A Story Teller's Case Study

Unlocking the Power of CLDR Person Name Formatting – A Solution for Formatting Names in a Globalized World

By Mike McKenna, Chair of CLDR Person Names Subcommittee
June 2023

BACKGROUND

How a person's name is displayed and used can convey respect, familiarity, or even be interpreted as rude if used improperly. That's why it's important to format names correctly, especially because naming practices vary across the globe. In many cultures, names can indicate gender, status, birthplace, nationality, ethnicity, religion, and more.

Until now, there have been no good standards for how to format people's names in various contexts. A number of Unicode members wanted to address this problem and provide a mechanism that anyone could use to format people's names in a wide variety of applications, such as contact lists, air travel, ticketing applications, CRMs, social media, and any other application that asks for user information and presents it back to the user or others.

UNICODE® PERSON NAME FORMATTING

The Unicode Person Name Formats defines patterns used to take a person's name and format it correctly in a given language or locale depending on a chosen context. With the Unicode Common Locale Data Repository (CLDR), locale codes and name sequences can be selected to create a specific pattern for formatting personal names — including preferences for formal, informal, or abbreviated versions. As a result, designers and developers can correctly display names according to the user's native locale and culture, especially important when integrating names in different character scripts, such as Japanese, Chinese, or Russian.

The paper below provides an idea of the implications for user experience and application design — it uses folk tale characters as personas attending an international storytelling conference as an illustration of the many contexts in which names can be formatted through CLDR Person Names.

- There are many different ways to present a person's name : an email greeting is very different from the name on a diploma or prestigious award

- Formatting a name correctly requires consideration of context, usage, and length
- CLDR PersonName formats will help implementers create more intuitive experiences that respect the cultures of their audience.

The Unicode Consortium added Person Name formatting to CLDR in version 42 and was subsequently refined and enhanced for v43, which was released in April 2023. In CLDR v43, with the help of linguists from around the world, we completed data for formatting people's names. Its formal name is “Unicode Technical Standard #35 Unicode Locale Data Markup Language (LDML); Part 8: Person Names". In ICU, the PersonNameFormatter class is available with “draft” status (ICU-22287) in ICU 73.

You can now see details of each defined PersonName format with example names in most locale languages in the CLDR Survey Tool (examples: en-US, nl-NL, ja-JP, fr-FR).

HOW WILL PERSON NAME FORMATTING HELP DEVELOPERS?

How will Person Name Formats help you? We published a summary background document when it was released in v42. But perhaps a little story will help illustrate:

Hello! My name is Johann Schmidt, but I prefer to be called “John”. You might know me by my longer name "John Jacob Jinglehiemer Schmidt". Let me introduce you to my friends Rip van Winkle III, Urashima Tarō (浦島太郎), Vasilisa Tsarevich (Царевна-лягушка), Shahrazad Shahryar (شهرزاد قصه‌گو), Belle-Belle Fortuné Bourbon-Busset (Belle-Belle ou Le Chevalier Fortuné), and Don Quijote (Don Quijote de la Mancha). We are going to a storytelling conference later this year in The Netherlands, and I have to organize everything. You would think it would be really straightforward, but the attendees are real sticklers for detail, and they are always pleased when their cultures are respected. They are pretty famous after all! Next year, the conference will be in Japan, which will be fun!

Things I have to worry about:

- Informal (or formal) email greetings to each attendee. Hopefully in their own language.
- How their names are presented when listed alongside their talks, and even how I should refer to them when introducing them to the audience, and how to welcome them in person
- What form should the names take on name badges for informal social gatherings?
- How should they be listed in the conference attendee list?
- We will be using an instant messaging app and I'll need to know how to present their names in abbreviated form for the text avatars in the chat sessions.
- I also have to make travel arrangements, so I'll need to know their full names when booking airfare.

Luckily, I have CLDR Person Name formats to help me out. I'll need the full name, informal name, and the preferred region they are from, as that will influence how their name gets displayed. I can see from both the spec, and ICU PersonName.NameField, that I can gather their title, credentials or professional letters, given names, surnames, and if needed generational info such as “Jr." or “III". I can also see that I have "modifiers" through the PersonName.FieldModifier enum where I can set values or overrides like capitalization, if the name entry is informal and if the surname has a prefix (aka tussenvoegsel) like my friend Van Winkle. Let's start by gathering all their name information. My interface is pretty simple, so I'll be limiting the number of available titles to Mr./Mrs./Ms./Mx./Dr./Prof. CLDR will not help me with the local versions of these terms, or list of standard terms yet, so for now, I will gather that information elsewhere.

Last year, the conference was in New York, and I have this information already:

NameField with FieldModifier	title	given	given-informal	given2	surname-prefix	surname-core	surname	surname2	generation	credentials	name locale
Belle-Belle	Mx.	Fortuné		Belle-Belle			Bourbon-Busset			chevalier	fr
Don	Mr.	Alonso	Don	Quijote			Quijano	de la Mancha		Esq.	es
John	Mr.	Johann	John	Jacob Jinglehiemer			Schmidt			MA, PMP	en
Rip	Mr.	Ripse	Rip		van	Winkle	*van Winkle*		III	Ph.D.	nl
Sharazad	Mrs.	Sharazad	Shari				Shahryar			Queen of Persia	fa
Urashima	Dr.	Tarō	Taro				Urashima				ja
Vasilisa	Ms.	Vasilisa	Froggi	Lyagushkavna			Tsarevna			Princess	ru

When formatting names, we need to consider what is the “formatting locale” and what is in the name object. That is, the name object can not only provide fields of data (such as the surname), but also other information, including the locale of the name, and the `preferredOrder` for the name. That is, the name object can provide an override for the ordering of surname vs given name. For example, Dr. Urashima is from Japan, and his name locale is “ja”. The NameOrder for “ja” is “surnameFirst”. Even if that is applicable for the Japanese script, we can use it to set the value of the `preferredOrder` for the name when expressed in the Latin script. Therefore, when his Latin-rendered name is formatted, the “surnameFirst” formats will be used even though the default for the formatting locale may be “givenFirst”. In other words, in an English or Dutch locale, we will see “Dr. Urashima Tarō”, not “Dr. Tarō Urashima”.

The PersonName formats provide different formats depending on length, formality, and usage — whether I am *referring* to someone (normative case), or *addressing* someone (vocative case), or formatting names for sorting. However, it does *not* provide a default format and it does not provide culturally appropriate equivalents of common titles, honorifics, or credentials. I will have to do my own homework as to which is best in each case.

I want to treat the attendees of the storytelling conference with respect, so I will follow the formality conventions of each person's language. The United States is pretty informal. The Netherlands is a bit more reserved, and I must show appropriate respect in Japan, especially if Zeus (*Ζεύς*) shows up. While CLDR v43 does not have default formality and length for locales, this *is* being added in v44.

Use Case 1: Ensuring names are accurate for air travel

This is actually the easiest part, and uses the CLDR structure, but with hard-coded patterns according to the ICAO Doc 9303 standard for international travel documents.

The result is just:

legal name for air travel	
surnames (primary name)	**given names (secondary name)**
{surname} {surname2}	{given} {given2}
Bourbon-Busset	Fortuné Belle-Belle
Quijano de la Mancha	Alonso Quijote
Schmidt	Johann Jacob Jinglehiemer
van Winkle	Ripse
Shahryar	Shahrazad
Urashima	Tarō
Tsarevna	Vasilisa Lyagushkavna

Use Case 2: Emailing speakers and attendees

For sending email greetings, I am *addressing* the attendee and using a *short* format since we will be sending several messages over time. For the United States, I will use an *informal* format, and for the Netherlands and Japan, I will use *formal*. I'll want to keep track of which formats I use for each context (*edit: starting in CLDR v44, I will have default formalities which will remove the extra work I am doing here*). I will use *short-addressing-informal* in the U.S. and *short-addressing-formal* in NL and JP. I have provided culturally-appropriate equivalent honorific titles, with permission from the attendees. *pName* is replaced with the formatted person names below. The Japanese renditions of the names were created by referring to existing Japanese Wikipedia pages, then a little extra translation if needed, with the final results approved by each attendee.

email Greetings \ welcome	en-US	nl-NL	ja-JP
Message	“Hi {pName}!”	“Hallo {pName}”	“こんにちは、{pName}”
length	short	short	short
usage	**addressing**	**addressing**	**addressing**
formality	**informal**	**formal**	**formal**
default format	*{given-informal}*	*{title} {surname}*	*{surname}{title}*
	Fortuné	Mx. Bourbon-Busset	ブルボン・ビュセットさん
	Don	Dhr. Quijano	キハーノさん
	John	Dhr. Schmidt	シュミットさん
	Rip	Dhr. Van Winkle	ヴァン・ウィンクル博士
	Shari	Mw. Shahryar	シャフリアールさま
	Taro	Dr. Urashima	浦島博士
	Froggi	Mw. Tsarevna	ツァロヴナさま

So far, it's pretty straight-forward. The name requested (*given-informal* or *surname*) was the name used. One locale-specific formatting note: in Dutch, the *surname-prefix* or *tussenvoegsels* like “van” are lower case when following the given name, but are in proper name case (*-initialCap*) when standing alone or immediately following the honorific title.

Use Case 3: Publishing the names in the Program

When we put the presenter's names in the program, we want to be sure we treat them with respect. But we don't need every last detail of their name. Since we are mentioning their names for reference, this would be the nominative case, or *usage=referring*. Medium, formal seems to fit well for this case.

presenter in program	en-US	nl-NL	ja-JP
Message	“Presented by {pName}”	“Gepresenteerd door …"	“…によるプレゼン”
length	medium	medium	medium
usage	**referring**	**referring**	**referring**
formality	**formal**	**formal**	**formal**
default format	*{given} {given2-initial} {surname} {generation}, {credentials}*	*{given} {given2-initial} {surname} {generation} {credentials}*	*{surname}{given}{title}*
	Fortuné B. B. Bourbon-Busset, chevalier	Fortuné B. B. Bourbon-Busset chevalier	フォルチュネ・ブルボン・ビュセットさん
	Alonso Q. Quijano, Esq.	Alonso Q. Quijano Esq.	アロンソ・キハーノさん
	Johann J. J. Schmidt, MA, PMP	Johann J. J. Schmidt MA, PMP	ジョン・シュミットさん
	Ripse van Winkle III, Ph.D.	Ripse van Winkle III Ph.D.	リップ・ヴァン・ウィンクル・三世博士
	Shahrazad Shahryar, Queen of Persia	Shahrazad Shahryar Koningin van Perzië	シェヘラザード・シャフリアールさま
	Urashima Tarō	Urashima Tarō	浦島太郎博士
	Vasilisa L. Tsarevna, Princess	Vasilisa L. Tsarevna Prinses	ヴァシリーサ・ツァロヴナさま

Now it starts to get interesting! First, you can see in en-US that Urashima has his surname first. Where initials for names are needed, this is done automatically with every word in the name initialized. In France, Fortuné would have Belle-Belle initialized as "B.-B." (*edit: French hyphenated initials will be included in CLDR v44*). In ja-JP, it may not be obvious, but the names from regions outside Japan are formatted with the givenFirst formats. “ジョン・シュミットさん”== “John・Smith-san”. Notice the middle dot “・” that was automatically inserted between each part of the name. This is handled through an auxiliary specifier in the CLDR format data called the *foreignSpaceReplacement*.

The *foreignSpaceReplacement* is used in languages that do not usually have spaces between parts of a name, and is included to help delineate between parts of names coming from other regions and languages. A famous example is “Albert Einstein” ⇒ "アルベルト・アインシュタイン" in Japanese and ⇒ "阿尔伯特・爱因斯坦" in Chinese

Use Case 4: Introducing the speakers at the event

We want to give prompts to the facilitators for each session to help them introduce the speakers. It looks like *usage=referring, length=medium* would be appropriate, with *formality=informal* for the U.S. and *formal* for Japan. I could go either way for Dutch since the attendees know each other pretty well.

introduction by facilitator	en-US	nl-NL	ja-JP
Message	“Please welcome {pName}”	“Gelieve {pName} welkom te heten”	“{pName}、いらっしゃいませ”
length	medium	medium	medium
usage	referring	referring	referring
formality	**informal**	**informal**	**formal**
default format	*{given-informal} {surname}*	*{given-informal} {surname}*	*{surname}{given}{title}*
	Fortuné Bourbon-Busset	- same as en-US -	フォルチュネ・ブルボン・ビュセットさん
	Don Quijano	""	アロンソ・キハーノさん
	John Schmidt	""	ジョン・シュミットさん
	Rip van Winkle	""	リップ・ヴァン・ウィンクル・三世博士
	Shari Shahryar	""	シェヘラザード・シャフリアールさま
	Urashima Taro	""	浦島太郎博士
	Froggi Tsarevna	""	ヴァシリーサ・ツァロヴナさま

If we wanted to make the names more familiar to the audience, we could maintain a parallel set of name objects in our attendee data to highlight their “stage names”, and use those to override the default formatting from CLDR PersonName. An example might

be Don's “stage name” data could be simple and use the *long-referring-formal* format to get “Alonso (Don Quixote) Quijano”. My “stage name” data may look like this:

given	given2	surname	name locale
Belle-Belle	(Belle-Belle ou Le Chevalier Fortuné)	Fortuné	fr
Alonso	(Don Quixote de la Mancha)	Quijano	es
John	(Jacob Jinglehiemer)	Schmidt	en
Rip		van Winkle	nl
Shahrazad	(the Storyteller)	Shahryar	fa
Tarō		Urashima	ja
Vasilisa	(Princess Frog)	Tsarevna	ru

Use Case 5: Supporting chat avatars in instant messaging

We use a proprietary messaging tool during the conference, and each attendee is represented by the initials of their name. We'll use the *usage=monogram* format which creates a sequence using the first letter of each name part. The *long-monogram-formal* formats give the most information to reduce confusion, but are still short enough to not clutter up the chat sessions.

chat avatar	en-US	nl-NL
length	long	long
usage	monogram	monogram
formality	formal	formal
default format	*{given-monogram-allCaps} {given2-monogram-allCaps} {surname-monogram-allCaps}*	*{given-monogram-allCaps} {given2-monogram-allCaps} {surname-prefix-monogram} {surname-core-monogram-allCaps}*
	FBB	FBB
	AQQ	AQQ
	JJS	JJS
	RV	**RvW**
	SS	SS
	UT	UT
	VLT	VLT

Here, we see the differing treatment of the *tussenvoegsel* in Dutch vs English. In en-US, “Van Winkle” is considered to be the full surname, and we only use “RV” for Rip van Winkle. However, in Dutch, the *tussenvoegsel* “van” is considered distinctive enough that it is kept, and we get “RvW” as the generated monogram. Since Japanese and Chinese use Han characters, monograms don't make much sense, so we will use the Latin versions of the name objects for them.

Use Case 6: Accurately sorting the attendee lists

The name order is usually automatic based on the format locale or the name locale, but there is an explicit “`SORTING`” attribute that can be set to get names formatted so they will sort as expected for an index or ordered list. The usage is always “referring”. Just to get an idea, here are sorted lists for short-informal, long-formal, and short formal for en-US, nl-NL, and ja-JP respectively. I added the Latin surname to the Japanese to help with cross-language searching.

Sorting/Index	en-US	nl-NL	ja-JP
length	short	long	short
usage	referring	referring	referring
formality	informal	formal	formal
default format	*{surname}, {given-informal}*	*{surname-core}, {given} {given2} {surname-prefix}*	*{surname} {given}*
	Bourbon-Busset, Fortuné	Bourbon-Busset, Fortuné Belle-Belle	ヴァン・ウィンクル・リップ (van Winkle)
	Quijano, Don	Quijano, Alonso Quijote	ウラシ・タロウ (Urashima)
	Schmidt, John	Schmidt, Johann Jacob Jinglehiemer	キハーノ・アロンソ (Quijano)
	Shahryar, Shari	Shahryar, Shahrazad	シャフリアール・シェヘラザード (Shahryar)
	Tsarevna, Froggi	Tsarevna, Vasilisa Lyagushkavna	シュミット・ジョン (Schmidt)
	Urashima, Taro	Urashima, Tarō	ツァロヴナ・ヴァシリーサ (Tsarevna)
	van Winkle, Rip	Winkle, Ripse van	ブルボン・ビュセット・フォルチュネ (Bourbon-Busset)

Use Case 7: Generating the name badges

Finally, I want to print name badges for the attendees to wear during social-networking events. Since they will be *informal* social gatherings, we will use *short-**addressing**-**informal*** in large type, followed by *short-**referring**-**formal*** in small type on the line below, and if they have supplied their name in a native script other than the conference locales' then place that on the last line in parentheses. We would have the following for en-US and nl-NL.

Fortuné F.B.B. Bourbon-Busset
Don A.Q. Quijano
John J.J.J. Schmidt
Rip R. van Winkle
Shari S. Shahryar (شهرزاد شهریار)
Taro Urashima T. (浦島太郎博士)
Froggi V.L. Tsarevna (*Василиса Царевна*)

CONCLUSION

As we have seen with John Schmidt, Rip van Winkle, Urashima Tarō and their storytelling friends, there are many different ways to present a person's name and doing it correctly requires the correct context, usage, and length settings. With the help of CLDR PersonName formats, it should remove some of the mystery and help implementers create more intuitive experiences that respect the cultures of their audience.

For CLDR v43, an extensive amount of linguistic and in-region checking has been done, many errors and incomplete formats have been corrected, and the name data has been thoroughly reviewed.

Beyond CLDR v43, there are several additional features planned for CLDR v44, including (but not limited to):

- limited handling of mixed declensions for languages such as Lithuanian
- simple name validation
- default formality and length per locale
- better handling of languages that don't use spaces between words

And we'll continue to enhance the support beyond CLDR v44.

ADDITIONAL RESOURCES

For more information, see

- *Unicode Technical Standard #35 Unicode Locale Data Markup Language (LDML); Part 8: Person Names*
- CLDR Survey Tool : Locale | Miscellaneous | PersonName
- ICU4j interface PersonName

Past IUC presentations that may be helpful

- IUC 44 (video)
  - CLDR: What's in a Personal Name?
- IUC 45 (video)
  - CLDR and Person Names - The Saga Continues

Image credit: "Hello my name is" https://commons.wikimedia.org/wiki/File:Hello_my_name_is_sticker.svg

ABOUT THE AUTHOR

mike mckenna — ɐuuǝʞɔom ǝʞıɯ

Mike McKenna is the Chair of the CLDR Person Name Formatting Subcommittee and a Globalization Architect and Head of Globalization Engineering at Square Inc. He is responsible for next generation localization and internationalization infrastructure as well as evangelizing a Globalization-First culture across Square

ABOUT THE UNICODE CONSORTIUM

For 30 years, Unicode has coordinated the efforts of a world-wide team of volunteer programmers and linguists to standardize, evolve, and maintain a global software foundation that allows virtually every computer system and service to help people connect using their native language. This has real world consequences.

Today's global economy runs on networks that reach billions of people around the globe. An airline reservation system in Ireland processes a reservation made by a traveler in Swahili. Databases, commerce engines, websites and shipping systems handle local names, addresses, and text in hundreds of languages from Latin to Cyrillic to Hindi to Japanese – all thanks to Unicode standards and code.

Unicode Consortium Quick Facts

- Founded in 1988, incorporated in 1991
- Public benefit, 501(c)3 non-profit organization
- Open source standards, data, and software development
- Orchestrates the contributions of 100s of professionals, expert volunteers, and language experts
- 30+ organizational members across corporate, academic, and governmental institutions
- Funded by membership dues and donations

Website
Unicode CLDR Project
Technical Quick Start Guide
How to Become a Member (for Organizations)
How to Become a Member (for Individuals)

Contact Unicode

611 Gateway Blvd.
Suite 120
South San Francisco, CA 94080
info@unicode.org