 4.0 International Lizenz.

Unicode-Wert U+….	Glyph	Unicode-Zeichenname
00C5	Å	LATIN CAPITAL LETTER A WITH RING ABOVE
00C6	Æ	LATIN CAPITAL LETTER AE
00C7	Ç	LATIN CAPITAL LETTER C WITH CEDILLA
00C8	È	LATIN CAPITAL LETTER E WITH GRAVE
00C9	É	LATIN CAPITAL LETTER E WITH ACUTE
00CA	Ê	LATIN CAPITAL LETTER E WITH CIRCUMFLEX
00CB	Ë	LATIN CAPITAL LETTER E WITH DIAERESIS
00CC	Ì	LATIN CAPITAL LETTER I WITH GRAVE
00CD	Í	LATIN CAPITAL LETTER I WITH ACUTE
00CE	Î	LATIN CAPITAL LETTER I WITH CIRCUMFLEX
00CF	Ï	LATIN CAPITAL LETTER I WITH DIAERESIS
00D0	Ð	LATIN CAPITAL LETTER ETH
00D1	Ñ	LATIN CAPITAL LETTER N WITH TILDE
00D2	Ò	LATIN CAPITAL LETTER O WITH GRAVE
00D3	Ó	LATIN CAPITAL LETTER O WITH ACUTE
00D4	Ô	LATIN CAPITAL LETTER O WITH CIRCUMFLEX
00D5	Õ	LATIN CAPITAL LETTER O WITH TILDE
00D7	×	MULTIPLICATION SIGN
00D8	Ø	LATIN CAPITAL LETTER O WITH STROKE
00D9	Ù	LATIN CAPITAL LETTER U WITH GRAVE
00DA	Ú	LATIN CAPITAL LETTER U WITH ACUTE
00DB	Û	LATIN CAPITAL LETTER U WITH CIRCUMFLEX
00DD	Ý	LATIN CAPITAL LETTER Y WITH ACUTE
00DE	Þ	LATIN CAPITAL LETTER THORN
00E0	à	LATIN SMALL LETTER A WITH GRAVE
00E1	á	LATIN SMALL LETTER A WITH ACUTE
00E2	â	LATIN SMALL LETTER A WITH CIRCUMFLEX
00E3	ã	LATIN SMALL LETTER A WITH TILDE
00E5	å	LATIN SMALL LETTER A WITH RING ABOVE
00E6	æ	LATIN SMALL LETTER AE
00E7	ç	LATIN SMALL LETTER C WITH CEDILLA
00E8	è	LATIN SMALL LETTER E WITH GRAVE

„“ des Bundesverwaltungsamts ist lizenziert unter einer Creative Commons Namensnennung 4.0 International Lizenz.

Unicode-Wert U+….	Glyph	Unicode-Zeichenname
00E9	é	LATIN SMALL LETTER E WITH ACUTE
00EA	ê	LATIN SMALL LETTER E WITH CIRCUMFLEX
00EB	ë	LATIN SMALL LETTER E WITH DIAERESIS
00EC	ì	LATIN SMALL LETTER I WITH GRAVE
00ED	í	LATIN SMALL LETTER I WITH ACUTE
00EE	î	LATIN SMALL LETTER I WITH CIRCUMFLEX
00EF	ï	LATIN SMALL LETTER I WITH DIAERESIS
00F0	ð	LATIN SMALL LETTER ETH
00F1	ñ	LATIN SMALL LETTER N WITH TILDE
00F2	ò	LATIN SMALL LETTER O WITH GRAVE
00F3	ó	LATIN SMALL LETTER O WITH ACUTE
00F4	ô	LATIN SMALL LETTER O WITH CIRCUMFLEX
00F5	õ	LATIN SMALL LETTER O WITH TILDE
00F7	÷	DIVISION SIGN
00F8	ø	LATIN SMALL LETTER O WITH STROKE
00F9	ù	LATIN SMALL LETTER U WITH GRAVE
00FA	ú	LATIN SMALL LETTER U WITH ACUTE
00FB	û	LATIN SMALL LETTER U WITH CIRCUMFLEX
00FD	ý	LATIN SMALL LETTER Y WITH ACUTE
00FE	þ	LATIN SMALL LETTER THORN
00FF	ÿ	LATIN SMALL LETTER Y WITH DIAERESIS
0100	Ā	LATIN CAPITAL LETTER A WITH MACRON
0101	ā	LATIN SMALL LETTER A WITH MACRON
0102	Ă	LATIN CAPITAL LETTER A WITH BREVE
0103	ă	LATIN SMALL LETTER A WITH BREVE
0104	Ą	LATIN CAPITAL LETTER A WITH OGONEK
0105	ą	LATIN SMALL LETTER A WITH OGONEK
0106	Ć	LATIN CAPITAL LETTER C WITH ACUTE
0107	ć	LATIN SMALL LETTER C WITH ACUTE
010A	Ċ	LATIN CAPITAL LETTER C WITH DOT ABOVE
010B	ċ	LATIN SMALL LETTER C WITH DOT ABOVE
010C	Č	LATIN CAPITAL LETTER C WITH CARON

„“ des Bundesverwaltungsamts ist lizenziert unter einer Creative Commons Namensnennung 4.0 International Lizenz.

Unicode-Wert U+….	Glyph	Unicode-Zeichenname
010D	č	LATIN SMALL LETTER C WITH CARON
010E	Ď	LATIN CAPITAL LETTER D WITH CARON
010F	ď	LATIN SMALL LETTER D WITH CARON
0110	Đ	LATIN CAPITAL LETTER D WITH STROKE
0111	đ	LATIN SMALL LETTER D WITH STROKE
0112	Ē	LATIN CAPITAL LETTER E WITH MACRON
0113	ē	LATIN SMALL LETTER E WITH MACRON
0114	Ĕ	LATIN CAPITAL LETTER E WITH BREVE
0115	ĕ	LATIN SMALL LETTER E WITH BREVE
0116	Ė	LATIN CAPITAL LETTER E WITH DOT ABOVE
0117	ė	LATIN SMALL LETTER E WITH DOT ABOVE
0118	Ę	LATIN CAPITAL LETTER E WITH OGONEK
0119	ę	LATIN SMALL LETTER E WITH OGONEK
011A	Ě	LATIN CAPITAL LETTER E WITH CARON
011B	ě	LATIN SMALL LETTER E WITH CARON
011E	Ğ	LATIN CAPITAL LETTER G WITH BREVE
011F	ğ	LATIN SMALL LETTER G WITH BREVE
0120	Ġ	LATIN CAPITAL LETTER G WITH DOT ABOVE
0121	ġ	LATIN SMALL LETTER G WITH DOT ABOVE
0122	Ģ	LATIN CAPITAL LETTER G WITH CEDILLA
0123	ģ	LATIN SMALL LETTER G WITH CEDILLA
0126	Ħ	LATIN CAPITAL LETTER H WITH STROKE
0127	ħ	LATIN SMALL LETTER H WITH STROKE
0128	Ĩ	LATIN CAPITAL LETTER I WITH TILDE
0129	ĩ	LATIN SMALL LETTER I WITH TILDE
012A	Ī	LATIN CAPITAL LETTER I WITH MACRON
012B	ī	LATIN SMALL LETTER I WITH MACRON
012C	Ĭ	LATIN CAPITAL LETTER I WITH BREVE
012D	ĭ	LATIN SMALL LETTER I WITH BREVE
012E	Į	LATIN CAPITAL LETTER I WITH OGONEK
012F	į	LATIN SMALL LETTER I WITH OGONEK
0130	İ	LATIN CAPITAL LETTER I WITH DOT ABOVE

„“ des Bundesverwaltungsamts ist lizenziert unter einer Creative Commons Namensnennung 4.0 International Lizenz.

Unicode-Wert U+….	Glyph	Unicode-Zeichenname
0131	ı	LATIN SMALL LETTER DOTLESS I
0134	Ĵ	LATIN CAPITAL LETTER J WITH CIRCUMFLEX
0135	ĵ	LATIN SMALL LETTER J WITH CIRCUMFLEX
0136	Ķ	LATIN CAPITAL LETTER K WITH CEDILLA
0137	ķ	LATIN SMALL LETTER K WITH CEDILLA
0138	ĸ	LATIN SMALL LETTER KRA
0139	Ĺ	LATIN CAPITAL LETTER L WITH ACUTE
013A	ĺ	LATIN SMALL LETTER L WITH ACUTE
013B	Ļ	LATIN CAPITAL LETTER L WITH CEDILLA
013C	ļ	LATIN SMALL LETTER L WITH CEDILLA
013D	Ľ	LATIN CAPITAL LETTER L WITH CARON
013E	ľ	LATIN SMALL LETTER L WITH CARON
013F	Ŀ	LATIN CAPITAL LETTER L WITH MIDDLE DOT
0140	ŀ	LATIN SMALL LETTER L WITH MIDDLE DOT
0141	Ł	LATIN CAPITAL LETTER L WITH STROKE
0142	ł	LATIN SMALL LETTER L WITH STROKE
0143	Ń	LATIN CAPITAL LETTER N WITH ACUTE
0144	ń	LATIN SMALL LETTER N WITH ACUTE
0145	Ņ	LATIN CAPITAL LETTER N WITH CEDILLA
0146	ņ	LATIN SMALL LETTER N WITH CEDILLA
0147	Ň	LATIN CAPITAL LETTER N WITH CARON
0148	ň	LATIN SMALL LETTER N WITH CARON
0149	ŉ	LATIN SMALL LETTER N PRECEDED BY APOSTROPHE
014A	Ŋ	LATIN CAPITAL LETTER ENG
014B	ŋ	LATIN SMALL LETTER ENG
014C	Ō	LATIN CAPITAL LETTER O WITH MACRON
014D	ō	LATIN SMALL LETTER O WITH MACRON
014E	Ŏ	LATIN CAPITAL LETTER O WITH BREVE
014F	ŏ	LATIN SMALL LETTER O WITH BREVE
0150	Ő	LATIN CAPITAL LETTER O WITH DOUBLE ACUTE

„“ des Bundesverwaltungsamts ist lizenziert unter einer Creative Commons Namensnennung 4.0 International Lizenz.

Unicode-Wert U+….	Glyph	Unicode-Zeichenname
0151	ő	LATIN SMALL LETTER O WITH DOUBLE ACUTE
0152	Œ	LATIN CAPITAL LIGATURE OE
0153	œ	LATIN SMALL LIGATURE OE
0154	Ŕ	LATIN CAPITAL LETTER R WITH ACUTE
0155	ŕ	LATIN SMALL LETTER R WITH ACUTE
0156	Ŗ	LATIN CAPITAL LETTER R WITH CEDILLA
0157	ŗ	LATIN SMALL LETTER R WITH CEDILLA
0158	Ř	LATIN CAPITAL LETTER R WITH CARON
0159	ř	LATIN SMALL LETTER R WITH CARON
015A	Ś	LATIN CAPITAL LETTER S WITH ACUTE
015B	ś	LATIN SMALL LETTER S WITH ACUTE
015E	Ş	LATIN CAPITAL LETTER S WITH CEDILLA
015F	ş	LATIN SMALL LETTER S WITH CEDILLA
0160	Š	LATIN CAPITAL LETTER S WITH CARON
0161	š	LATIN SMALL LETTER S WITH CARON
0162	Ţ	LATIN CAPITAL LETTER T WITH CEDILLA
0163	ţ	LATIN SMALL LETTER T WITH CEDILLA
0164	Ť	LATIN CAPITAL LETTER T WITH CARON
0165	ť	LATIN SMALL LETTER T WITH CARON
0166	Ŧ	LATIN CAPITAL LETTER T WITH STROKE
0167	ŧ	LATIN SMALL LETTER T WITH STROKE
0168	Ũ	LATIN CAPITAL LETTER U WITH TILDE
0169	ũ	LATIN SMALL LETTER U WITH TILDE
016A	Ū	LATIN CAPITAL LETTER U WITH MACRON
016B	ū	LATIN SMALL LETTER U WITH MACRON
016E	Ů	LATIN CAPITAL LETTER U WITH RING ABOVE
016F	ů	LATIN SMALL LETTER U WITH RING ABOVE
0170	Ű	LATIN CAPITAL LETTER U WITH DOUBLE ACUTE
0171	ű	LATIN SMALL LETTER U WITH DOUBLE ACUTE
0172	Ų	LATIN CAPITAL LETTER U WITH OGONEK

„“ des Bundesverwaltungsamts ist lizenziert unter einer Creative Commons Namensnennung 4.0 International Lizenz.

Unicode-Wert U+….	Glyph	Unicode-Zeichenname
0173	ų	LATIN SMALL LETTER U WITH OGONEK
0174	Ŵ	LATIN CAPITAL LETTER W WITH CIRCUMFLEX
0175	ŵ	LATIN SMALL LETTER W WITH CIRCUMFLEX
0176	Ŷ	LATIN CAPITAL LETTER Y WITH CIRCUMFLEX
0177	ŷ	LATIN SMALL LETTER Y WITH CIRCUMFLEX
0178	Ÿ	LATIN CAPITAL LETTER Y WITH DIAERESIS
0179	Ź	LATIN CAPITAL LETTER Z WITH ACUTE
017A	ź	LATIN SMALL LETTER Z WITH ACUTE
017B	Ż	LATIN CAPITAL LETTER Z WITH DOT ABOVE
017C	ż	LATIN SMALL LETTER Z WITH DOT ABOVE
017D	Ž	LATIN CAPITAL LETTER Z WITH CARON
017E	ž	LATIN SMALL LETTER Z WITH CARON
018F	Ə	LATIN CAPITAL LETTER SCHWA
01A0	Ơ	LATIN CAPITAL LETTER O WITH HORN
01A1	ơ	LATIN SMALL LETTER O WITH HORN
01AF	Ư	LATIN CAPITAL LETTER U WITH HORN
01B0	ư	LATIN SMALL LETTER U WITH HORN
01B7	Ʒ	LATIN CAPITAL LETTER EZH
01CD	Ǎ	LATIN CAPITAL LETTER A WITH CARON
01CE	ǎ	LATIN SMALL LETTER A WITH CARON
01CF	Ǐ	LATIN CAPITAL LETTER I WITH CARON
01D0	ǐ	LATIN SMALL LETTER I WITH CARON
01D1	Ǒ	LATIN CAPITAL LETTER O WITH CARON
01D2	ǒ	LATIN SMALL LETTER O WITH CARON
01D3	Ǔ	LATIN CAPITAL LETTER U WITH CARON
01D4	ǔ	LATIN SMALL LETTER U WITH CARON
01DE	Ǟ	LATIN CAPITAL LETTER A WITH DIAERESIS AND MACRON
01DF	ǟ	LATIN SMALL LETTER A WITH DIAERESIS AND MACRON
01E4	Ǥ	LATIN CAPITAL LETTER G WITH STROKE
01E5	ǥ	LATIN SMALL LETTER G WITH STROKE

„“ des Bundesverwaltungsamts ist lizenziert unter einer Creative Commons Namensnennung 4.0 International Lizenz.

Unicode-Wert U+….	Glyph	Unicode-Zeichenname
01E6	Ǧ	LATIN CAPITAL LETTER G WITH CARON
01E7	ǧ	LATIN SMALL LETTER G WITH CARON
01E8	Ǩ	LATIN CAPITAL LETTER K WITH CARON
01E9	ǩ	LATIN SMALL LETTER K WITH CARON
01EA	Ǫ	LATIN CAPITAL LETTER O WITH OGONEK
01EB	ǫ	LATIN SMALL LETTER O WITH OGONEK
01EC	Ǭ	LATIN CAPITAL LETTER O WITH OGONEK AND MACRON
01ED	ǭ	LATIN SMALL LETTER O WITH OGONEK AND MACRON
01EE	Ǯ	LATIN CAPITAL LETTER EZH WITH CARON
01EF	ǯ	LATIN SMALL LETTER EZH WITH CARON
01F0	ǰ	LATIN SMALL LETTER J WITH CARON
01F4	Ǵ	LATIN CAPITAL LETTER G WITH ACUTE
01F5	ǵ	LATIN SMALL LETTER G WITH ACUTE
01FA	Ǻ	WITH RING ABOVE AND ACUTE
01FB	ǻ	LATIN SMALL LETTER A WITH RING ABOVE AND ACUTE
01FC	Ǽ	LATIN CAPITAL LETTER AE WITH ACUTE
01FD	ǽ	LATIN SMALL LETTER AE WITH ACUTE
01FE	Ǿ	LATIN CAPITAL LETTER O WITH STROKE AND ACUTE
01FF	ǿ	LATIN SMALL LETTER O WITH STROKE AND ACUTE
0218	Ș	LATIN CAPITAL LETTER S WITH COMMA BELOW
0219	ș	LATIN SMALL LETTER S WITH COMMA BELOW
021A	Ț	LATIN CAPITAL LETTER T WITH COMMA BELOW
021B	ț	LATIN SMALL LETTER T WITH COMMA BELOW
021E	Ȟ	LATIN CAPITAL LETTER H WITH CARON
021F	ȟ	LATIN SMALL LETTER H WITH CARON
022A	Ȫ	LATIN CAPITAL LETTER O WITH DIAERESIS AND MAC RON
022B	ȫ	LATIN SMALL LETTER O WITH DIAERESIS AND MACRON
022E	Ȯ	LATIN CAPITAL LETTER O WITH DOT ABOVE
022F	ȯ	LATIN SMALL LETTER O WITH DOT ABOVE

„“ des Bundesverwaltungsamts ist lizenziert unter einer Creative Commons Namensnennung 4.0 International Lizenz.

Unicode-Wert U+….	Glyph	Unicode-Zeichenname
0230	Ȱ	LATIN CAPITAL LETTER O WITH DOT ABOVE AND MAC RON
0231	ȱ	LATIN SMALL LETTER O WITH DOT ABOVE AND MACRON
0232	Ȳ	LATIN CAPITAL LETTER Y WITH MACRON
0233	ȳ	LATIN SMALL LETTER Y WITH MACRON
0259	ə	LATIN SMALL LETTER SCHWA
0292	ʒ	LATIN SMALL LETTER EZH
1E02	Ḃ	LATIN CAPITAL LETTER B WITH DOT ABOVE
1E03	ḃ	LATIN SMALL LETTER B WITH DOT ABOVE
1E0A	Ḋ	LATIN CAPITAL LETTER D WITH DOT ABOVE
1E0B	ḋ	LATIN SMALL LETTER D WITH DOT ABOVE
1E10	Ḑ	LATIN CAPITAL LETTER D WITH CEDILLA
1E11	ḑ	LATIN SMALL LETTER D WITH CEDILLA
1E1E	Ḟ	LATIN CAPITAL LETTER F WITH DOT ABOVE
1E1F	ḟ	LATIN SMALL LETTER F WITH DOT ABOVE
1E20	Ḡ	LATIN CAPITAL LETTER G WITH MACRON
1E21	ḡ	LATIN SMALL LETTER G WITH MACRON
1E24	Ḥ	LATIN CAPITAL LETTER H WITH DOT BELOW
1E25	ḥ	LATIN SMALL LETTER H WITH DOT BELOW
1E26	Ḧ	LATIN CAPITAL LETTER H WITH DIAERESIS
1E27	ḧ	LATIN SMALL LETTER H WITH DIAERESIS
1E30	Ḱ	LATIN CAPITAL LETTER K WITH ACUTE
1E31	ḱ	LATIN SMALL LETTER K WITH ACUTE
1E40	Ṁ	LATIN CAPITAL LETTER M WITH DOT ABOVE
1E41	ṁ	LATIN SMALL LETTER M WITH DOT ABOVE
1E44	Ṅ	LATIN CAPITAL LETTER N WITH DOT ABOVE
1E45	ṅ	LATIN SMALL LETTER N WITH DOT ABOVE
1E56	Ṗ	LATIN CAPITAL LETTER P WITH DOT ABOVE
1E57	ṗ	LATIN SMALL LETTER P WITH DOT ABOVE
1E60	Ṡ	LATIN CAPITAL LETTER S WITH DOT ABOVE
1E61	ṡ	LATIN SMALL LETTER S WITH DOT ABOVE
1E62	Ṣ	LATIN CAPITAL LETTER S WITH DOT BELOW

„“ des Bundesverwaltungsamts ist lizenziert unter einer Creative Commons Namensnennung 4.0 International Lizenz.

Unicode-Wert U+….	Glyph	Unicode-Zeichenname
1E63	ṣ	LATIN SMALL LETTER S WITH DOT BELOW
1E6A	Ṫ	LATIN CAPITAL LETTER T WITH DOT ABOVE
1E6B	ṫ	LATIN SMALL LETTER T WITH DOT ABOVE
1E80	Ẁ	LATIN CAPITAL LETTER W WITH GRAVE
1E81	ẁ	LATIN CAPITAL LETTER W WITH ACUTE
1E82	Ẃ	LATIN CAPITAL LETTER W WITH ACUTE
1E83	ẃ	LATIN SMALL LETTER W WITH ACUTE
1E84	Ẅ	LATIN CAPITAL LETTER W WITH DIAERESIS
1E85	ẅ	LATIN SMALL LETTER W WITH DIAERESIS
1E8C	Ẍ	LATIN CAPITAL LETTER X WITH DIAERESIS
1E8D	ẍ	LATIN SMALL LETTER X WITH DIAERESIS
1E8E	Ẏ	LATIN CAPITAL LETTER Y WITH DOT ABOVE
1E8F	ẏ	LATIN SMALL LETTER Y WITH DOT ABOVE
1E90	Ẑ	LATIN CAPITAL LETTER Z WITH CIRCUMFLEX
1E91	ẑ	LATIN SMALL LETTER Z WITH CIRCUMFLEX
1E92	Ẓ	LATIN CAPITAL LETTER Z WITH DOT BELOW
1E93	ẓ	LATIN SMALL LETTER Z WITH DOT BELOW
1E9E	ß	LATIN CAPITAL LETTER SHARP S
1EA0	Ạ	LATIN CAPITAL LETTER A WITH DOT BELOW
1EA1	ạ	LATIN SMALL LETTER A WITH DOT BELOW
1EA2	Ả	LATIN CAPITAL LETTER A WITH HOOK ABOVE
1EA3	ả	LATIN SMALL LETTER A WITH HOOK ABOVE
1EA4	Ấ	LATIN CAPITAL LETTER A WITH CIRCUMFLEX AND ACUTE
1EA5	ấ	LATIN SMALL LETTER A WITH CIRCUMFLEX AND ACUTE
1EA6	Ầ	LATIN CAPITAL LETTER A WITH CIRCUMFLEX AND GRAVE
1EA7	ầ	LATIN SMALL LETTER A WITH CIRCUMFLEX AND GRAVE
1EAA	Ẫ	LATIN CAPITAL LETTER A WITH CIRCUMFLEX AND TILDE
1EAB	ẫ	LATIN SMALL LETTER A WITH CIRCUMFLEX AND TILDE
1EAC	Ậ	LATIN CAPITAL LETTER A WITH CIRCUMFLEX AND DOT BELOW

„“ des Bundesverwaltungsamts ist lizenziert unter einer Creative Commons Namensnennung 4.0 International Lizenz.

Unicode-Wert U+….	Glyph	Unicode-Zeichenname
1EAE	Ắ	LATIN CAPITAL LETTER A WITH BREVE AND ACUTE
1EAF	ắ	LATIN SMALL LETTER A WITH BREVE AND ACUTE
1EB0	Ằ	LATIN CAPITAL LETTER A WITH BREVE AND GRAVE
1EB1	ằ	LATIN SMALL LETTER A WITH BREVE AND GRAVE
1EB2	Ẳ	LATIN CAPITAL LETTER A WITH BREVE AND HOOK ABOVE
1EB3	ẳ	LATIN SMALL LETTER A WITH BREVE AND HOOK ABOVE
1EB4	Ẵ	LATIN CAPITAL LETTER A WITH BREVE AND TILDE
1EB5	ẵ	LATIN SMALL LETTER A WITH BREVE AND TILDE
1EB6	Ặ	LATIN CAPITAL LETTER A WITH BREVE AND DOT BELOW
1EB7	ặ	LATIN SMALL LETTER A WITH BREVE AND DOT BELOW
1EB8	Ẹ	LATIN CAPITAL LETTER E WITH DOT BELOW
1EB9	ẹ	LATIN SMALL LETTER E WITH DOT BELOW
1EBA	Ẻ	LATIN CAPITAL LETTER E WITH HOOK ABOVE
1EBB	ẻ	LATIN SMALL LETTER E WITH HOOK ABOVE
1EBC	Ẽ	LATIN CAPITAL LETTER E WITH TILDE
1EBD	ẽ	LATIN SMALL LETTER E WITH TILDE
1EBE	Ế	LATIN CAPITAL LETTER E WITH CIRCUMFLEX AND ACUTE
1EBF	ế	LATIN SMALL LETTER E WITH CIRCUMFLEX AND ACUTE
1EC0	Ề	LATIN CAPITAL LETTER E WITH CIRCUMFLEX AND GRAVE
1EC1	ề	LATIN SMALL LETTER E WITH CIRCUMFLEX AND GRAVE
1EC4	Ễ	LATIN CAPITAL LETTER E WITH CIRCUMFLEX AND TILDE
1EC5	ễ	LATIN SMALL LETTER E WITH CIRCUMFLEX AND TILDE
1EC6	Ệ	LATIN CAPITAL LETTER E WITH CIRCUMFLEX AND DOT BELOW
1EC7	ệ	LATIN SMALL LETTER E WITH CIRCUMFLEX AND DOT BELOW
1EC8	Ỉ	LATIN CAPITAL LETTER I WITH HOOK ABOVE
1EC9	ỉ	LATIN SMALL LETTER I WITH HOOK ABOVE
1ECA	Ị	LATIN CAPITAL LETTER I WITH DOT BELOW
1ECB	ị	LATIN SMALL LETTER I WITH DOT BELOW

„“ des Bundesverwaltungsamts ist lizenziert unter einer Creative Commons Namensnennung 4.0 International Lizenz.

Unicode-Wert U+….	Glyph	Unicode-Zeichenname
1ECC	Ọ	LATIN CAPITAL LETTER O WITH DOT BELOW
1ECD	ọ	LATIN SMALL LETTER O WITH DOT BELOW
1ECE	Ỏ	LATIN CAPITAL LETTER O WITH HOOK ABOVE
1ECF	ỏ	LATIN SMALL LETTER O WITH HOOK ABOVE
1ED0	Ố	LATIN CAPITAL LETTER O WITH CIRCUMFLEX AND ACUT
1ED1	ố	LATIN SMALL LETTER O WITH CIRCUMFLEX AND ACUTE
1ED2	Ồ	LATIN CAPITAL LETTER O WITH CIRCUMFLEX AND GRAV
1ED3	ồ	LATIN SMALL LETTER O WITH CIRCUMFLEX AND GRAVE
1ED6	Ỗ	LATIN CAPITAL LETTER O WITH CIRCUMFLEX AND TILDE
1ED7	ỗ	LATIN SMALL LETTER O WITH CIRCUMFLEX AND TILDE
1ED8	Ộ	LATIN CAPITAL LETTER O WITH CIRCUMFLEX AND DOT BELOW
1ED9	ộ	LATIN SMALL LETTER O WITH CIRCUMFLEX AND DOT BELOW
1EDA	Ớ	LATIN CAPITAL LETTER O WITH HORN AND ACUTE
1EDB	ớ	LATIN SMALL LETTER O WITH HORN AND ACUTE
1EDC	Ờ	LATIN CAPITAL LETTER O WITH HORN AND GRAVE
1EDD	ờ	LATIN SMALL LETTER O WITH HORN AND GRAVE
1EE4	Ụ	LATIN CAPITAL LETTER U WITH DOT BELOW
1EE5	ụ	LATIN SMALL LETTER U WITH DOT BELOW
1EE6	Ủ	LATIN CAPITAL LETTER U WITH HOOK ABOVE
1EE7	ủ	LATIN SMALL LETTER U WITH HOOK ABOVE
1EE8	Ứ	LATIN CAPITAL LETTER U WITH HORN AND ACUTE
1EE9	ứ	LATIN SMALL LETTER U WITH HORN AND ACUTE
1EEA	Ừ	LATIN CAPITAL LETTER U WITH HORN AND GRAVE
1EEB	ừ	LATIN SMALL LETTER U WITH HORN AND GRAVE
1EEC	Ử	LATIN CAPITAL LETTER U WITH HORN AND HOOK ABOVE
1EED	ử	LATIN SMALL LETTER U WITH HORN AND HOOK ABOVE
1EEE	Ữ	LATIN CAPITAL LETTER U WITH HORN AND TILDE
1EEF	ữ	LATIN SMALL LETTER U WITH HORN AND

„“ des Bundesverwaltungsamts ist lizenziert unter einer Creative Commons Namensnennung 4.0 International Lizenz.

Unicode-Wert U+….	Glyph	Unicode-Zeichenname
		TILDE
1EF0	Ự	LATIN CAPITAL LETTER U WITH HORN AND DOT BELOW
1EF1	ự	LATIN SMALL LETTER U WITH HORN AND DOT BELOW
1EF2	Ỳ	LATIN CAPITAL LETTER Y WITH GRAVE
1EF3	ỳ	LATIN SMALL LETTER Y WITH GRAVE
1EF4	Ỵ	LATIN CAPITAL LETTER Y WITH DOT BELOW
1EF5	ỵ	LATIN SMALL LETTER Y WITH DOT BELOW
1EF6	Ỷ	LATIN CAPITAL LETTER Y WITH HOOK ABOVE
1EF7	ỷ	LATIN SMALL LETTER Y WITH HOOK ABOVE
1EF8	Ỹ	LATIN CAPITAL LETTER Y WITH TILDE
1EF9	ỹ	LATIN SMALL LETTER Y WITH TILDE
20AC	€	EURO SIGN

Tabelle 2: Zusätzliche Zeichen

„“ des Bundesverwaltungsamts ist lizenziert unter einer Creative Commons Namensnennung 4.0 International Lizenz.

7. Quellenverzeichnis

[DGSG]

Oracle: Globalization Support Guide 10g Release 2 (10.2) Dezember 2005. http://download.oracle.com/docs/cd/B19306_01/server.102/b14225.pdf.

[SAGA40]

SAGA Version 4.0 – Standards und Architekturen für E-Government-Anwendungen; Publikation der KBSt; März 2008.
http://www.kbst.bund.de/saga.

[SLMapping]

Mapping-Tabelle für die Transformation von String.Latin-Zeichen 10_Bausteine/Sonderzeichen/Mappingtabelle.xls.

[Transskriptionsregeln]

Transkriptions-Regeln 10_Bausteine/Sonderzeichen/Transkriptionsregeln.xls.

[XOEVStringLatin]

Handbuch zur Entwicklung XÖV-konformer IT-Standards (Anhang A) http://www.xoev.de/sixcms/media.php/13/2010-03-02-Handbuch-final.pdf . (Zugriff am 11.12.2014).

„“ des Bundesverwaltungsamts ist lizenziert unter einer Creative Commons Namensnennung 4.0 International Lizenz.

8. Abbildungsverzeichnis

Abbildung 1: Datentyp für umgeschriebene Texte.....13
Abbildung 2: Komponente Transkription.....14

„“ des Bundesverwaltungsamts ist lizenziert unter einer Creative Commons Namensnennung 4.0 International Lizenz.

9. Tabellenverzeichnis

Tabelle 1: Attribute des Datentyps „TransText“ 14

Tabelle 2: Zusätzliche Zeichen 34