Internationalisierung bei der Software-Entwicklung

Thomas Tampe

Leiter Solution Center
Logics Software GmbH
München

Code once, run anywhere …

… in jeder Sprache und in jedem Land …

- in Deutsch oder Französisch …
- in England oder den USA …
- in Türkisch oder Polnisch …
- in Griechenland oder Bulgarien …
- in Hebräisch oder Farsi …
- in Japan oder China …

Agenda

- Begriffe und Definitionen
- Zeichensätze – Historie und Grundlagen
- Internationalisierung und Lokalisierung
- Anwendungsarchitekturen
  - Checkliste
  - Werkzeuge
  - Beispiel
- Quellen und Links

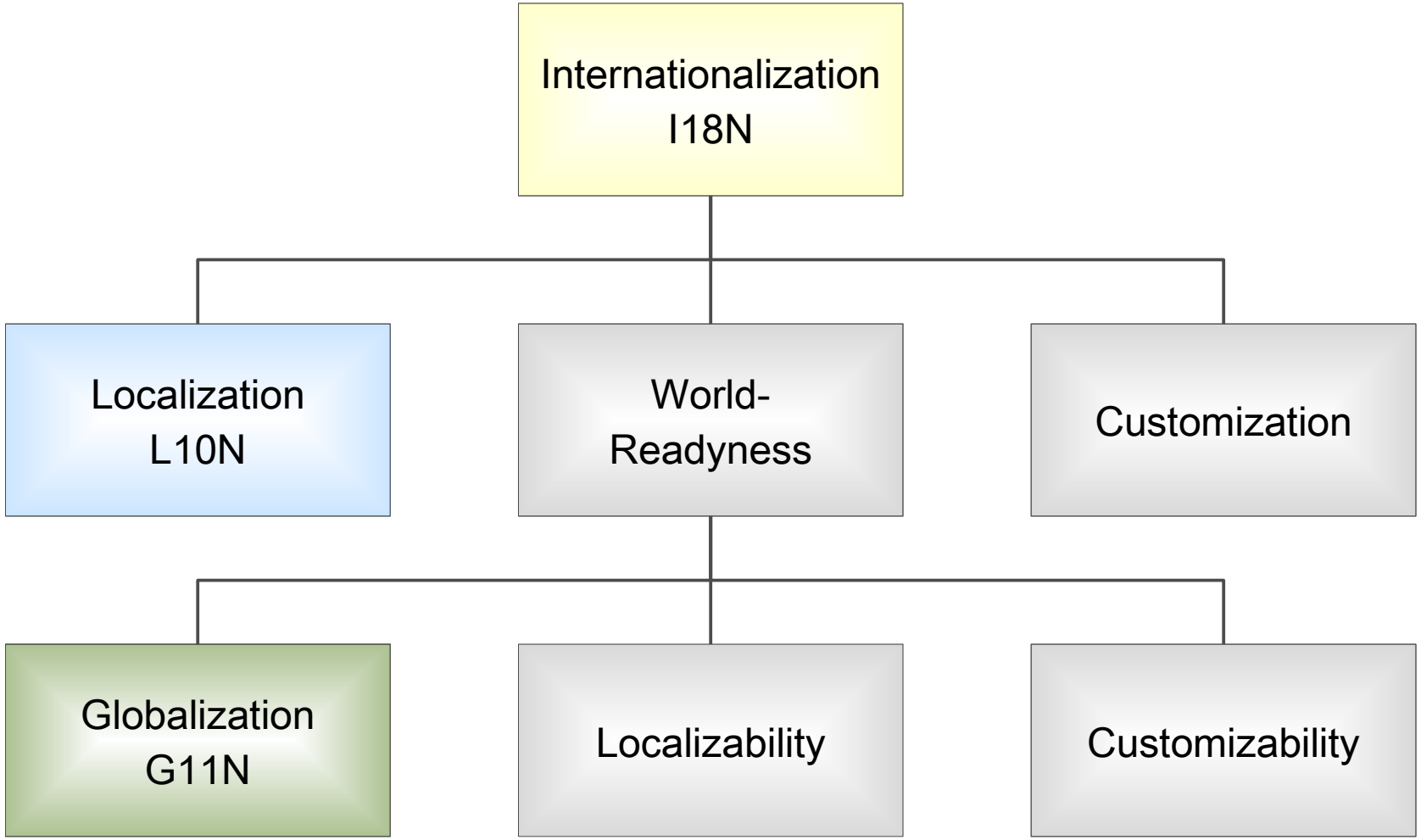
Agenda

- Begriffe und Definitionen
- Zeichensätze – Historie und Grundlagen
- Internationalisierung und Lokalisierung
- Anwendungsarchitekturen
  - Checkliste
  - Werkzeuge
  - Beispiel
- Quellen und Links

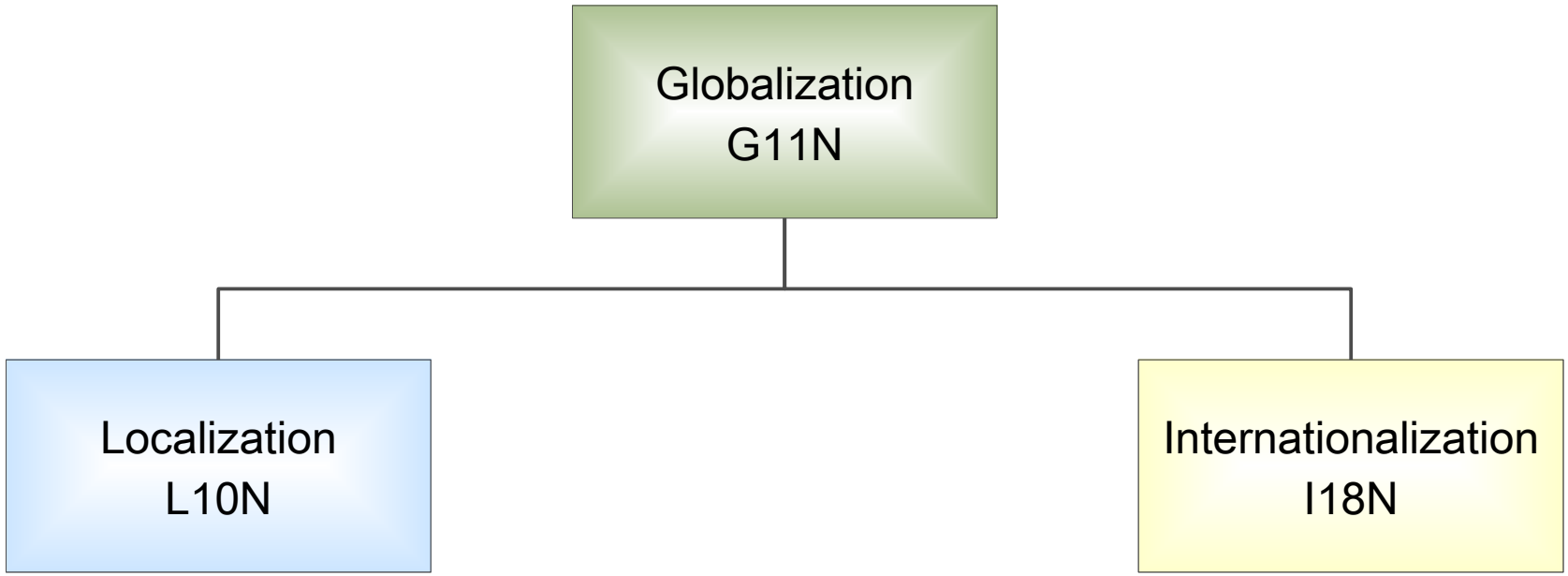
Begriffe und Definitionen

- Internationalisierung
  - I18N = InternationalizatioN
  - Grundlage der Lokalisierung
- Lokalisierung
  - L10N = LocalizatioN
  - Anpassung an individuelle Sprache und Region
- Globalisierung
  - G11N = GlobalizatioN

Microsoft ist anders ...



… als alle anderen



Agenda

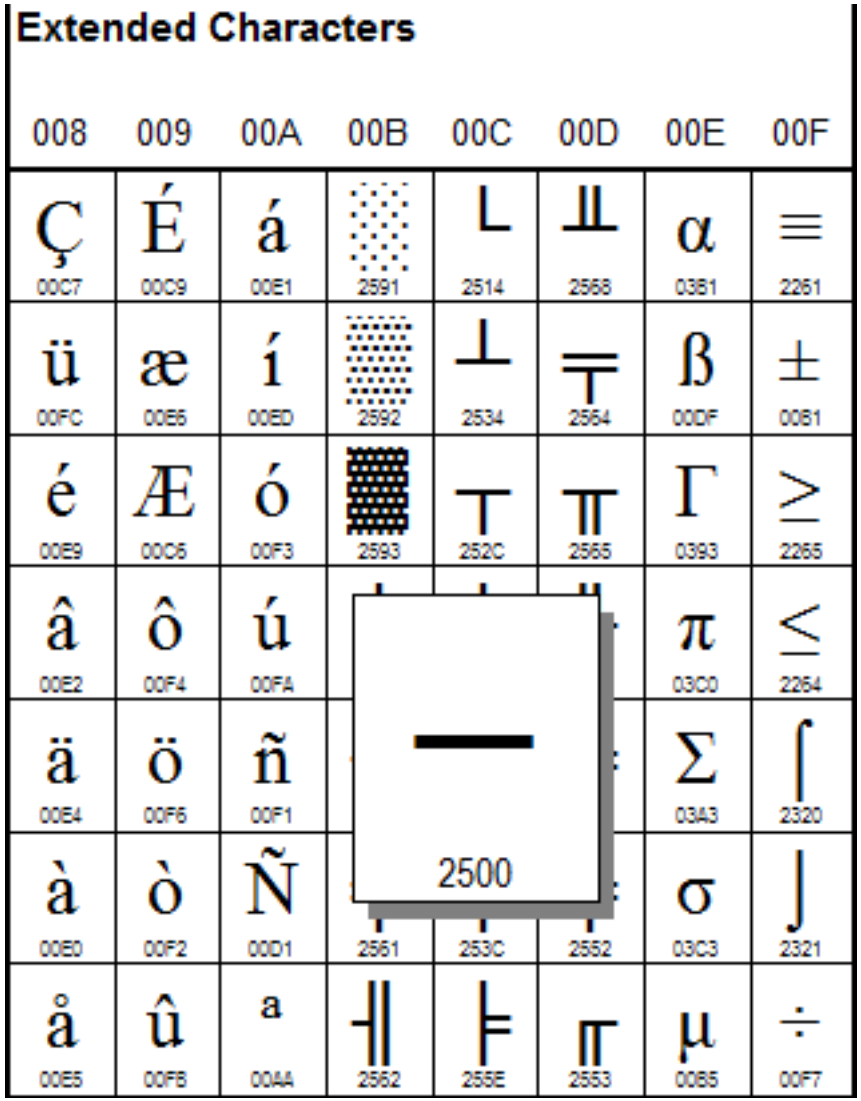
- Begriffe und Definitionen
- Zeichensätze – Historie und Grundlagen
- Internationalisierung und Lokalisierung
- Anwendungsarchitekturen
  - Checkliste
  - Werkzeuge
  - Beispiel
- Quellen und Links

Zeichensätze – Historie

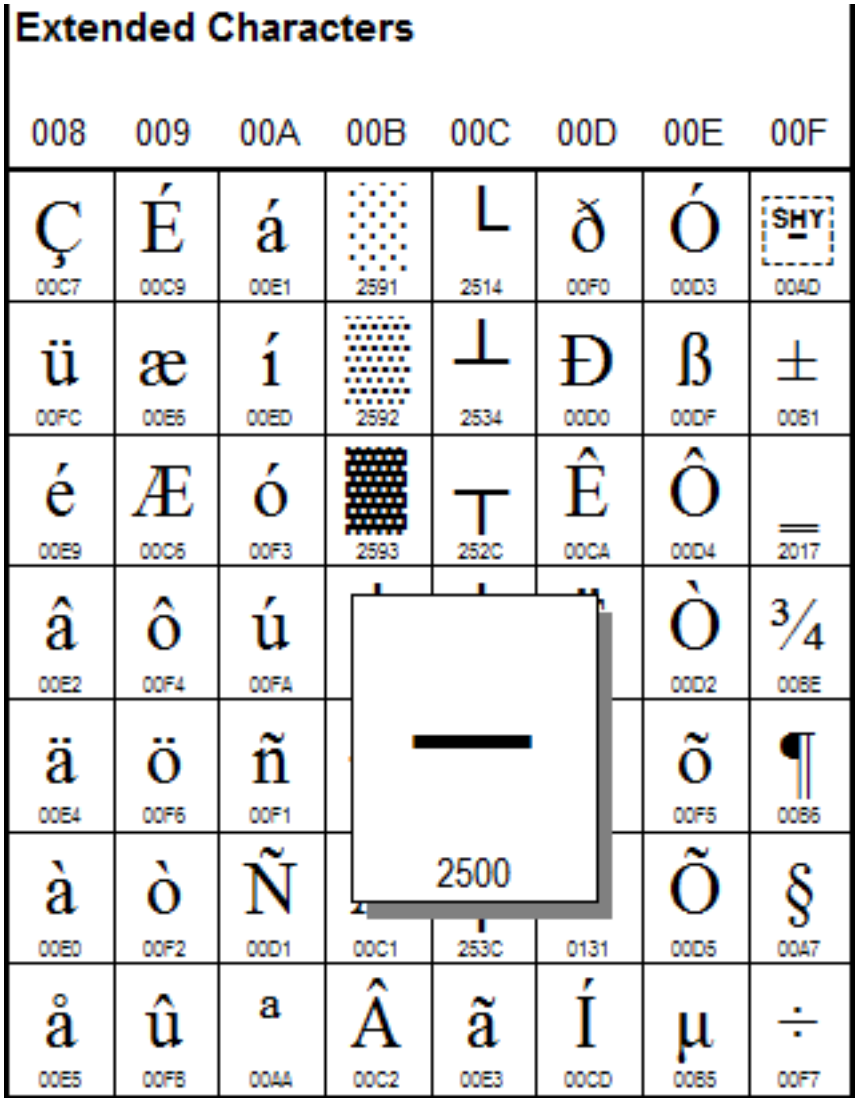
- ASCII (7 Bit)
  - Nur für englische Texte geeignet
- Ein Byte = Ein Zeichen
  - OEM Codepages
  - ANSI Standards
  - ISO-8859-X
- Unicode
  - Code-Points (U+0000 … U+10FFFF)
  - Oft eingeschränkt auf U+0000 … U+FFFF

OEM Codepages

437 (OEM – Vereinigte Staaten)



850 (OEM – Multilingual Lateinisch 1)



ANSI Standards

1252 (ANSI – Lateinisch I)

Extended Characters

008	009	00A	00B	00C	00D	00E	00F
€ 20AC	DCS 0090	NB SP 00A0	° 00B0	À 00C0	Đ 00D0	à 00E0	ð 00F0
XXX 0081	‘ 2018	¡ 00A1	± 00B1	Á 00C1	Ñ 00D1	á 00E1	ñ 00F1
‚ 201A	’ 2019	¢ 00A2	² 00B2	Â 00C2	Ò 00D2	â 00E2	ò 00F2
ƒ 0192	“ 201C	£ 00A3				ã 00E3	ó 00F3
„ 201E	” 201D	¤ 00A4				ä 00E4	ô 00F4
… 2026	• 2022	¥ 00A5				å 00E5	õ 00F5
† 2020	– 2013	¦ 00A6	¶ 00B6	Æ 00C6	Ö 00D6	æ 00E6	ö 00F6

Ä 00C4

1251 (ANSI – Kyrillisch)

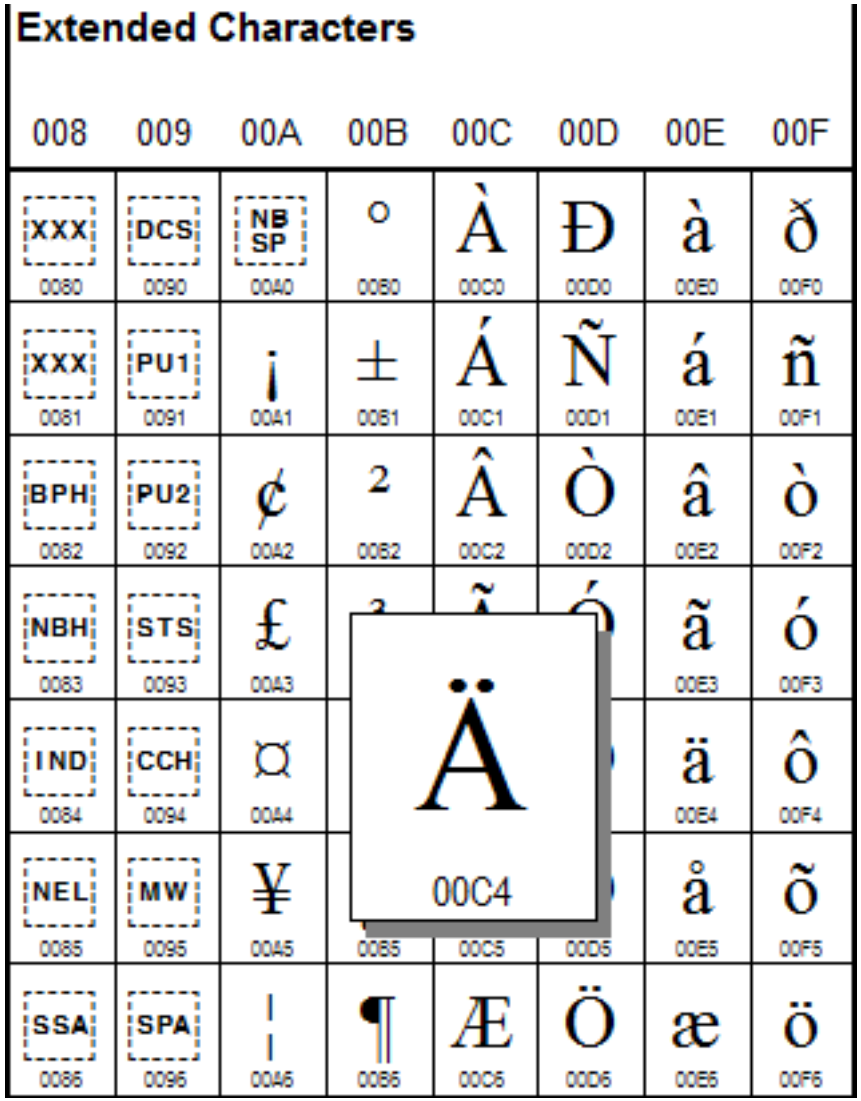
Extended Characters

008	009	00A	00B	00C	00D	00E	00F
Ъ 0402	ђ 0452	NB SP 00A0	° 00B0	А 0410	Р 0420	а 0430	р 0440
Ѓ 0403	‘ 2018	Ў 040E	± 00B1	Б 0411	С 0421	б 0431	с 0441
‚ 201A	’ 2019	ў 045E	І 0406	В 0412	Т 0422	в 0432	т 0442
ѓ 0453	“ 201C	Ј 0408				г 0433	у 0443
„ 201E	” 201D	¤ 00A4				д 0434	ф 0444
… 2026	• 2022	Ґ 0490		Е 0415	Х 0425	е 0435	х 0445
† 2020	– 2013	¦ 00A6	¶ 00B6	Ж 0416	Ц 0426	ж 0436	ц 0446

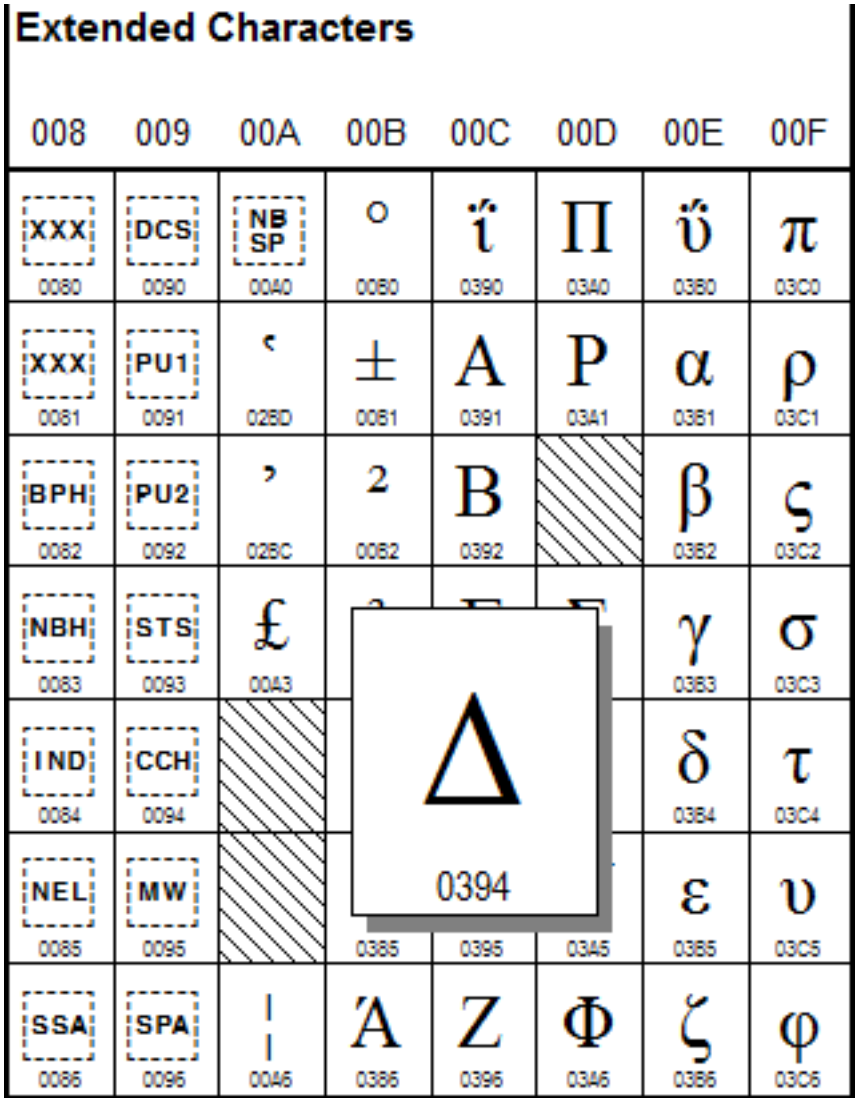
Д 0414

ISO-8859-X

ISO 8859-1 (Lateinisch I)



ISO 8859-7 (Griechisch)



Unicode

Control		Latin-1 Supplement					
008	009	00A	00B	00C	00D	00E	00F
XXX 0080	DCS 0090	NB SP 00A0	° 00B0	À 00C0	Đ 00D0	à 00E0	ð 00F0
XXX 0081	PU1 0091	¡ 00A1	± 00B1	Á 00C1	Ñ 00D1	á 00E1	ñ 00F1
BPH 0082	PU2 0092	¢ 00A2	² 00B2	Â 00C2	Ò 00D2	â 00E2	ò 00F2
NBH 0083	STS 0093	£ 00A3				ã 00E3	ó 00F3
IND 0084	CCH 0094	¤ 00A4				ä 00E4	ô 00F4
NEL 0085	MW 0095	¥ 00A5	0085	00C5	00D5	å 00E5	õ 00F5
SSA 0086	SPA 0096	¦ 00A6	¶ 00B6	Æ 00C6	Ö 00D6	æ 00E6	ö 00F6

Ä
00C4

Cyrillic							
040	041	042	043	044	045	046	047
È 0400	А 0410	Р 0420	а 0430	р 0440	è 0450	Ѡ 0460	Ψ 0470
Ё 0401	Б 0411	С 0421	б 0431	с 0441	ё 0451	ѡ 0461	ψ 0471
Ђ 0402	В 0412	Т 0422	в 0432	т 0442	ђ 0452	Ѣ 0462	Ѳ 0472
Ѓ 0403	Г 0413	У 0423				ѣ 0463	ѳ 0473
Є 0404	Д 0414	Ф 0424				Ѥ 0464	Ѵ 0474
S 0405	Е 0415	Х 0425	0435	0445	0455	ѥ 0465	ѵ 0475
І 0406	Ж 0416	Ц 0426	ж 0436	ц 0446	і 0456	Ѧ 0466	Ѷ 0476

ф
0444

Grundlagen

Darstellung	Zeichensatz	Wert	Bedeutung
─	437 OEM	**0xC4**	Blockgrafik – einfache horizontale Linie
Д	1251 ANSI	**0xC4**	Kyrillischer Großbuchstabe DE
Δ	ISO-8859-7	**0xC4**	Griechischer Großbuchstabe Delta
Ä	ISO-8859-1	**0xC4**	Lateinischer Großbuchstabe A Umlaut
A	Unicode	U+0040	Lateinischer Großbuchstabe A
Α	Unicode	U+0391	Griechischer Großbuchstabe Alpha
А	Unicode	U+0410	Kyrillischer Großbuchstabe A

- Zeichensatz und Wert legen Bedeutung fest
- Codierung dient Speicherung und Transport
- Font (Schriftart) stellt Zeichen dar
- There Ain't No Such Thing As Plain Text. (Joel Spolsky)

Beispiel – Codierung

- Erwartete Textdarstellung

  l'élection présidentielle française

- Text codiert als UTF-8 aber als ISO-8859-1 angezeigt

  lâ€™Ã©lection prÃ©sidentielle franÃ§aise

- Text codiert als ISO-8859-1 aber als UTF-8 angezeigt

  l?鎖ction pr鳩dentielle fran誓se

http://www.alanwood.net/unicode/htmlunicode.html

Unicode / UTF–8

- Aktueller Standard
  - Derzeit 17 * 2^{16} = 1.114.112 Zeichen
  - Zeichen U+0000 … U+00FF identisch zu ISO-8859-1
- Codierung UTF–8
  - maximal 2^{31} Zeichen
  - Variable Länge (1 – 6 Byte pro Zeichen)

Zeichen	Buchstabe y	Buchstabe Ä	Zeichen ©	Eurozeichen €
Unicode	U+0079	U+00C4	U+00A9	U+20AC
Unicode binär	00000000 01111001	00000000 11000100	00000000 10101001	00100000 10101100
UTF-8 binär	01111001	11000011 10000100	11000010 10101001	11100010 10000010 10101100
UTF-8	0x79	0xC3 0x84	0xC2 0xA9	0xE2 0x82 0xAC

UTF-8

- Kompatibel mit 7-Bit ASCII
  - Codierung als einzelnes Byte
  - NUL (0x00) markiert das Ende einer Zeichenreihe
- Überprüfbare Codierung
  - Teilsequenzen können erkannt werden
  - Anfang einer Sequenz ist ermittelbar
- Unterstützung in modernen Betriebssystemen
  - Windows (NT, 2000, XP, Vista, 7)
  - Unix (Solaris, AIX, HP-UX, …)
  - Linux (Debian, Ubuntu, SUSE, …)

Agenda

- Begriffe und Definitionen
- Zeichensätze – Historie und Grundlagen
- Internationalisierung und Lokalisierung
- Anwendungsarchitekturen
  - Checkliste
  - Werkzeuge
  - Beispiel
- Quellen und Links

Internationalisierung

- Intern Unicode (U+0000 … U+FFFF)
  - Transport und Speicherung als UTF-8
- Texte zur Laufzeit laden
  - Verschiedene Pluralformen
  - Platzhalter für Werte
  - Anzeigeformate für Datum / Uhrzeit, Zahlen, …
- Medien zur Laufzeit laden
  - Symbole
  - Audio- und Video-Clips

Internationalisierung

- Eingabemethoden
  - Tastaturkürzel für Menübefehle
  - Unabhängigkeit vom Tastaturlayout
- Layout der Benutzeroberfläche
  - Unterschiedliche Länge der Texte
  - Schreibrichtung (Bidirektionale Texte)
- Druckausgaben
  - Installierte Druckertreiber
  - Unterstützte Papierformate

Internationalisierung

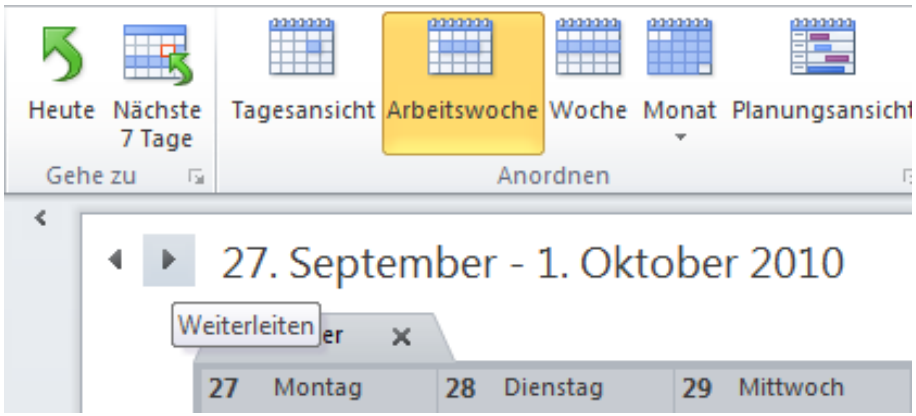
- Interpretation der Eingabedaten
  - Gibt es auch andere Ziffern als 0 … 9 ?
  - Was sind Großbuchstaben / Kleinbuchstaben ?
  - Wo ist das Ä einzusortieren ?
  - Ist ¼ das gleiche wie 1/4 ?
- Bedingte Übersetzung für grundlegende Unterschiede eventuell sinnvoll
  - bidirektionale Sprachen (Hebräisch, Arabisch, …)
  - ostasiatische Sprachen (Chinesisch, Japanisch, …)

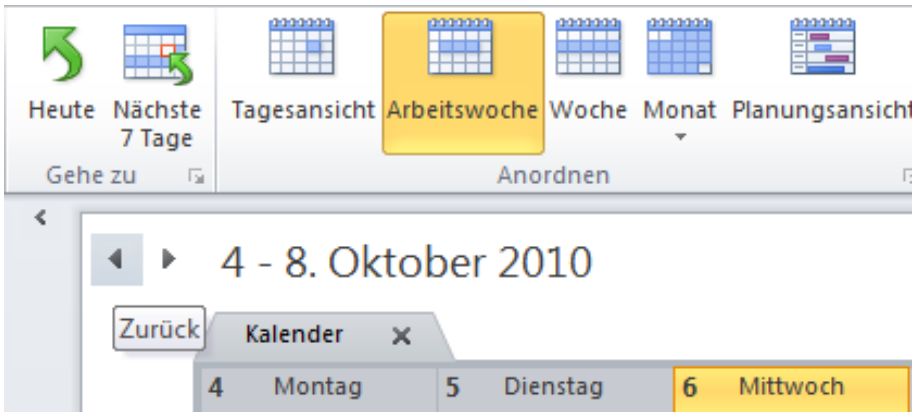
Lokalisierung

- Für jede Locale (Culture) separat
- Übersetzen der (externen) Texte
- Mediale Inhalte bereitstellen
- Falls nicht im Laufzeitsystem verfügbar
  - Formate für Zeit / Datum, Zahlen, …
  - Währungssymbole
  - Umrechnungsfunktionen für Maßeinheiten
  - Algorithmen für Sortierung
  - Umwandlung Großbuchstaben / Kleinbuchstaben

Lokalisierung

- Setzt sorgfältige Internationalisierung voraus
  - Beispiel aus Microsoft Outlook 2010





  - Kontextsensitive Bezeichnungen verwenden
  - Konsistente Datumsformate definieren

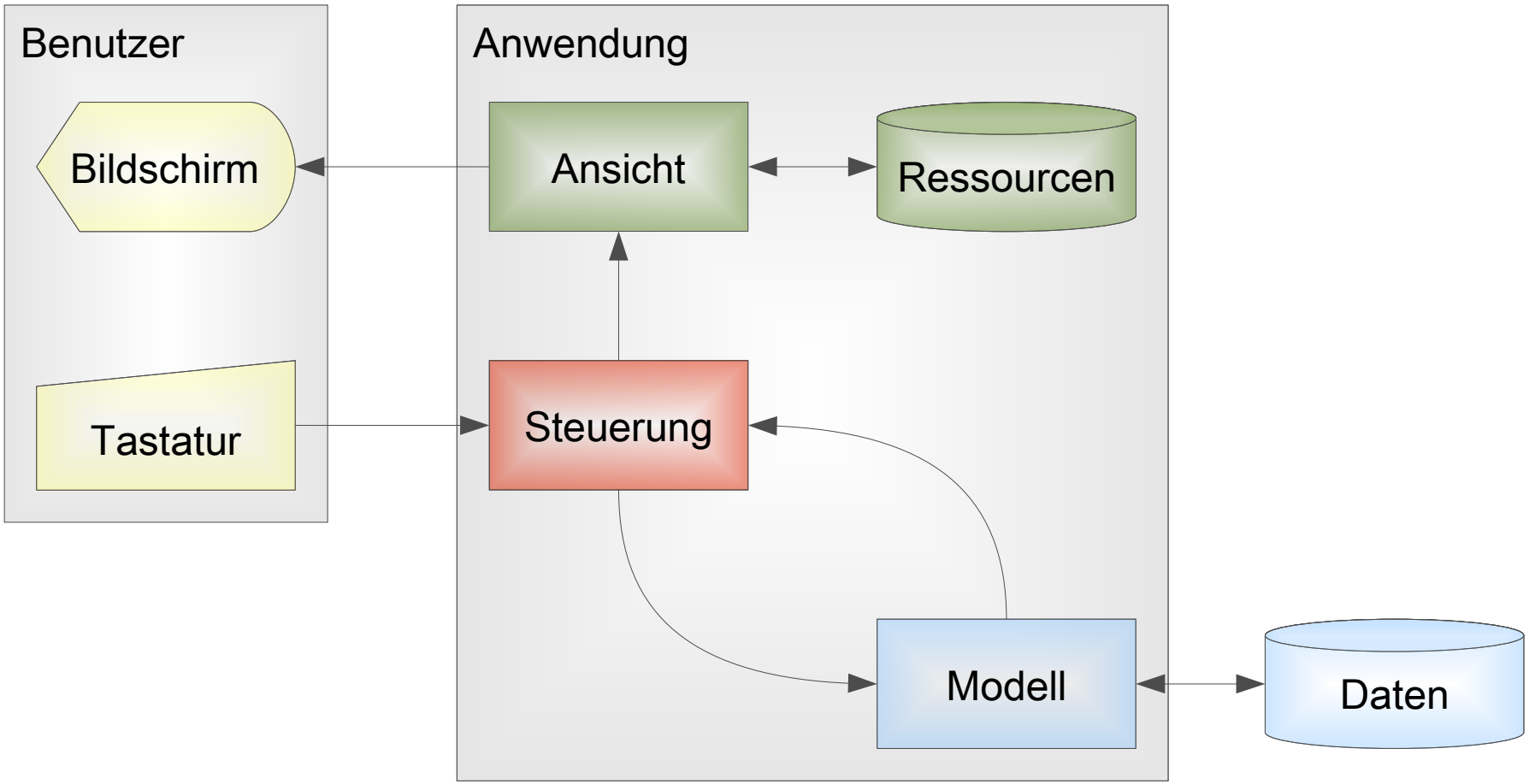
Agenda

- Begriffe und Definitionen
- Zeichensätze – Historie und Grundlagen
- Internationalisierung und Lokalisierung
- Anwendungsarchitekturen
  - Checkliste
  - Werkzeuge
  - Beispiel
- Quellen und Links

Model – View – Controller (MVC)

- Trennt Darstellung von Anwendungslogik
- Controller
  - Koordiniert Model und View
- Model
  - Benötigt keine Information zu Sprache / Region
  - Verarbeitet abstrakte Daten
- View
  - Kennt Sprache und Region des Anwenders
  - Stellt abstrakte Daten dar

Desktop Anwendung



Datenbankschnittstelle

- Einzige externe Schnittstelle
- Zugriff über Persistenz-Frameworks
  - TopLink, Hibernate (Java)
  - ADO.NET Entity Framework, NHibernate (.NET)
- Automatische Konvertierung in Objekte
- Entwurfswerkzeuge
  - NetBeans, Eclipse (Java)
  - Visual Studio 2010 (.NET)
- Kein Thema für Internationalisierung

Desktop Anwendung

- Sprache und Region aus Benutzerumgebung
  - Locale (Java) oder Culture (.NET)
    - de_DE ist Deutsch (Deutschland)
    - es_US wäre Spanisch (Vereinigte Staaten)
  - Sprache wichtiger als Region
- Tastaturbelegung oft umschaltbar
  - Symbolische Tastennamen verwenden
  - mögliche Tastenkombinationen berücksichtigen

Desktop Anwendung – Checkliste

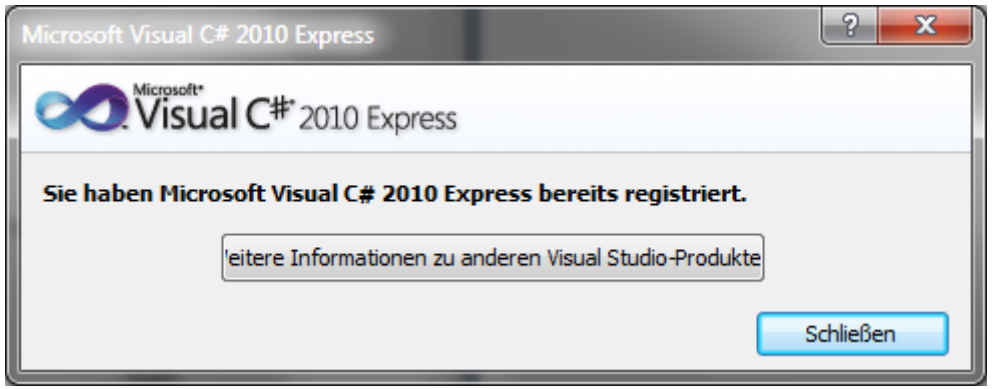
- Fenster
  - Titelzeile
- Menüs
  - Titel und Einträge
  - Tastaturkürzel
- Symbolleisten
  - Symbole
  - Texte für Info-Fenster
- Statuszeile
- Dialogfelder
  - Meldungstexte
- Formate
  - Zeit / Datum
  - Zahlen, Währung
- Kontrollelemente
  - Beschriftungen
  - Inhalte
- Klänge und Bilder

Desktop Anwendung – Werkzeuge

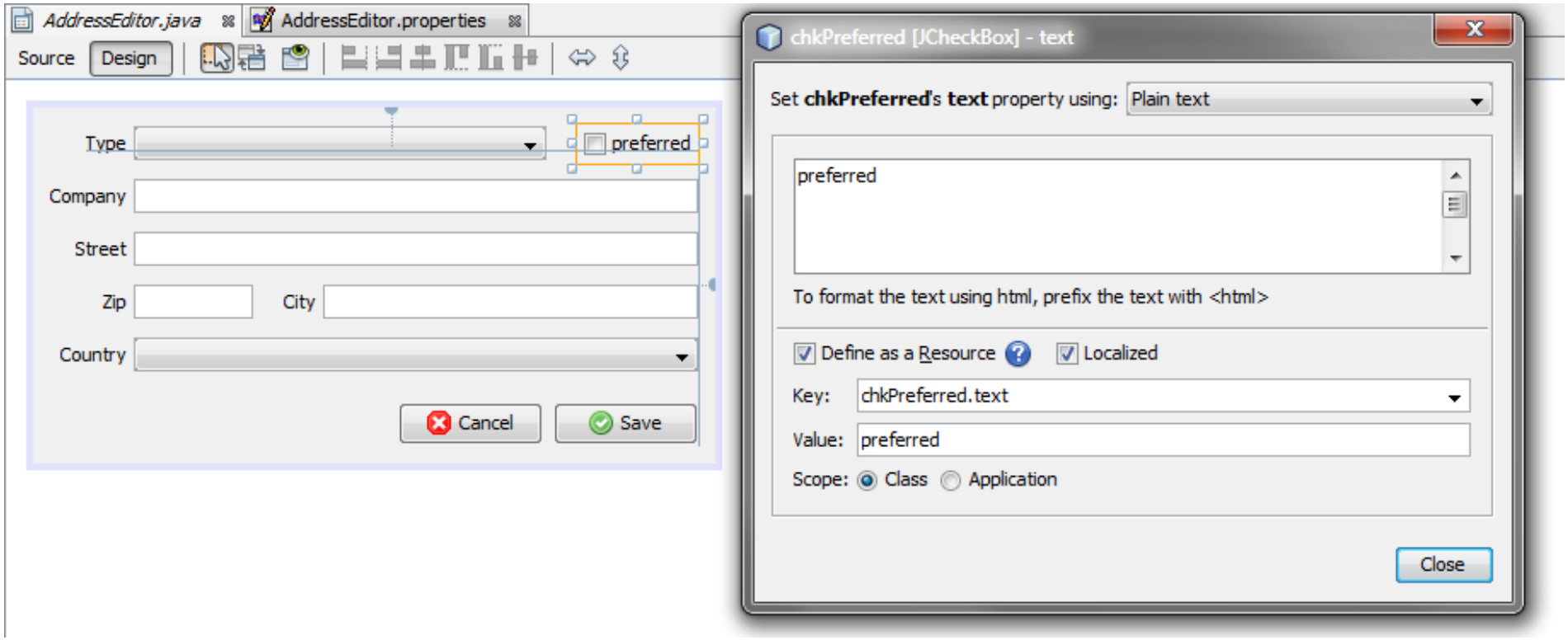
- Integrierte Entwicklungsumgebungen
  - NetBeans, Eclipse (Java)
  - Visual Studio (.NET)
  - Ressourcen automatisch beim Entwurf anlegen
    - *.properties oder ResourceBundle (Java)
    - *.resx, *.resource (.NET)
- Ressource Editoren
  - Windows Forms Resource Editor (.NET)

Desktop Anwendung – Layouts

- Flexible Layouts
  - Größe zur Laufzeit an Inhalt anpassen
- Feste Layouts
  - Größe zum Übersetzungszeitpunkt an Inhalt anpassen



Desktop Anwendung – NetBeans

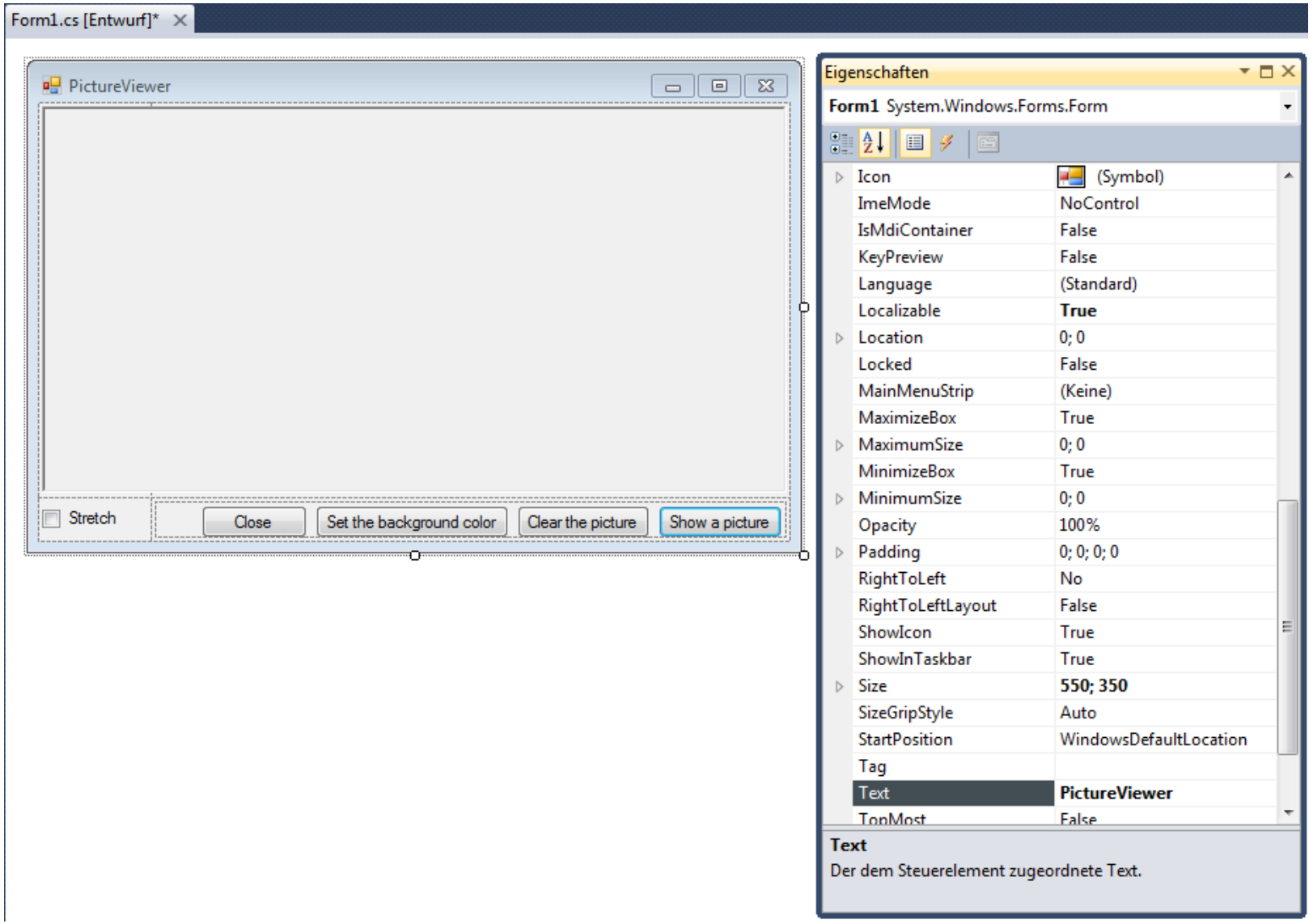


Desktop Anwendung – NetBeans

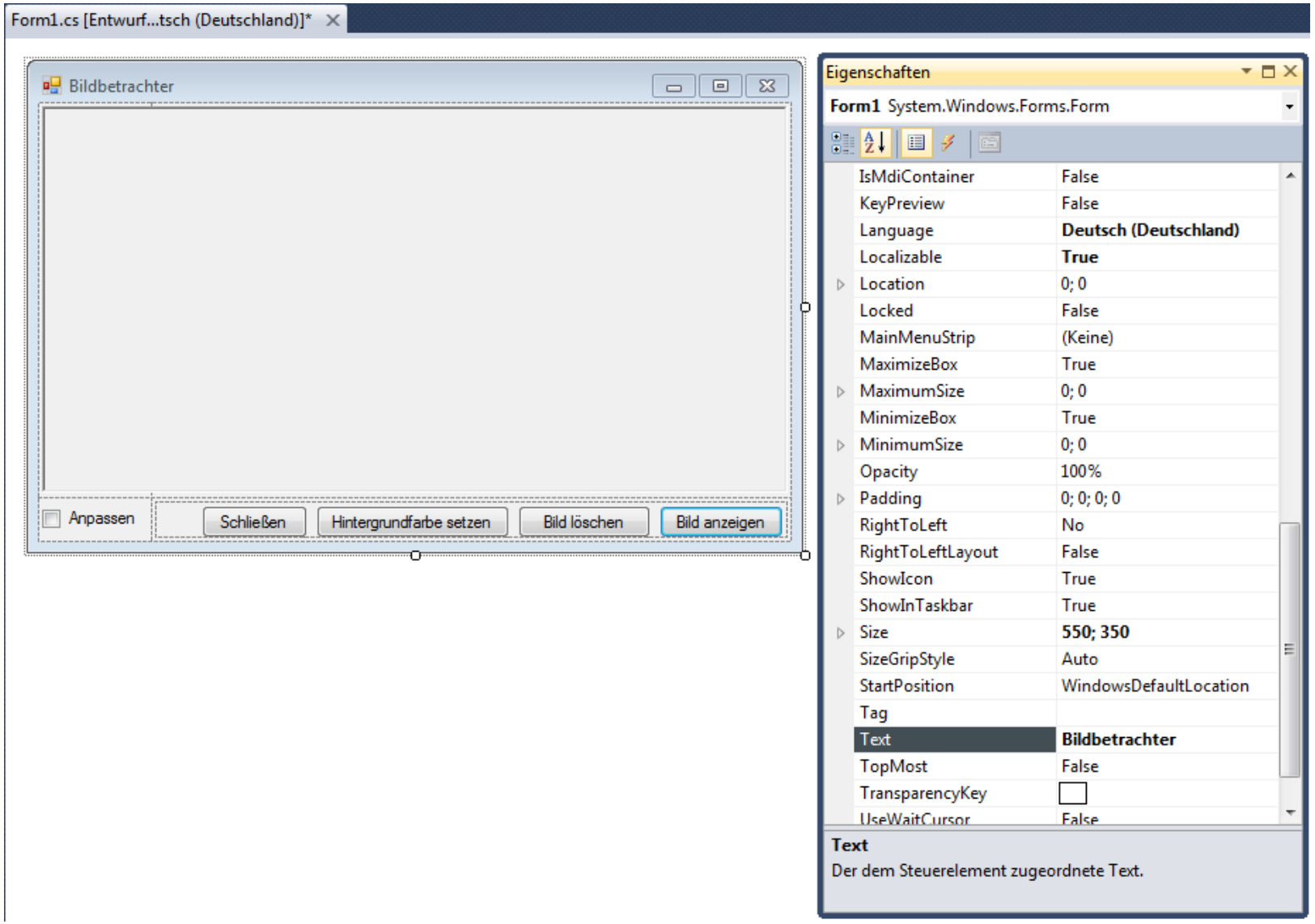
AddressEditor.java | AddressEditor.properties

Key	default language	de - Deutsch
addressEditor.title	Edit Address details	Adressdetails bearbeiten
lblType.text	Type	Typ
chkPreferred.text	preferred	bevorzugt
lblCompany.text	Company	Firma
lblStreet.text	Street	Straße
lblZipcode.text	Zip	PLZ
lblCity.text	City	Ort
lblCountry.text	Country	Land
save.Action.text	Save	Speichern
save.Action.shortDescription	Save the Address data	Speichert die Adresse
save.Action.accelerator	pressed ENTER	pressed ENTER
save.Action.icon	/de/tampe/address/resources/accept.png	/de/tampe/address/resources/accept.png
save.Action.smallIcon	/de/tampe/address/resources/accept.png	/de/tampe/address/resources/accept.png
cancel.Action.text	Cancel	Abbrechen
cancel.Action.shortDescription	Cancel editing the Address data	Bricht die Bearbeitung der Adressdetai...
cancel.Action.accelerator	pressed ESCAPE	pressed ESCAPE
cancel.Action.icon	/de/tampe/address/resources/cancel.png	/de/tampe/address/resources/cancel.png
cancel.Action.smallIcon	/de/tampe/address/resources/cancel.png	/de/tampe/address/resources/cancel.png

Desktop Anwendung - Visual Studio



Desktop Anwendung - Visual Studio



Beispiel – Markieren und Kopieren

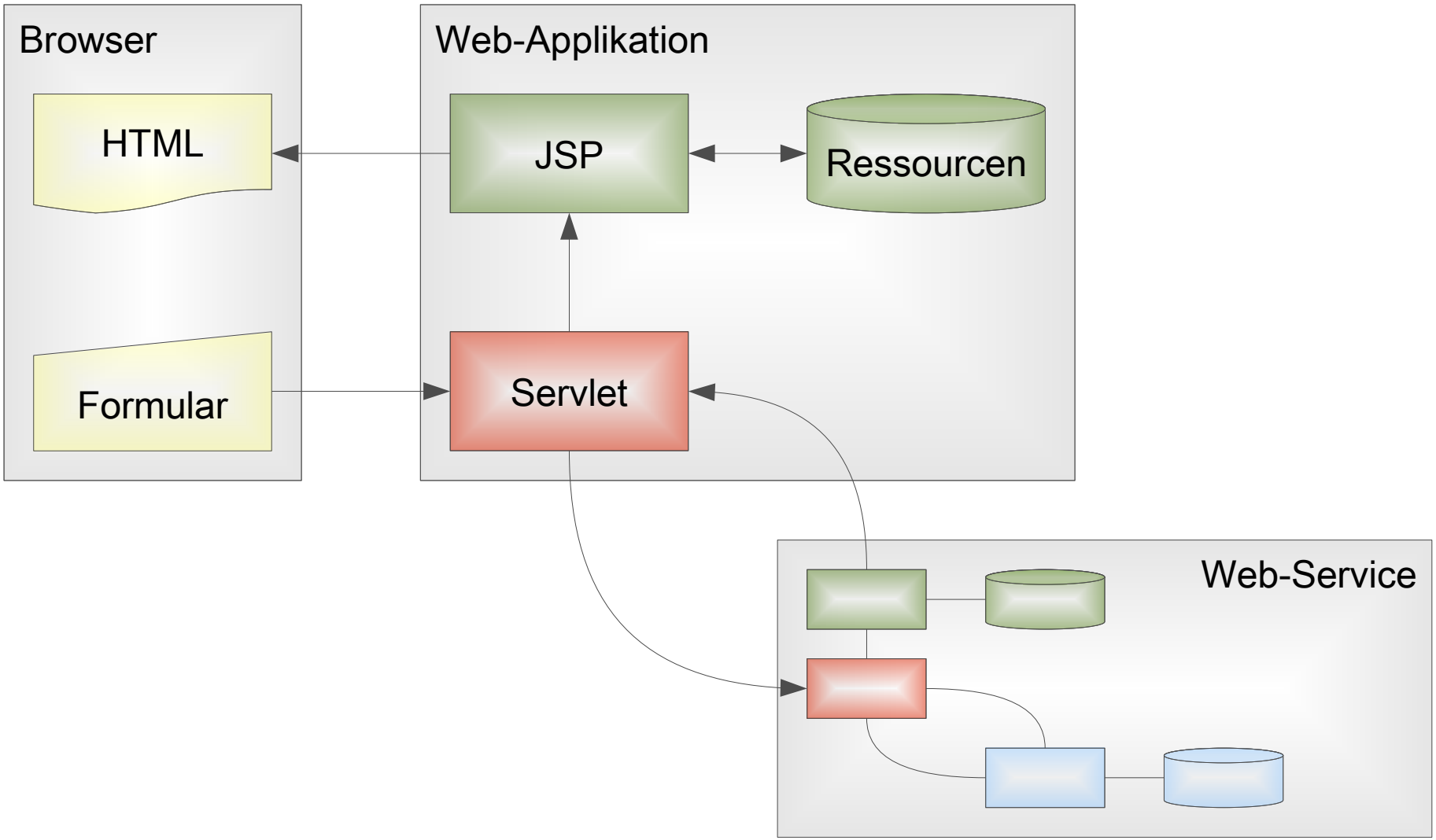
Normaler Text	Buchstaben und Ziffern LTR
Zeichenreihe	ABCD 1234 EFGH
Darstellung	ABCD 1234 EFGH
Zeichenreihe	ABCD 1234 EFGH
Zu kopierender Text	CD 1234 EF

Bidirektionaler Text	Buchstaben RTL, Ziffern LTR
Zeichenreihe	ABCD 1234 EFGH
Darstellung	DCBA 1234 HGFE
Zeichenreihe	ABCD 1234 EFGH
Zu kopierender Text	AB 1234 GH

Agenda

- Begriffe und Definitionen
- Zeichensätze – Historie und Grundlagen
- Internationalisierung und Lokalisierung
- Anwendungsarchitekturen
  - Checkliste
  - Werkzeuge
  - Beispiel
- Quellen und Links

Verteilte Anwendung



HTTP – Schnittstellen

- Client teilt Codierung mit
  - Accept-Charset: utf-8
  - <form accept-charset="utf-8" …>
- Server teilt Codierung und Typ mit
  - HTML-Dokumente
  - Content-Type: text/html; charset=utf-8
  - <meta http-equiv="content-type" content="text/html; charset=utf-8">
  - XML-Dokumente
  - Content-Type: application/xml; charset=utf-8
  - <?xml version="1.0" encoding="UTF-8" …>

Verteilte Anwendung

- Bevorzugte Sprache und Region aus HTTP-Header oder Parameter
  - Accept-Language: de-DE, en-US;q=0.5
  - … &lang=de-DE …
- Service-Aufruf aus URL und / oder Parametern
  - http://server.domain/controller/service.jsp?param1=value1& …
- Service-spezifische JSP / ASP aus Sprache und Region

Verteilte Anwendung – Checkliste

- HTML-Seiten
  - Texte
- Stylesheets
  - Symbole
- Scripts
  - Meldungstexte
  - Tastaturkürzel
- Klänge
- Bilder
- Ressourcen
  - Texte
  - Zeit- / Datums- und Zahlformate
- Frameworks

Verteilte Anwendung – Werkzeuge

- Integrierte Entwicklungsumgebungen
  - NetBeans, Eclipse, Visual Studio
- Editor für HTML / JSP / ASP
  - DreamWeaver, Kompozer, …
  - Nur sichtbare Teile der Seite übersetzen
- Programm zum Dateivergleich
  - ExamDiff, WinMerge, …
  - Strukturänderungen der Seite übernehmen
- Windows Forms Resource Editor (.NET)

Beispiel – Online-Shop

- Bestellung per Internet auch aus dem Ausland
- Rechnung mit ausgewiesener Umsatzsteuer

- Model liefert nur Netto-Betrag
- View berechnet Steuer- und Brutto-Betrag

- Controller liefert Sprache und Region
- Model liefert Netto-, Steuer- und Brutto-Betrag
- Land und Steuersatz gehören zum Model

- Rechnung wird von View erstellt
  - Text, Datumsformate, Währungsformate, ...

Agenda

- Begriffe und Definitionen
- Zeichensätze – Historie und Grundlagen
- Internationalisierung und Lokalisierung
- Anwendungsarchitekturen
  - Checkliste
  - Werkzeuge
  - Beispiel
- Quellen und Links

Quellen und Links

Unicode and Character Sets http://www.joelonsoftware.com/articles/Unicode.html
UTF-8, a transformation format of ISO 10646 http://www.ietf.org/rfc/rfc3629.txt
Unibook™ Character Browser http://unicode.org/unibook/
Encoding and Localization http://msdn.microsoft.com/en-us/library/h6270d0z(v=VS.100).aspx
Microsoft .NET Internationalization http://msdn.microsoft.com/en-us/goglobal/bb688096.aspx
ASP.NET Wiki: Internationalization http://wiki.asp.net/page.aspx/55/internationalization/
ADO.NET Entity Framework http://msdn.microsoft.com/en-us/library/bb399572.aspx
Open Source Persistence Frameworks in C# http://csharp-source.net/open-source/persistence
Robust internationalization with GTK+ http://www.ibm.com/developerworks/aix/library/au-internatlgtk/index.html
Create Great-Looking GUIs With NetBeans IDE 5.5 http://java.sun.com/developer/technicalArticles/tools/nb_guibuilder/

An old joke in the internationalization community goes like this:

A person who speaks three languages is called trilingual.
And a person who speaks two languages is called bilingual.
So what do you call someone who only speaks one language?

American.

http://www.eclipse.org/articles/Article-Internationalization/how2I18n.html

Thomas Tampe

thomas.tampe@arcor.de
+49 (172) 9 57 17 05

t.tampe@logics.de
+49 (89) 55 24 04 32