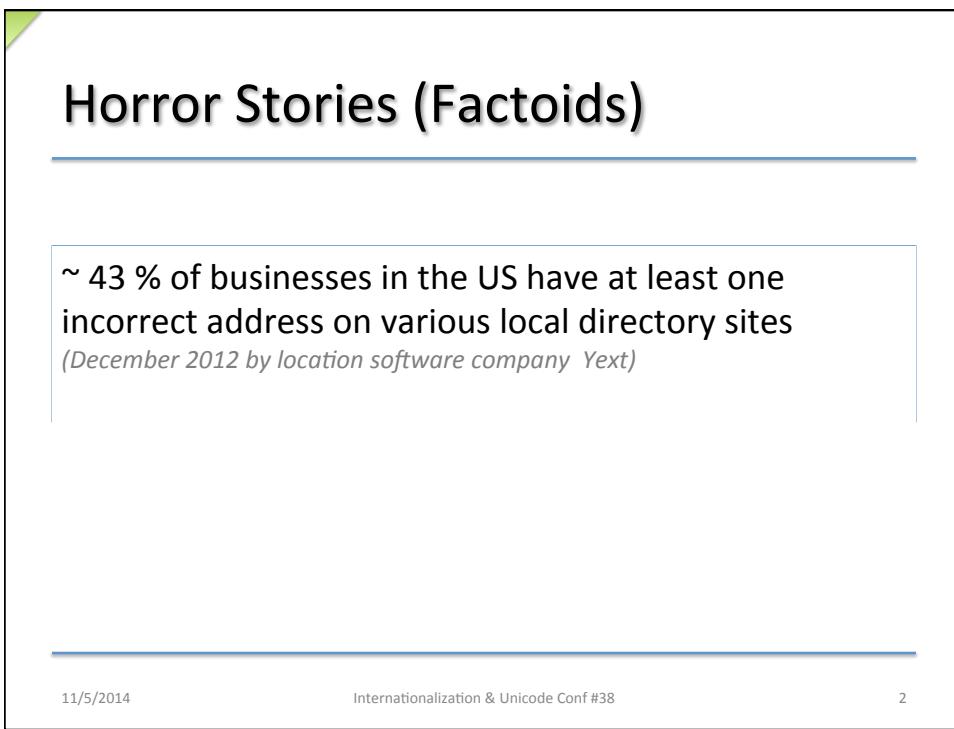


10 DOWNING STREET  
LONDON SW1A 2AA

ELM ST.

## Addresses, Nightmares and Standards

Erwin Hom, Aarti Ashok  
PayPal



## Horror Stories (Factoids)

---

~ 43 % of businesses in the US have at least one incorrect address on various local directory sites  
*(December 2012 by location software company Yext)*

---

11/5/2014      Internationalization & Unicode Conf #38      2

## Where PayPal does business

### Africa

	Algeria		Angola		Benin		Botswana		Burkina Faso
	Burundi		Cameroon		Cape Verde		Chad		Comoros
	Côte d'Ivoire		Democratic Republic of the Congo		Djibouti		Egypt		Eritrea
	Ethiopia		Gabon Republic		Gambia		Guinea		Guinea-Bissau
	Kenya		Lesotho		Madagascar		Malawi		Mali
	Mauritania		Mauritius		Mayotte		Morocco		Mozambique
	Namibia		Niger		Nigeria		Republic of the Congo		Reunion
	Rwanda		Saint Helena		Sao Tome and Principe		Senegal		Seychelles
	Sierra Leone		Somalia		South Africa		Swaziland		Tanzania
	Togo		Tunisia		Uganda		Zambia		Zimbabwe

### Americas

	Anguilla		Antigua and Barbuda		Argentina		Aruba		Bahamas
	Barbados		Belize		Bermuda		Bolivia		Brazil
	British Virgin Islands		Canada		Cayman Islands		Chile		Colombia
	Costa Rica		Dominica		Dominican Republic		Ecuador		El Salvador

## Problem

- Handle Postal addresses from **200+ countries and 27 languages (PayPal)**
- Address capture for different use-cases
  - customer sign-up, shipping, billing, regulatory compliance, etc.
- Address display for multiple scenarios
  - Single line, multi-line, PO Box, customized
- Coarse grained and fine grained address formats
- Bad Address Data

## Topics

*Why do we need address standards?*

*Address Entry Forms*

*Address Display Formats*

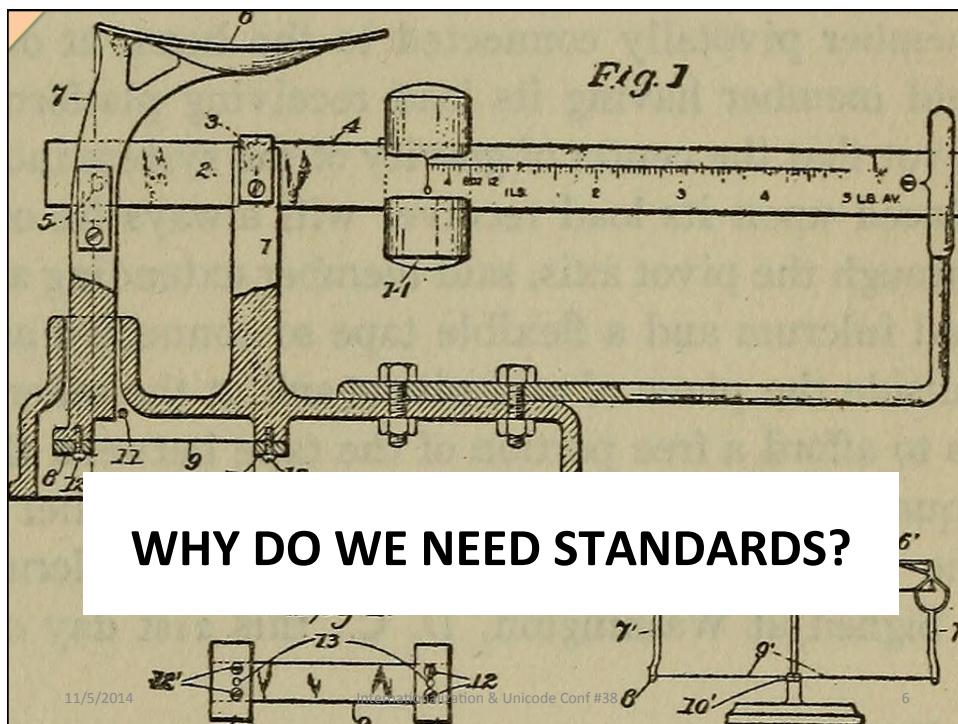
*Address Data*

*Arriving at an address framework*

11/5/2014

Internationalization & Unicode Conf #38

5



## WHY DO WE NEED STANDARDS?

- Is there a single address entry form that can be used globally?
- Why not just localize the field labels of the US address entry forms?
- What is the purpose of a postal address standard?

11/5/2014

Internationalization &amp; Unicode Conf #38

7

## WHY DO WE NEED STANDARDS

Same US address written in different ways



11/5/2014

Internationalization &amp; Unicode Conf #38

8

## WHY DO WE NEED STANDARDS

Same country different address types

Home delivery:

MR JOHN DOE  
123 MAGNOLIA ST  
HEMPSTEAD NY 11550-1234  
UNITED STATES OF AMERICA

Additional recipient information:

JOHN DOE  
ABC COMPANY  
1401 MAIN ST  
FALLS CHURCH VA 22042-1441  
UNITED STATES OF AMERICA

Delivery to PO Box:

MISS JANICE SMITH  
PO Box 34  
DULUTH MN 55803-0034  
UNITED STATES OF AMERICA

Rural route:

H E BROWN  
RR 3 BOX 9  
CANTON OH 44730-9521  
UNITED STATES OF AMERICA

11/5/2014

Internationalization & Unicode Conf #38

9

## WHY DO WE NEED STANDARDS?

US address vs addresses from other countries

168 Fangbang Middle Rd  
Huangpu, Shanghai  
P. R. China

500008  
INDIA

P. R. China  
上海市黄浦区  
(Shanghai) (Huangpu)  
方浜中路168号  
(Fangbang Middle Rd)

11/5/2014

Internationalization & Unicode Conf #38

10

## WHY DO WE NEED STANDARDS?

Providing support for postal addresses can be overwhelming because:

- Multiple address layouts within a country
- Non-standard address components
  - directional or suffix information
- Postal address formats differ from country to country
- Future proofing

11/5/2014

Internationalization & Unicode Conf #38

11

## WHY DO WE NEED STANDARDS?

### **Benefits of using standards**

- Address data can be used across projects, disciplines and applications
- Address data can be used across organizations
- Reduces duplicative efforts and lowers product costs
- Increases consistency
- Allows scalability
- Conflicts are more easily resolved
- Improves address quality

11/5/2014

Internationalization & Unicode Conf #38

12

## WHY DO WE NEED STANDARDS?

Examples of postal address standards

<b>International Standards:</b> <ul style="list-style-type: none"> <li>Universal Postal Union's <a href="#">S42 Standard</a></li> <li>ISO 19 19160 <a href="#">Address Standards</a></li> <li><a href="#">OASIS</a></li> </ul>	<b>Country Standards:</b> <ul style="list-style-type: none"> <li><a href="#">USPS (Pub 28)</a></li> <li><a href="#">Federal Geo Data Committee</a></li> <li><a href="#">BS 7666</a></li> <li><a href="#">Australia</a></li> <li><a href="#">Canada</a></li> <li><a href="#">Singapore</a></li> <li><a href="http://www.swisspost.ch">www.swisspost.ch</a></li> </ul>
--	--

11/5/2014      Internationalization & Unicode Conf #38      13

		<b>PERSONAL INFORMATION</b>			ACCOUNT NO.
Are in or are relying on property in a community property state or if you are applying for a secured credit or joint account.		SOCIAL SECURITY NO.	MARITAL STATUS: CHECK ONE If you reside in or are relying on property in a community property state or if you are applying for a secured credit or joint account. <input type="checkbox"/> MARRIED <input type="checkbox"/> SEPARATED <input type="checkbox"/> UNMARRIED		
LAST NAME (JR./SR.)		FIRST NAME		INITIAL	LAST NAME (JR./SR.)
YEARS AT THIS ADDRESS		CURRENT STREET ADDRESS			APT. NO.
DRIVER'S LICENSE NO.		CITY		STATE	ZIP
DENTS (NOT INCLUDING		DATE OF BIRTH	HOME PHONE (      )	NO. OF DEPENDENTS (NOT INCLUDING YOURSELF)	
<b>EMPLOYMENT INCOME</b>					
MONTHLY S PHONE ) DATE	Commit a paystub with this Application)	GROSS MONTHLY SALARY \$	WORK PHONE (      )	POSITION / TYPE OF WORK	START DATE

**ADDRESS ENTRY FORMS**

11/5/2014      Internationalization & Unicode Conf #38      14

## ADDRESS ENTRY FORMS

Multiple layout types

---

Coarse Grained Address Capture	Fine Grained Address Capture
<input type="text" value="Street Address Line 1"/> * <input type="text" value="Street Address Line 2"/> <input type="text" value="City"/> * <input type="text" value="State"/> <input type="text" value="ZIP Code"/> *	<input type="text" value="Number"/> * <input type="text" value="Street"/> * <input type="text" value="Building Name"/> <input type="text" value="Additional Guidance"/> <input type="text" value="Floor"/> <input type="text" value="Apartment"/> <input type="text" value="City"/> * <input type="text" value="State"/> <input type="text" value="ZIP Code"/> *

PO Box Address Capture
<input type="text" value="PO Box"/> * <input type="text" value="City"/> * <input type="text" value="State"/> <input type="text" value="ZIP Code"/> *

---

11/5/2014      Internationalization & Unicode Conf #38      15

## ADDRESS CAPTURE FORMATS

Different layouts for different scripts

---

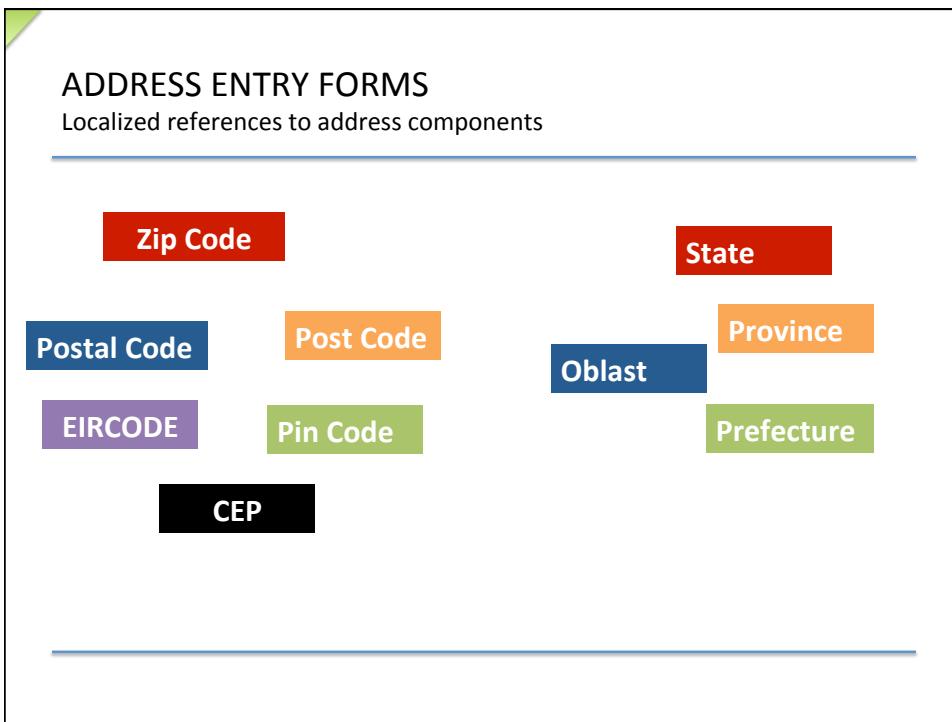
Native Script
<input type="text" value="郵便番号 (Postal Code - 0 0 0 - 0 0)"/> * <input type="text" value="都道府県 (Prefecture - to/dō/fu/ken)"/> * <input type="text" value="郡 (County - Gun)"/> <input type="text" value="市 (City)"/> * <input type="text" value="区 (District - Ku)"/> <input type="text" value="地名・団地名 (Area/Block/Complex name)"/> * <input type="text" value="丁目 (Chome Number - 0 - 0 - 0)"/> * <input type="text" value="ビル名 (Building Name)"/> 'フロ <input type="text" value="additionalGuidance"/> <input type="text" value="Contact or organization"/>

---

Latin Script
<input type="text" value="Contact or organization"/> <input type="text" value="additionalGuidance"/> <input text"="" type="text" value="District (ku)"/> <input type="text" value="City (shi)"/> * <input type="text" value="County (Rural Area - gen)"/> <input type="text" value="Prefecture (to/dō/fu/ken)"/> * <input type="text" value="Postal Code - 000-0000"/>

---

11/5/2014      Internationalization & Unicode Conf #38      16



## ADDRESS ENTRY FORMS

Getting it right — using lists (dropdown, scrolling)

<b>Provide drop-downs:</b> <ul style="list-style-type: none"> <li>• List of countries</li> <li>• List of principal sub divisions</li> <li>• sorted</li> <li>• available in multiple languages</li> </ul>	<b>Sorted list of US States</b> <small>▼ Click to expand the list</small> <ul style="list-style-type: none"> <li>AL Alabama</li> <li>AK Alaska</li> <li>AZ Arizona</li> <li>AR Arkansas</li> <li>CA California</li> <li>CO Colorado</li> <li>CT Connecticut</li> <li>DE Delaware</li> <li>DC District of Columbia</li> <li>FL Florida</li> <li>GA Georgia</li> <li>HI Hawaii</li> <li>ID Idaho</li> <li>IL Illinois</li> <li>IN Indiana</li> <li>IA Iowa</li> <li>KS Kansas</li> <li>KY Kentucky</li> <li>LA Louisiana</li> <li>ME Maine</li> <li>MD Maryland</li> <li>MA Massachusetts</li> <li>MN Michigan</li> <li>MN Minnesota</li> </ul>	<b>Sorted list of countries</b> <small>Ваша страна или регион</small> <table border="1" style="width: 100%; border-collapse: collapse;"> <tr><td style="padding: 2px;">Россия</td></tr> <tr><td style="padding: 2px;">Албания</td></tr> <tr><td style="padding: 2px;">Алжир</td></tr> <tr><td style="padding: 2px;">Ангола</td></tr> <tr><td style="padding: 2px;">Андорра</td></tr> <tr><td style="padding: 2px;">Антигуа и Барбуда</td></tr> <tr><td style="padding: 2px;">Аргентина</td></tr> <tr><td style="padding: 2px;">Армения</td></tr> <tr><td style="padding: 2px;">Аруба</td></tr> <tr><td style="padding: 2px;">Багамы</td></tr> <tr><td style="padding: 2px;">Барбадос</td></tr> <tr><td style="padding: 2px;">Бахрейн</td></tr> <tr><td style="padding: 2px;">Белиз</td></tr> <tr><td style="padding: 2px;">Белоруссия</td></tr> <tr><td style="padding: 2px;">Бельгия</td></tr> <tr><td style="padding: 2px;">Бенин</td></tr> <tr><td style="padding: 2px;">Бермудские о-ва</td></tr> <tr><td style="padding: 2px;">Болгария</td></tr> <tr><td style="padding: 2px;">Боливия</td></tr> </table>	Россия	Албания	Алжир	Ангола	Андорра	Антигуа и Барбуда	Аргентина	Армения	Аруба	Багамы	Барбадос	Бахрейн	Белиз	Белоруссия	Бельгия	Бенин	Бермудские о-ва	Болгария	Боливия
Россия																					
Албания																					
Алжир																					
Ангола																					
Андорра																					
Антигуа и Барбуда																					
Аргентина																					
Армения																					
Аруба																					
Багамы																					
Барбадос																					
Бахрейн																					
Белиз																					
Белоруссия																					
Бельгия																					
Бенин																					
Бермудские о-ва																					
Болгария																					
Боливия																					
List of countries and List of principal sub divisions: <a href="#">ISO 3166-2</a>																					

## ADDRESS ENTRY FORMS

Getting it right — verifying the postal code

```

<postCodeRegex territoryId="GB">GIR[ ]?0AA|((AB|AL|B|BA|BB|BD|BH|BL|BN|BS|BT|CA|CB|CF|CH|CM|CO|CR|CT|CV|CW|DA|DD|DE|DG|DH|DL|DN|DT|D|(\d|\dA-Z){2}\d{2}|(ABD-HJLN-UW-Z){2})|BFPO[ ]?\d{1,4}</postCodeRegex>
<postCodeRegex territoryId="IE">IE\d{2} \d{2} \d{2} \d{2}</postCodeRegex>
<postCode>Country</postCode> United States of America
<postCode>Zip Code format</postCode> NNNNN (optionally NNNNN-NNNN)
<postCode>Zip Code Reg-ex</postCode> d{5}([ -]\d{4})?
<postCode>Territory</postCode> IPRSTV-Z]\d</p>
<postCodeRegex territoryId="AU">\d{4}</postCodeRegex>
<postCodeRegex territoryId="IT">\d{5}</postCodeRegex>
<postCodeRegex territoryId="CH">\d{4}</postCodeRegex>
<postCodeRegex territoryId="AT">\d{4}</postCodeRegex>
<postCodeRegex territoryId="ES">\d{5}</postCodeRegex>
<postCodeRegex territoryId="NL">\d{4}[ ]?[A-Z]{2}</postCodeRegex>
<postCodeRegex territoryId="BE">\d{4}</postCodeRegex>
<postCodeRegex territoryId="DK">\d{4}</postCodeRegex>
<postCodeRegex territoryId="SE">\d{3}[ ]?\d{2}</postCodeRegex>
<postCodeRegex territoryId="NO">\d{4}</postCodeRegex>
-----><!-->
  
```

<http://unicode.org/repos/cldr/trunk/common/supplemental/postalCodeData.xml>

11/5/2014 Internationalization & Unicode Conf #38 19

## ADDRESS ENTRY FORMS

Parse key address fields

Street Address Line 1 → Flat Floor Building

Street Address Line 2 → Street Number Street Name

2798 West Main Street #12 Minneapolis MN 23876	Flat	12
	Street Number	2798
	Street Name	West Main Street
	City	Minneapolis
	State	MN
	Postal Code	23876

11/5/2014 Internationalization & Unicode Conf #38 20

## ADDRESS ENTRY FORMS

### Address Normalization

---

Address Correction and Standardization

2211 North First Street  
S Jose  
CA 11201

2211 N 1<sup>st</sup> St  
San Jose  
CA 95131-2021

---

11/5/2014      Internationalization & Unicode Conf #38      21

## ADDRESS ENTRY FORMS

### Auto fill

---

Legal business name

Business phone

Street address (no PO Box)

San Jose

CA

95112

---

11/5/2014      Internationalization & Unicode Conf #38      22

## ADDRESS ENTRY FORMS

### Address Validation

#### ➤ Basic Format Validation

- Using Regex
- Max length
- Data type checks

#### ➤ Extended Validation

- Does the address exist?
- Does the postal code match with city and state?
- Is the house number valid?

11/5/2014

Internationalization & Unicode Conf #38

23

## ADDRESS ENTRY FORMS

### Summary

- Different address capture layouts for different scenarios
  - Type: Multi-line vs single line, business vs residential
  - Coarse grain, fine grain
  - Scope: domestic, international
  - Script: native vs Latin
- Mandatory/optional fields for each locale
- Use drop-down lists to ease input
- Address Validation
- Address Normalization
- Ability to parse key address fields
- Localized references to address components

11/5/2014

Internationalization & Unicode Conf #38

24



11/5/2014

Internationalization &amp; Unicode Conf #38

25

## ADDRESS DISPLAY FORMATS

### Variation in the Order of Address Elements

#### Sample Austrian Address

- Street Name before Street No
- Postal Code before City

**Rennbahnweg 25/2/15**

<Street Name>      <St. No.>/<Block>/<Floor>

**1220 WIEN**

<PostalCode>      <City>

#### Sample UK Address

- Street No before Street Name
- City before Post Code

**4 Catherine Drive**

<St. No>      <Street Name>

**RICHMOND**

<City>

**TW9 2BX**

<PostCode>

11/5/2014

Internationalization &amp; Unicode Conf #38

26

## ADDRESS DISPLAY FORMATS

Constructing an address for display

Construction of a multi-line domestic address (fine grain)

```
<Contact><BR>
<StreetNo><space><StreetName><space><Flat><BR>
<City><space><State><space><ZipCode>
```

What the end user sees

```
Mr John Doe
123 Magnolia St Apt 2
Hempstead NY 11550
```

11/5/2014

Internationalization & Unicode Conf #38

27

## ADDRESS DISPLAY FORMATS

Display address layouts

Construction of a single line international address (coarse grain)

```
<AddressLine1><comma><space><AddressLine2><comma><space>
<City><space><State><space><ZipCode><space><COUNTRY>
```

What the end user sees

```
Apt 12, 123 Magnolia St, Hempstead NY 11550 UNITED STATES
OF AMERICA
```

11/5/2014

Internationalization & Unicode Conf #38

28

## ADDRESS DISPLAY FORMATS

### Zero Width Space

Construction of Single-line Japanese address with ZWSP

```
<Postal Code><ZWSP><City><ZWSP><Ward><ZWSP>
<District><ZWSP><Block><ZWSP><Room>
```

What the end user sees

〒100-8994<ZWSP>東京都<ZWSP>中央区<ZWSP>八重洲<ZWSP>一丁目<ZWSP>5番3号

Single line format wrapped using ZWSP

〒100-8994	Tōkyō-to
東京都	Chūō-ku
中央区	Yaesu
八重洲	1 Chōme
一丁目5番3号	5-ban 3-gō

11/5/2014 Internationalization & Unicode Conf #38 29

## ADDRESS DISPLAY FORMATS

### Wrap Up or Wrap Down

MS MILDRED DOE  
12 E BUSINESS LN **STE 209**  
KRYTON TN 38188-0002

Wrap Up Street No

MS MILDRED DOE  
**STE 209**  
12 E BUSINESS LN  
KRYTON TN 38188-0002

MS MILDRED DOE  
BUSINESS LN STE 209  
KRYTON TN  
**38188-0002**

Wrap Down Zip Code

MS MILDRED DOE  
BUSINESS LN STE 209  
KRYTON TN  
**38188-0002**

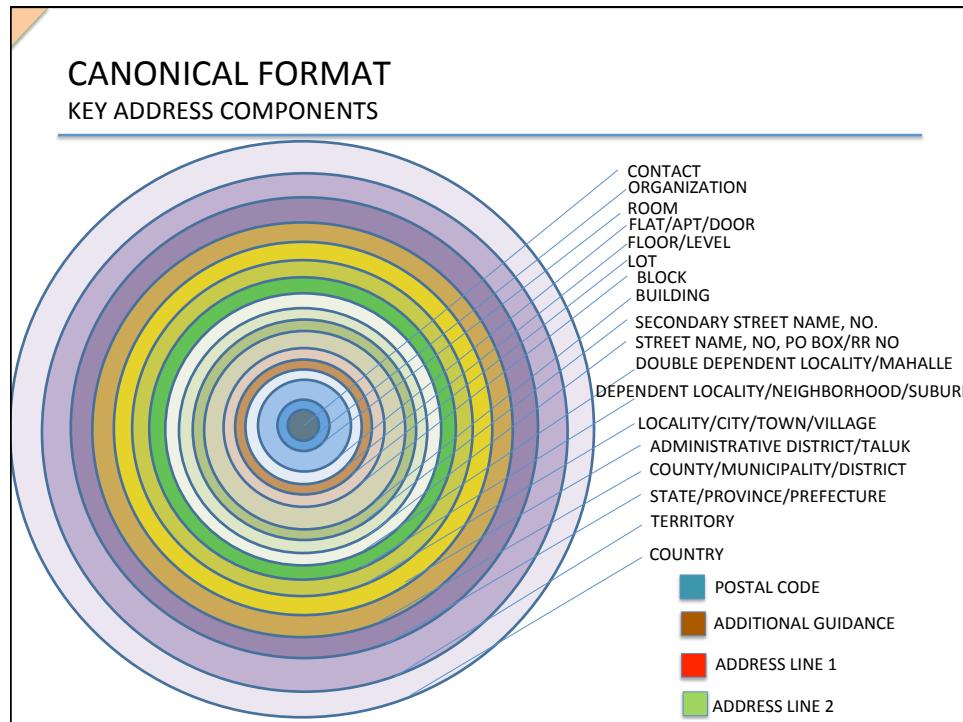
Guidelines on when to wrap up or wrap down a US address can be found [here](#)

## ADDRESS DISPLAY

### Summary

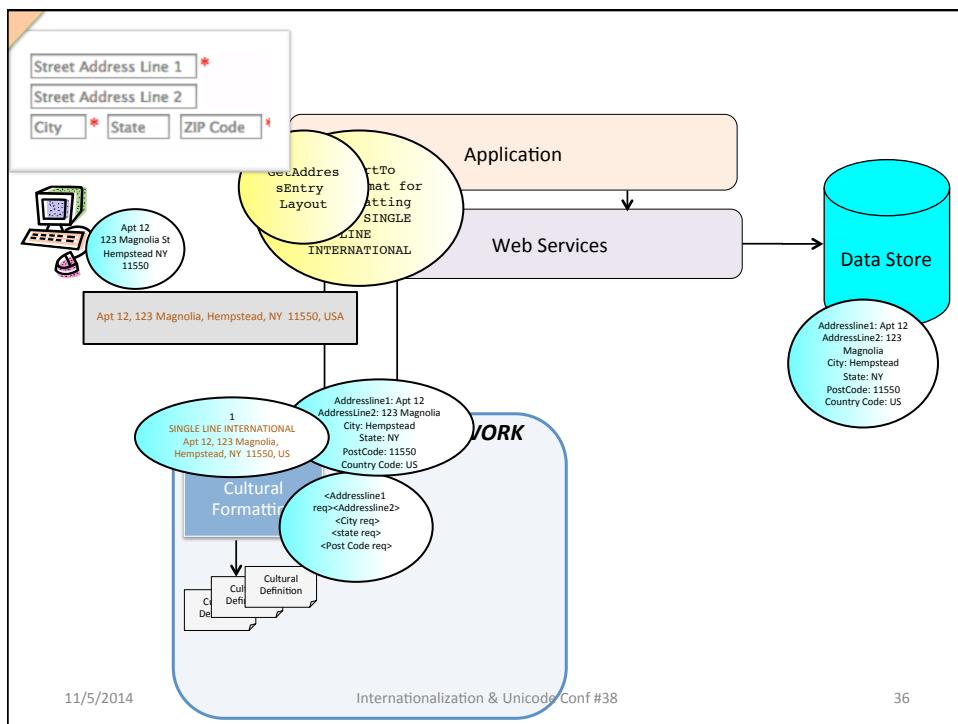
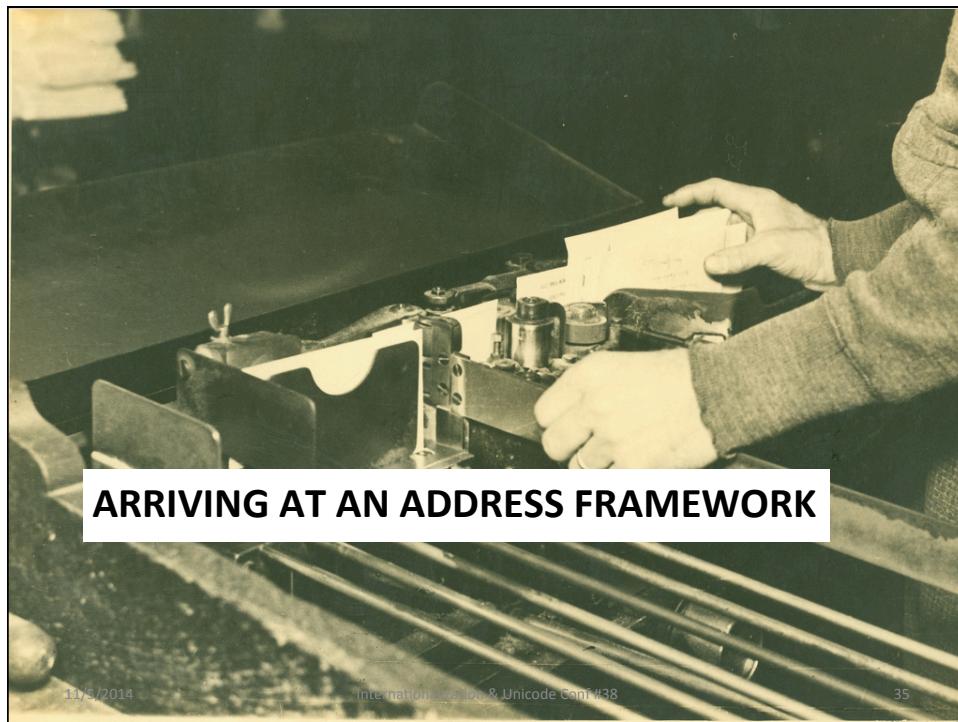
- The presentation order of address elements varies from country to country.
- Address assembly based on layout type:
  - Type: single-line, multiline, PO Box, business, residential
  - Coarse grain, Fine grain
  - Scope: domestic, international
  - Script: native vs Latin
- Formatting elements:
  - Casing
  - Address element separators (comma, space, ZWSP)
  - New Line
- Wrap Up or Wrap Down?

## ADDRESS DATA



## What is a Canonical Address Format?

- It is a single common schema to be used for interchange, processing and storage.
- It is locale-agnostic.
- All locale-specific fields should map to one or more fields of the canonical format.
- Composed of finite number of fields



## Postal Address Definition (DTD)

```
<!ELEMENT postalLayoutDefinition ( postalAddressLayout+ )>

<!-- POSTAL ADDRESS
=====
<!ELEMENT postalAddressLayout ( adrID?, isoTerritoryCode?, isPoBox?, line+ )>

<!ATTLIST postalAddressLayout type ( captureStreetAddress | capturePoBox | captureLine1Line2 | layout |
labelLayout | singleLine | line1Line2 | singleLine1Line2 ) #REQUIRED >

<!ATTLIST postalAddressLayout scope ( local | regional | international ) #REQUIRED >
<!-- Script tags from http://www.unicode.org/iso15924/iso15924-codes.html -->
<!ATTLIST postalAddressLayout script ( Latn | Grek | Cyrl | Jpan | Hira | Kana | Hrkt | Kore | Hang | Hans |
Hant | Bopo | Thai | Arab | Hebr ) #IMPLIED >
<!-- default separator value to use between elements in layouts -->
<!ATTLIST postalAddressLayout defaultSeparator CDATA " " >
<!ATTLIST postalAddressLayout casing ( lowercase | uppercase | titlecase ) #IMPLIED >
```

11/5/2014

Internationalization &amp; Unicode Conf #38

37

## Postal Address Definition (cont'd)

```
<!ELEMENT line ( addressee | contact | additionalGuidance | floor | flat | room |
subBuildingGroup | buildingName | streetAddress | streetNumber | streetName |
secondaryStreetNumber | secondaryStreetName | buildingGroup | blockName |
poBox | postOfficeName | suburb | district | administrativeDistrict | subLocality |
city | county | localityGroup | state | postcode | adrTerritory |
internationalTerritoryName | line1 | line2 | fixedText | sep | listsep | wrapOK |
special )+ >

<!-- What relative line number am I on, starting with the topmost line -->
<!ATTLIST line lineNo NMTOKEN #IMPLIED >

<!-- does the line wrap DOWN or wrap UP if too long? wrap at wrapOK tokens -->
<!ATTLIST line wrap ( up | down ) "down" >
```

11/5/2014

Internationalization &amp; Unicode Conf #38

38

## Postal Address Definition (cont'd)

```
<!ELEMENT city ( #PCDATA ) >
<!ATTLIST city required ( true | false ) "true" >
<!ATTLIST city pos CDATA #IMPLIED >
<!ATTLIST city label CDATA #IMPLIED >
<!ATTLIST city regex CDATA #IMPLIED >
<!ATTLIST city casing ( lowercase | uppercase | titlecase ) #IMPLIED >
```

11/5/2014

Internationalization &amp; Unicode Conf #38

39

## US Postal Address Format Definition

```
<postalAddressLayout type="captureLine1Line2" scope="local" casing="titlecase">
    <line lineNo="01">
        <line1 pos="01" label="Street Address Line 1"/>
    </line>
    <line lineNo="05">
        <line2 pos="01" label="Street Address Line 2"/>
    </line>
    <line lineNo="10">
        <city pos="01" required="true" label="City"/>
        <state pos="20" label="State" casing="uppercase"/>
        <postcode pos="30" label="ZIP Code" regex="\d{5}([ \-]\d{4})?"/>
    </line>
</postalAddressLayout>
```

11/5/2014

Internationalization &amp; Unicode Conf #38

40

## US address capture: coarse grain, line1Line2

Street Address Line 1	*			
Street Address Line 2				
City	*	State	*	ZIP Code

The input pattern for postcode is "\d{5}([ \-]\d{4})?"

11/5/2014

Internationalization &amp; Unicode Conf #38

41

## US address capture: coarse grain, line1Line2

### Address number and street

### More address information (optional)

City

State

State

ZIP code

Your ZIP code must be 5 or 9 digits.

11/5/2014

Internationalization &amp; Unicode Conf #38

42

## US address capture: fine-grain

'Number'\* 'Street'\*

Building Name

Additional Guidance

Floor

Apartment

City

\* State

ZIP Code

\*

The input pattern for postcode is "\d{5}([ \-]\d{4})?"

11/5/2014

Internationalization & Unicode Conf #38

43

## Japanese: line1Line2, local

郵便番号 (Postal Code - 000-0000) \*

都道府県 (Prefecture - to/dō/fu/ken) \*

郡 (County - Gun) 市 (City) \* 区 (District - Ku)

ストリート住所 1 (Street Address Line 1) \*

ストリート住所 2 (Street Address Line 2)

The input pattern for postcode is "\d{3}-\d{4}"

11/5/2014

Internationalization & Unicode Conf #38

44

## Japanese: Fine grain address

**Japanese Input**

郵便番号 (Postal Code - 000-0000)	*		
都道府県 (Prefecture - to/dō/fu/ken)	*		
郡 (County - Gun)	市 (City)	*	区 (District - Ku)
'地名・団地名 (Area/Block/Complex name)*' '丁目 (Chome Number - 0-0-0)*' 'ビル名 (Building Name)'			
additionalGuidance	... 'フロア (Floor)' 'ルーム (Room Number)'		
Contact or organization			

The input pattern for postcode is "\d{3}-\d{4}"

**Latin/International Input**

Contact or organization			
additionalGuidance			
'Floor' 'Room' 'Building Name' 'Chome Number - 0-0-0'* 'Street Name or Complex Name'			
District (ku)	City (shi)	*	County (Rural Area - gen)
Prefecture (to/dō/fu/ken)	*		
Postal Code - 000-0000	*		

The input pattern for postcode is "\d{3}-\d{4}"

## Key Takeaways

- Considerations for Address Entry:
  - ✓ Different layouts for different scenarios
  - ✓ Localized field labels
  - ✓ Address Validation
  - ✓ Address Normalization
- Considerations for Address Display:
  - ✓ Order address elements based on country and layout
  - ✓ Spacing, punctuation, casing and line breaks
  - ✓ Wrap up or down
- Use country standards to define the locale specific formats.
- Use of centralized canonical schema for Address data storage, processing and transmission



11/5/2014

Internationalization & Unicode Conf #38

47