

CC:AAM • 16 June 2022

## **OCLC Shared Entity Management Infrastructure: Language and script challenges in the creation of Work and Person entities**

**Becky Dean**

Product Manager

**Charlene Morrison**

Database Specialist

# Shared Entity Management Infrastructure project

- Two-year, \$2.436M Mellon grant, matched by OCLC
  - December 2019 - December 2021
- Collaboration with library community
- Data sources for WorldCat Works and WorldCat Persons
- Production infrastructure for WorldCat Entities
  - WorldCat Entities: Persistent identifiers for entities in aggregation
  - OCLC Meridian: Create and edit entities

---

# WORLDCAT WORK ENTITIES

---



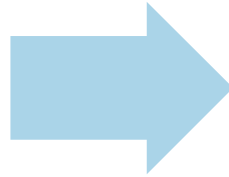
# Creation of WorldCat Work entities

- Used WorldCat bibliographic records
  - FRBR Clusters
  - Representative expressions
- Relationships to WorldCat Person entities
- Created over 130M WorldCat Work entities
  - 300+ languages represented

# WorldCat Work entities & scripts

## Phase 1

- Chinese
- Japanese
- Korean
- Hebrew
- Arabic
- Cyrillic



## Phase 2

- All scripts

# BCP-47 for scripts

## Tajik example

- tg-Arab: Tajik in Arabic script
- tg-Cyrl: Tajik in Cyrillic script
- tg-Latn: Tajik in Latin script

## Japanese example

- ja-Hani: Japanese written in kanji
- ja-Hira: Japanese written in hiragana
- ja-Hrkt: Japanese written in kana
- ja-Kana: Japanese written in katakana

# Label, title, and alias

- Language tag string
    - Literal: textual string
- +
- Language tag: A two-part element identifying the language and script of the literal
    - Name of the language associated with the literal (ISO 639-1)
    - Name of the script associated with the literal (ISO 15924)

# Label

- Label is the title of the expression

Language

Korean

Label

Madang ūl naon amt'ak

Translation



Language

Chinese

Label

Jianzhen he shang dong zheng chuan



# Title

- Title is the title of the expression

Descriptor	title
Value	마당을 나온 암탉 (Korean)

Translation 

Descriptor	title
Value	鉴真和尚东征传 (Chinese)

# Alias

- Other titles associated with the WC Work from the representative expression

Language	Georgian	Descriptor	title
Label	Antikuri kultura	Value	ანტიკური კულტურა (Georgian)

## Description

Aliases	Antikuri kultura Bizantia da Sak'art'velo
---------	---

---

# WORLDCAT PERSON ENTITIES

---

VIAF

Virtual International Authority File



# Creation of WorldCat Person entities

- Used aggregation data from the Virtual International Authority File (VIAF) for persons
- Included Wikidata data when the VIAF aggregation record contained a Wikidata link
- Created approximately 2M person entities
  - 1,159,953 VIAF aggregations include a Wikidata link
  - 230+ languages represented

# Connecting data in VIAF to WorldCat

- Authority data in VIAF
- Authority data used to control headings in bibliographic records in WorldCat
  - *Biblioteca Nacional de España*
  - *Bibliothèque et Archives Canada / Library and Archives Canada*
  - *Deutsche Nationalbibliothek*
  - *Koninklijke Bibliotheek*
  - *Library of Congress Name authority file*

# Label: Literal + “Language tags”

- Literal: the name given to identify an entity, typically, the most common name that the entity would be known
- +
- Language tag: A two-part element is required to create the literal that is in the object position of a triple statement
  - Name of the language associated with the literal (ISO 639-1)
  - Name of the country associated with the literal (ISO 3166)

# Concept of language in authority records

- Given that VIAF is an aggregation of 56 authority files from more than 30 countries around the world
- Given that VIAF contributors create authority records based on their established cataloging rules
- *Problem to solve: What information could/should be used to determine the **language name** and **country code** for a literal extracted from a VIAF aggregation document?*

# Challenges


- Assign a "language" using the United Nations official language(s) associated with the VIAF contributor's country
  - many countries have more than one official language
    - Canada (en-CA ; fr-CA)
    - Switzerland (de-CH ; fr-CH ; it-CH ; rm-CH)
  - some VIAF data files are not bounded by a country, rather, some files are other registries or focused projects
    - ISNI: an ISO standard that uniquely identifies organizations involved in creative activities
    - PERSEUS: about individuals associated with the history, literature, and culture of the Greco-Roman world




# VIAF: Yōko Ogawa


Ogawa, Yōko, 1962-





小川, 洋子, 1962- 

小川, 洋子 


오가와 요코 1962- 

小川洋子 

1962 , יוקו , אוגוה, - 

Ogawa, Yoko 

Ogawa, Yōko f. 1962 

Огава, Е. Еко 

# Label creation using VIAF: Yōko Ogawa

	Language(s)	Country	
National Library of Australia	English	Australia	"Yōko Ogawa"@en-AU
National Library of Brazil	Portuguese	Brazil	"Yōko Ogawa"@pt-BR
Library and Archives Canada	French	Canada	"Yōko Ogawa"@fr-CA
	English		"Yōko Ogawa"@en-CA
National Library and Archives of Quebec	French	Canada	"Yōko Ogawa"@fr-CA
	English		"Yōko Ogawa"@en-CA
National Library of Catalonia	Catalan	Catalonia	"Yōko Ogawa"@ca-ES
	Spanish		"Yōko Ogawa"@es-ES
National and University Library in Zagreb	Croatian	Croatia	"Yoko Ogawa"@hr-HR
National Library of the Czech Republic	Czech	Czech Republic	"Yōko Ogawa"@cs-CZ
DBC Digital	Danish	Denmark	"Yōko Ogawa"@da-DK
National Library of the Netherlands	Dutch	Netherlands	"Yōko Ogawa"@nl-NL
National Library of Estonia	Estonian	Estonia	"Yōko Ogawa"@et-EE
National Library of France	French	France	"Yōko Ogawa"@fr-FR
Sudoc [ABES], France	French	France	"Yōko Ogawa"@fr-FR
German National Library	German	Germany	"Yōko Ogawa"@de-DE

# Label creation using VIAF: Yōko Ogawa

	Language(s)	Country	
National Library of Greece	Greek	Greece	"Yōko Ogawa"@el-GR
National Library of Israel	Hebrew	Israel	"יוקו אוגוה"@he-IL
National Library of Lithuania	Lithuanian	Lithuania	"Yōko Ogawa"@lt-LT
National Diet Library	Japanese	Japan	"洋子 小川"@ja-JP
National Institute of Informatics	Japanese	Japan	"洋子 小川"@ja-JP
National Library of Korea	Korean	Korea	"오가와 요코"@ko-KR
BIBSYS	Norwegian	Norway	"Yōko Ogawa"@no-NO
NUKAT Center of Warsaw University Library	Polish	Poland	"Yōko Ogawa"@pl-PL
National Library of Poland	Polish	Poland	"Yōko Ogawa"@pl-PL
National Library of Portugal	Portuguese	Portugal	"Yōko Ogawa"@pt-PT
National Library of Russia	Russian	Russia	"Е. Огава"@ru-RU
National Library of Spain	Spanish	Spain	"Yōko Ogawa"@es-ES
National Library of Sweden	Swedish	Sweden	"Yōko Ogawa"@sv-SE
RERO – Library Network of Western Switzerland	German	Switzerland	"Yōko Ogawa"@de-CH
	French		"Yōko Ogawa"@fr-CH
	Italian		"Yōko Ogawa"@it-CH
Library of Congress	English	United States	"Yōko Ogawa"@en-US

# Labels added from Wikidata: Yōko Ogawa

Albanian	Indonesian	Dhivehi <i>use Divehi</i>
Arabic	Irish	British English
Bengali	Italian	Canadian English
Catalan	Papiamentu	Egyptian Arabic
Chinese	Persian	Greek
Estonian	Swedish	South Azerbaijani
Galician	Urdu	Western Punjabi

# FINAL THOUGHTS...

# What did we learn?

- There are more than 7,100 languages spoken in the world
- There are more than 30 scripts, and some languages have multiple scripts
- Some countries have up to 12 official languages
- Some languages found in Wikidata person descriptions that are not represented in WorldCat
- Cataloging policies for manifestations impact languages and their representation in WorldCat Work entities (e.g., translations)

# Links

- OCLC
  - FRBR Work-Set: <https://www.oclc.org/research/activities/frbr.html>
  - VIAF: <http://viaf.org/>
  - WorldCat Entities: [entities.oclc.org](https://entities.oclc.org)
  - WorldCat Entities and OCLC Meridian: [oclc.org/meridian-entities-press-release](https://oclc.org/meridian-entities-press-release)
- Standards
  - BCP-47: <https://tools.ietf.org/search/bcp47>
  - ISO 639-1: <https://www.iso.org/iso-639-language-codes.html>
  - ISO 3166: <https://www.iso.org/iso-3166-country-codes.html>
  - ISO-15924: <https://www.iso.org/standard/81905.html>

- Interested in feedback from language specialists and the non-Latin script community
- Send questions and feedback to:  
[linkeddata@oclc.org](mailto:linkeddata@oclc.org)

**Because  
what is  
known must  
be shared.®**