Language Commons

The Language Commons Wiki

Ed Bice (Meedan.net)
Steven Bird (University of Melbourne / University of Pennsylvania)
Kurt Bollacker (The Long Now Foundation)
Gary Simons (SIL International)
Laura Welcher (The Long Now Foundation – Rosetta Project)

LW

Outline of this talk:

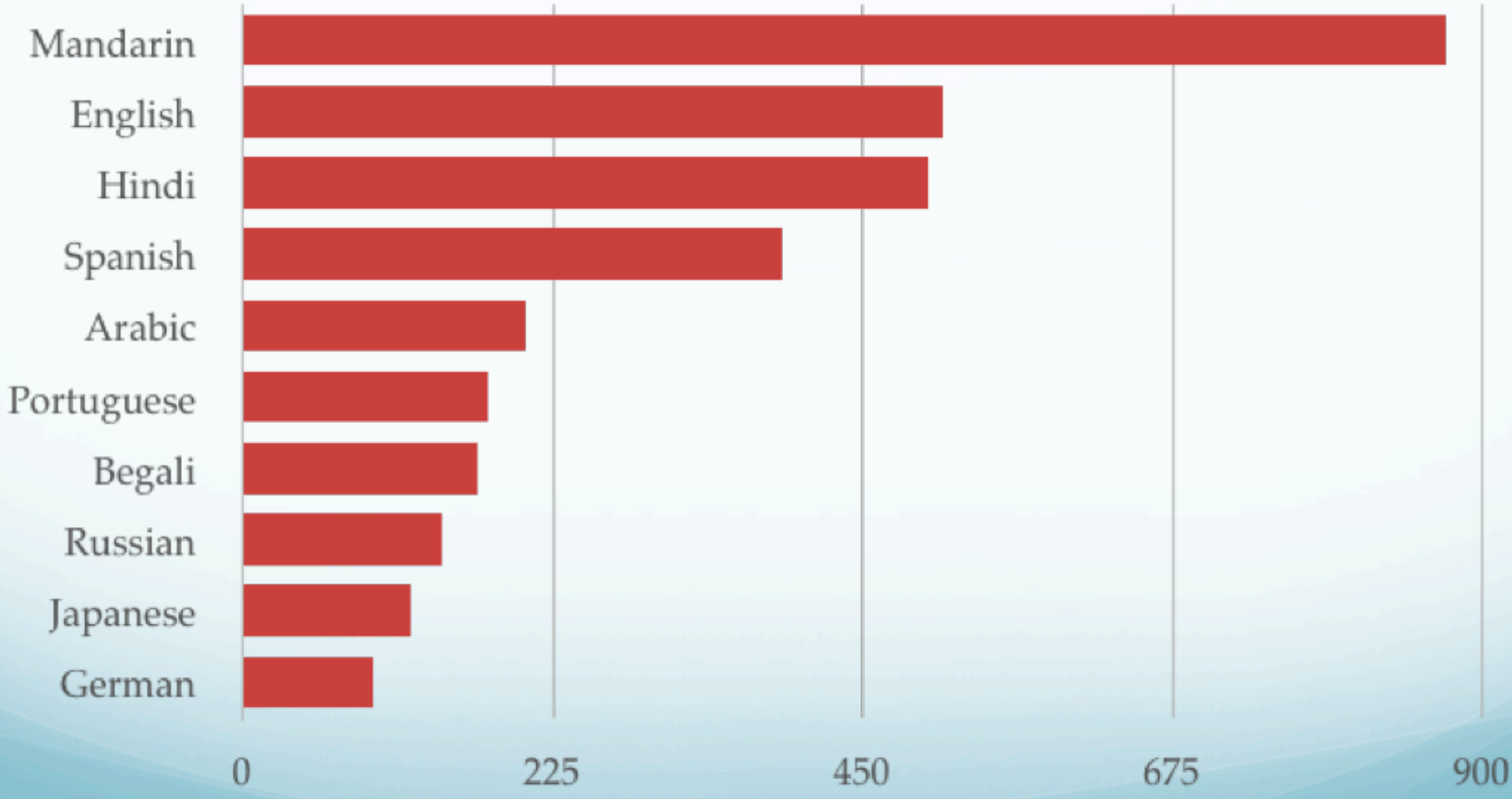
- Why build a wiki of all human language?
- How will the wiki work?
- A working prototype
- Questions and the future

LW

Why build a wiki of all human language?

LW

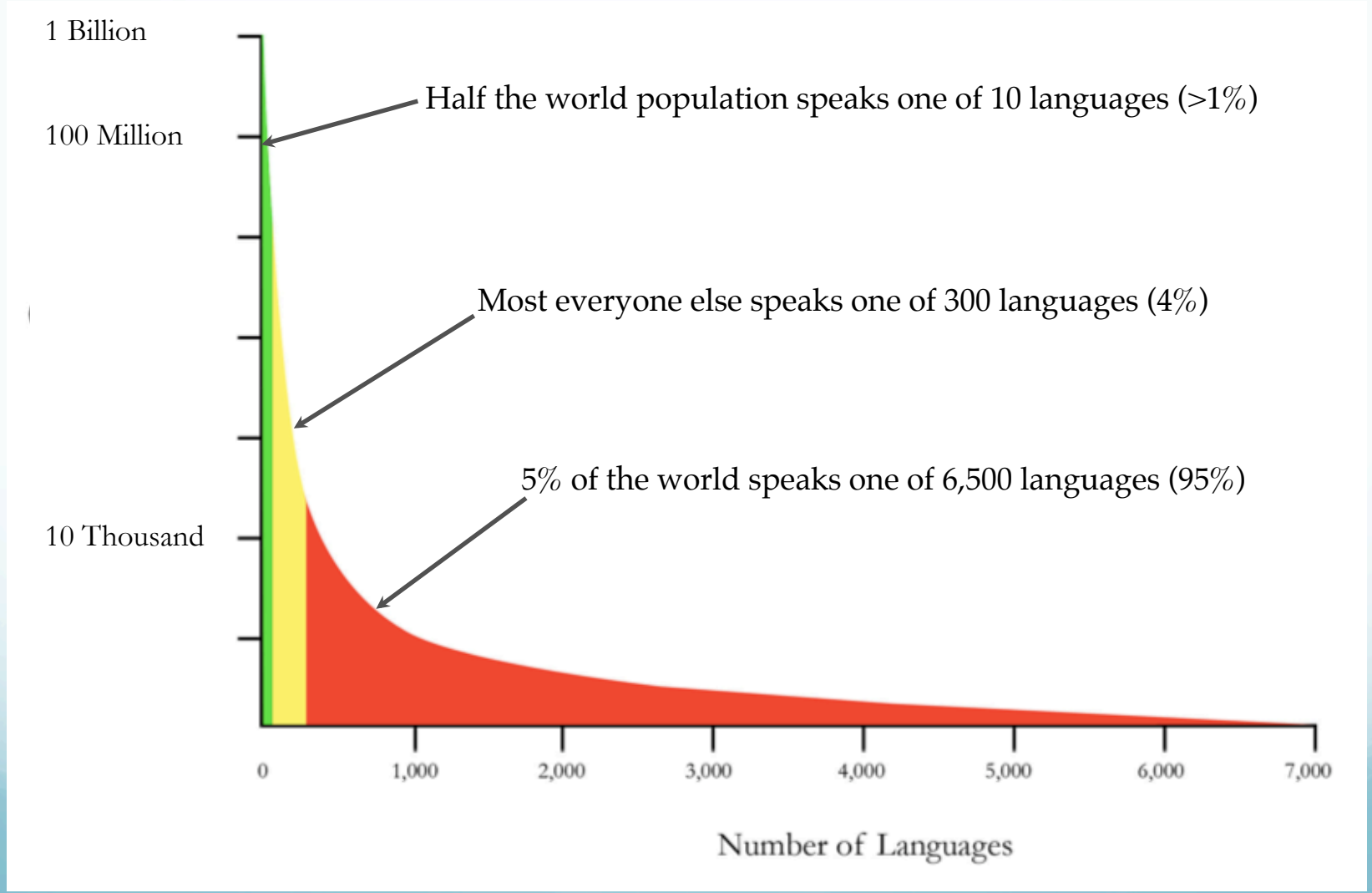
Top Ten languages by Native Speakers (Millions)



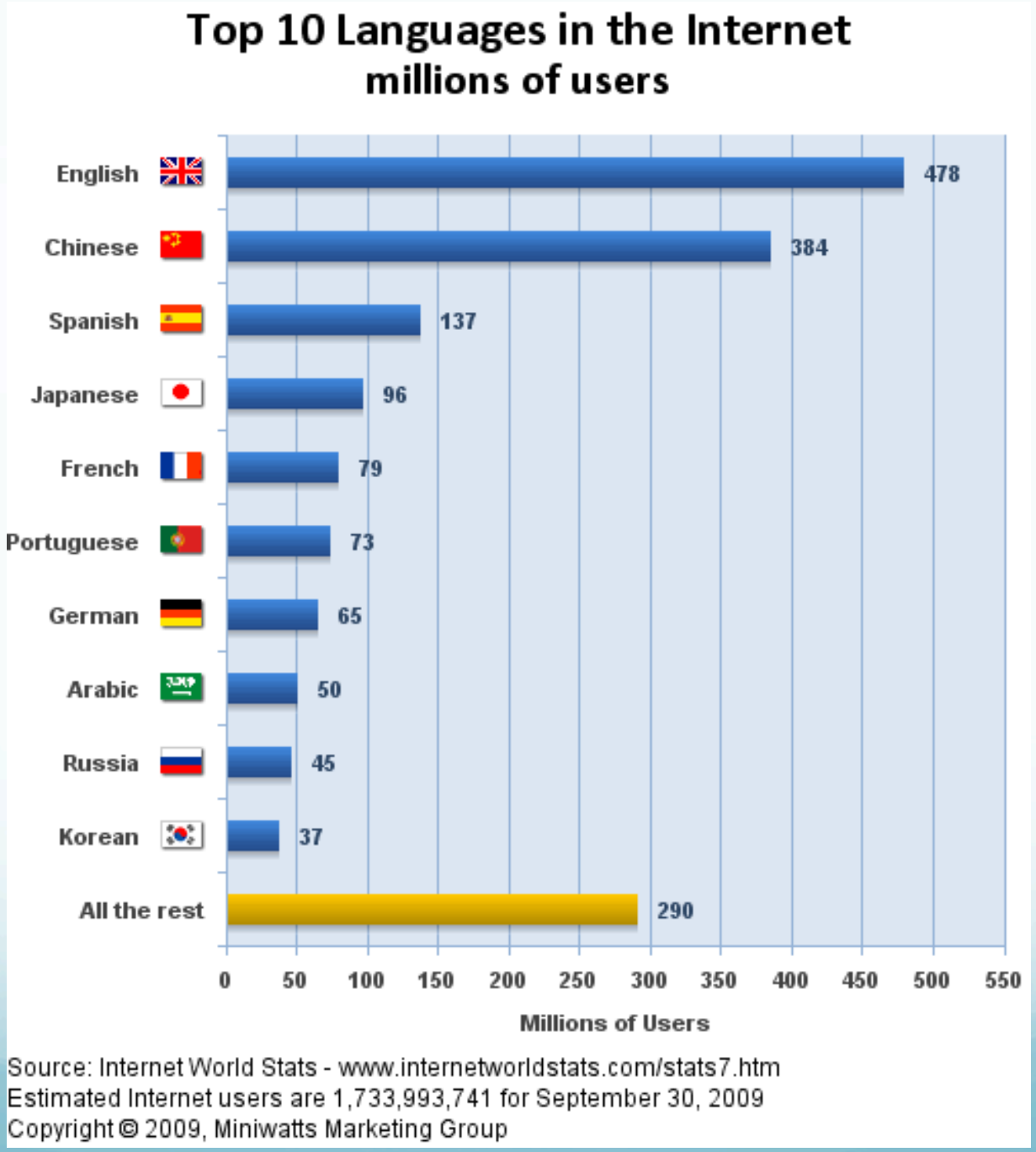
Data: The Ethnologue (2005) available at www.ethnologue.com

LW

“Long Tail” of Languages



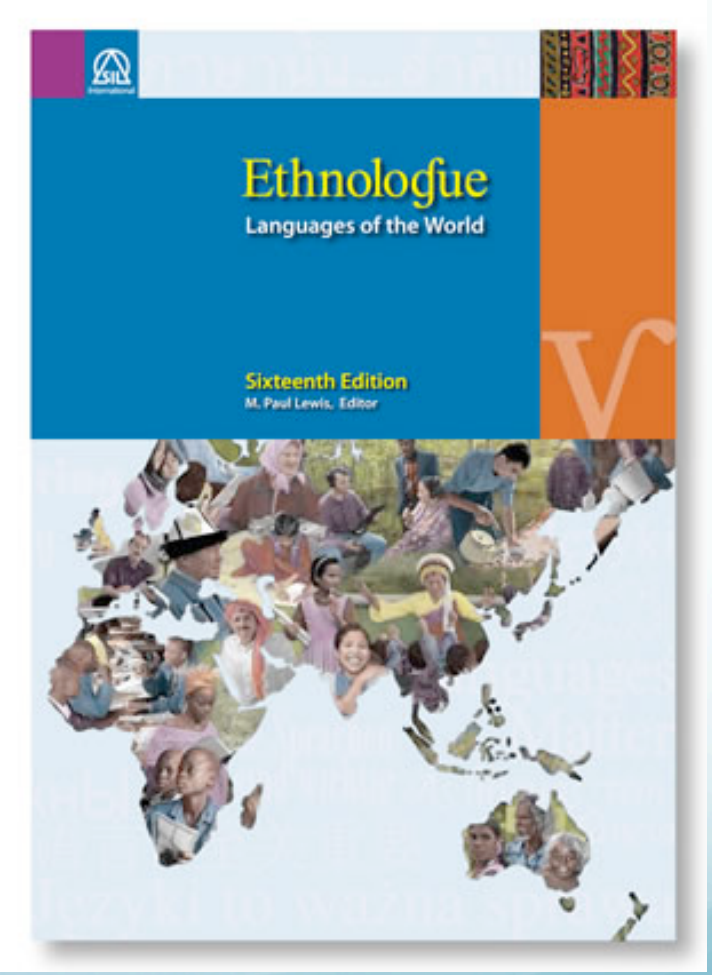
LW


Top 10 Languages in the Internet
millions of users
English
Chinese
Spanish
Japanese
French
Portuguese
German
Arabic
Russia
Korean
All the rest
478
384
137
96
79
73
65
50
45
37
290
0
50
100
150
200
250
300
350
400
450
500
550
Millions of Users
Source: Internet World Stats - www.internetworldstats.com/stats7.htm
Estimated Internet users are 1,733,993,741 for September 30, 2009
Copyright © 2009, Miniwatts Marketing Group

LW

Ethnologue

- 6900 language descriptions
- 41k language names
- ISO 639-3 codes
- statistical summaries
- ethnologue.com



SB

Open Language Archives Community

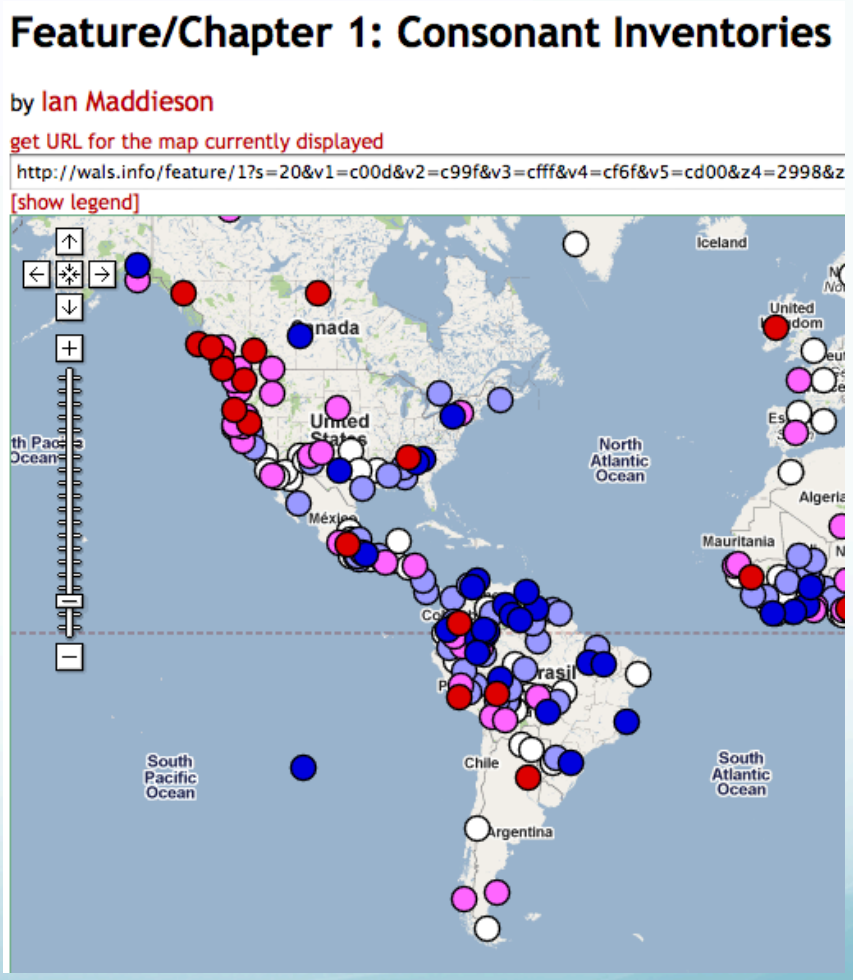
- digital / non-digital
- federated search
- page per language
- language-archives.org

Name	Value
Number of Archives	41
Archives with Fresh Catalogs	17
Archives with Five-star Metadata	9
Number of Resources	86042
Number of Resources Online	40193
Distinct Languages	7663
Distinct Linguistic Subfields	29
Distinct Linguistic Types	3
Distinct DCMI Types	9
Average Elements Per Record	17
Average Encoding Schemes Per Record	7.6
Average Metadata Quality Score	7.6
Last Updated	2010-06-10
Known Integrity Problems	74414

SB

World Atlas of Language Structures

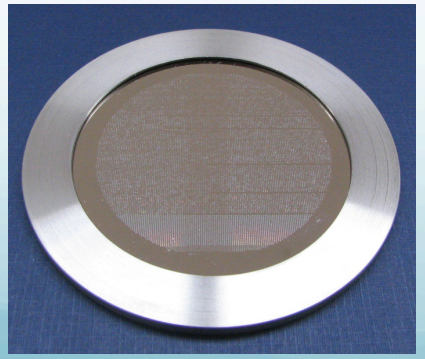
- 2650 languages
- 140 features (e.g. word order)
- 40 authors
- 6,100 references



SB

Rosetta Project

- ~ 2,500 languages
- 100,000+ text pages
- Audio and video recordings



SB

Large Scale Language Preservation Efforts

- Papua New Guinea
- >800 languages
- 100 voice recorders
- record, transcribe

SB

Proposal: All-Language Wiki

An aggregation and discovery portal for information and resources on **all** 6,900 human languages.

For use by:

- language speakers
- educators
- researchers
- general public

SB

Wiki Content

This will be the most comprehensive, accurate and accountable information available for any human language, available in a single location.

One page per language:

- **structured information** from dozens of external sources
- Further curated by:
  - **user-contributed material**
  - **expert commentary**

Linked index pages providing **taxonomic navigation** through 3,900 language families and subgroups

SB

How will it work?

Why can’t we just use the current Wikipedia?

KB

Current Wikipedia Article Structures

Lexington County, South Carolina

From Wikipedia, the free encyclopedia

Coordinates: 33.90°N 81.27°W

Lexington County is a county located in the U.S. state of South Carolina. In 2000, its population was 216,014. The 2008 Census estimate placed the population at 248,518.[1] Its county seat is Lexington.[2]

Lexington County is part of the Columbia, South Carolina, Metropolitan Statistical Area.

Contents [hide]

1 Geography
  1.1 Adjacent Counties
2 Demographics
3 Cities and towns
4 Transportation
5 References

Geography [edit]

According to the U.S. Census Bureau, the county has a total area of 758 square miles (1,963 km²), of which, 699 square miles (1,811 km²) of it is land and 59 square miles (152 km²) of it (7.74%) is water.

Adjacent Counties [edit]

- Richland County, South Carolina - east
- Orangeburg County, South Carolina - southeast
- Calhoun County, South Carolina - southeast
- Aiken County, South Carolina - southwest
- Saluda County, South Carolina - west
- Newberry County, South Carolina - northwest

Demographics [edit]

As of the census[3] of 2000, there were 216,014 people, 83,240 households, and 59,849 families residing in the county. The population density was 309 people per square mile (119/km²). There were 90,978 housing units at an average density of 130 per square mile (50/km²). The racial makeup of the county was 84.18% White, 12.63% Black or African American, 0.34% Native American, 1.05% Asian, 0.04% Pacific Islander, 0.79% from other races, and 0.98% from two or more races. 1.92% of the population were Hispanic or Latino of any race.

There were 83,240 households out of which 35.50% had children under the age of 18 living with them, 56.60% were married couples living together, 11.60% had a female householder with no husband present, and 28.10% were non-families. 22.50% of all households were made up of individuals and 6.90% had someone living alone who was 65 years of age or older. The average household size was 2.56 and the average family size was 3.01.

In the county, the population was spread out with 26.10% under the age of 18, 8.30% from 18 to 24, 31.60% from 25 to 44, 23.80% from 45 to 64, and 10.20% who were 65 years of age or older. The median age was 36 years. For every 100 females there were 94.50 males. For every 100 females age 18 and over, there

Lexington County, South Carolina

Location in the state of South Carolina

South Carolina's location in the U.S.

Seat	Lexington
Area	
- Total	758 sq mi (1,963 km²)
- Land	699 sq mi (1,810 km²)
- Water	59 sq mi (153 km²), 7.74%
Population	
- (2000)	216,014
- Density	308/sq mi (119/km²)
Founded	1804
Website	www.lex-co.com

KB

Current Wikipedia Article Structures

Free Form Text

Lexington County, South Carolina

From Wikipedia, the free encyclopedia

Coordinates: 33.90°N 81.27°W

Lexington County is a county located in the U.S. state of South Carolina. In 2000, its population was 216,014. The 2008 Census estimate placed the population at 248,518.[1] Its county seat is Lexington.[2]

Lexington County is part of the Columbia, South Carolina, Metropolitan Statistical Area.

Contents [hide]

1 Geography
  1.1 Adjacent Counties
2 Demographics
3 Cities and towns
4 Transportation
5 References

Lexington County, South Carolina

Location in the state of South Carolina

South Carolina's location in the U.S.

Seat	Lexington
Area	
- Total	758 sq mi (1,963 km²)
- Land	699 sq mi (1,810 km²)
- Water	59 sq mi (153 km²), 7.74%
Population	
- (2000)	216,014
- Density	308/sq mi (119/km²)
Founded	1804
Website	www.lex-co.com

Geography [edit]

According to the U.S. Census Bureau, the county has a total area of 758 square miles (1,963 km²), of which, 699 square miles (1,811 km²) of it is land and 59 square miles (152 km²) of it (7.74%) is water.

Adjacent Counties [edit]

- Richland County, South Carolina - east
- Orangeburg County, South Carolina - southeast
- Calhoun County, South Carolina - southeast
- Aiken County, South Carolina - southwest
- Saluda County, South Carolina - west
- Newberry County, South Carolina - northwest

Demographics [edit]

As of the census[3] of 2000, there were 216,014 people, 83,240 households, and 59,849 families residing in the county. The population density was 309 people per square mile (119/km²). There were 90,978 housing units at an average density of 130 per square mile (50/km²). The racial makeup of the county was 84.18% White, 12.63% Black or African American, 0.34% Native American, 1.05% Asian, 0.04% Pacific Islander, 0.79% from other races, and 0.98% from two or more races. 1.92% of the population were Hispanic or Latino of any race.

There were 83,240 households out of which 35.50% had children under the age of 18 living with them, 56.60% were married couples living together, 11.60% had a female householder with no husband present, and 28.10% were non-families. 22.50% of all households were made up of individuals and 6.90% had someone living alone who was 65 years of age or older. The average household size was 2.56 and the average family size was 3.01.

In the county, the population was spread out with 26.10% under the age of 18, 8.30% from 18 to 24, 31.60% from 25 to 44, 23.80% from 45 to 64, and 10.20% who were 65 years of age or older. The median age was 36 years. For every 100 females there were 94.50 males. For every 100 females age 18 and over, there

KB

Current Wikipedia Article Structures

Display Structures

Lexington County, South Carolina

From Wikipedia, the free encyclopedia

Coordinates: 33.90°N 81.27°W

Lexington County is a county located in the U.S. state of South Carolina. In 2000, its population was 216,014. The 2008 Census estimate placed the population at 248,518.[1] Its county seat is Lexington.[2]

Lexington County is part of the Columbia, South Carolina, Metropolitan Statistical Area.

Contents [hide]

1 Geography
  1.1 Adjacent Counties
2 Demographics
3 Cities and towns
4 Transportation
5 References

Geography [edit]

According to the U.S. Census Bureau, the county has a total area of 758 square miles (1,963 km²), of which, 699 square miles (1,811 km²) of it is land and 59 square miles (152 km²) of it (7.74%) is water.

Adjacent Counties [edit]

- Richland County, South Carolina - east
- Orangeburg County, South Carolina - southeast
- Calhoun County, South Carolina - southeast
- Aiken County, South Carolina - southwest
- Saluda County, South Carolina - west
- Newberry County, South Carolina - northwest

Demographics [edit]

As of the census[3] of 2000, there were 216,014 people, 83,240 households, and 59,849 families residing in the county. The population density was 309 people per square mile (119/km²). There were 90,978 housing units at an average density of 130 per square mile (50/km²). The racial makeup of the county was 84.18% White, 12.63% Black or African American, 0.34% Native American, 1.05% Asian, 0.04% Pacific Islander, 0.79% from other races, and 0.98% from two or more races. 1.92% of the population were Hispanic or Latino of any race.

There were 83,240 households out of which 35.50% had children under the age of 18 living with them, 56.60% were married couples living together, 11.60% had a female householder with no husband present, and 28.10% were non-families. 22.50% of all households were made up of individuals and 6.90% had someone living alone who was 65 years of age or older. The average household size was 2.56 and the average family size was 3.01.

In the county, the population was spread out with 26.10% under the age of 18, 8.30% from 18 to 24, 31.60% from 25 to 44, 23.80% from 45 to 64, and 10.20% who were 65 years of age or older. The median age was 36 years. For every 100 females there were 94.50 males. For every 100 females age 18 and over, there

Lexington County, South Carolina

Location in the state of South Carolina

South Carolina's location in the U.S.

Seat	Lexington
Area	
- Total	758 sq mi (1,963 km²)
- Land	699 sq mi (1,810 km²)
- Water	59 sq mi (153 km²), 7.74%
Population	
- (2000)	216,014
- Density	308/sq mi (119/km²)
Founded	1804
Website	www.lex-co.com

KB

Current Wikipedia Article Structures

Structured Schemas (Templates/Infoboxes)

Lexington County, South Carolina

From Wikipedia, the free encyclopedia

Coordinates: 33.90°N 81.27°W

Lexington County is a county located in the U.S. state of South Carolina. In 2000, its population was 216,014. The 2008 Census estimate placed the population at 248,518.[1] Its county seat is Lexington.[2]

Lexington County is part of the Columbia, South Carolina, Metropolitan Statistical Area.

Contents [hide]

1 Geography
  1.1 Adjacent Counties
2 Demographics
3 Cities and towns
4 Transportation
5 References

Geography [edit]

According to the U.S. Census Bureau, the county has a total area of 758 square miles (1,963 km²), of which, 699 square miles (1,811 km²) of it is land and 59 square miles (152 km²) of it (7.74%) is water.

Adjacent Counties [edit]

- Richland County, South Carolina - east
- Orangeburg County, South Carolina - southeast
- Calhoun County, South Carolina - southeast
- Aiken County, South Carolina - southwest
- Saluda County, South Carolina - west
- Newberry County, South Carolina - northwest

Demographics [edit]

As of the census[3] of 2000, there were 216,014 people, 83,240 households, and 59,849 families residing in the county. The population density was 309 people per square mile (119/km²). There were 90,978 housing units at an average density of 130 per square mile (50/km²). The racial makeup of the county was 84.18% White, 12.63% Black or African American, 0.34% Native American, 1.05% Asian, 0.04% Pacific Islander, 0.79% from other races, and 0.98% from two or more races. 1.92% of the population were Hispanic or Latino of any race.

There were 83,240 households out of which 35.50% had children under the age of 18 living with them, 56.60% were married couples living together, 11.60% had a female householder with no husband present, and 28.10% were non-families. 22.50% of all households were made up of individuals and 6.90% had someone living alone who was 65 years of age or older. The average household size was 2.56 and the average family size was 3.01.

In the county, the population was spread out with 26.10% under the age of 18, 8.30% from 18 to 24, 31.60% from 25 to 44, 23.80% from 45 to 64, and 10.20% who were 65 years of age or older. The median age was 36 years. For every 100 females there were 94.50 males. For every 100 females age 18 and over, there

Lexington County, South Carolina

Location in the state of South Carolina

South Carolina's location in the U.S.

Seat	Lexington
Area	
- Total	758 sq mi (1,963 km²)
- Land	699 sq mi (1,810 km²)
- Water	59 sq mi (153 km²), 7.74%
Population	
- (2000)	216,014
- Density	308/sq mi (119/km²)
Founded	1804
Website	www.lex-co.com

KB

These structures are insufficient for researchers, who:

- Need access to data en masse.
- Need strong citation structures.

(also relevant to WikiSpecies)

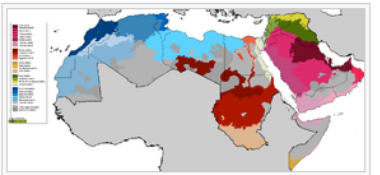
KB

Problem: Structures are display oriented, not semantic, and are used inconsistently

Example: Language Dialects

Language family	Sino-Tibetan • Sinitic • Chinese
Dialects	Mandarin Jin Huizhou Wu Hunanese Jiangxinese Hakka Yue (including Cantonese-Taishanese) Ping Shaojiang Northern Min Eastern Min (including Fuchow) Central Min Pu Xian Southern Min (including Amoy, Taiwanese)

The major dialect groups are:



Different Dialects of Arabic in the Arab World

Maghrebi Arabic [edit]

- Maghrebi Arabic includes Moroccan Arabic, Algerian Arabic, Algerian Saharan Arabic, Tunisian Arabic, and Libyan Arabic, and is spoken by around 45 million North Africans in Morocco, *Western Sahara*, Algeria, Tunisia, Libya, Niger, and western Egypt; it is often difficult for speakers of Middle Eastern Arabic varieties to understand. The Berber influence in these dialects varies in degree.[14]

Egyptian Arabic [edit]

- Egyptian Arabic, spoken by around 76 million in Egypt. It is one of the most understood varieties of Arabic, due in large part to the widespread distribution of Egyptian films and television shows throughout the Arabic speaking world. Closely related varieties are also spoken in Sudan.

Gulf Arabic [edit]

- Gulf Arabic, spoken by around 34 million people in Arab states and eastern Saudi Arabia.
- Iraqi Arabic, spoken by about 29 million people, with significant differences between the Arabian-like dialects of the south and the more conservative dialects of the north. Closely related varieties are also spoken in Iran, Syria, and Turkey.
- North Mesopotamian Arabic, spoken by around 7 million people in northern Iraq, northern Syria and southern Turkey.

Levantine Arabic [edit]

- Levantine Arabic includes North Levantine Arabic, South Levantine Arabic, and Cypriot Arabic. It is spoken by almost 35 million people in Lebanon, Syria, Jordan, The Palestinian territories, Israel, Cyprus, and Turkey. It is also called Mediterranean Arabic.

Other [edit]

Other varieties include:

List of dialects of the English language

From Wikipedia, the free encyclopedia

This article **does not cite any references or sources**.
Please help improve this article by adding citations to reliable sources. Unsourced material may be challenged and removed. *(December 2006)*

This is a **list of dialects of the English language.** Dialects are linguistic varieties which differ in pronunciation, vocabulary and grammar from each other and from Standard English (which is itself a dialect).

Dialects can be usefully defined as "sub-forms of languages which are, in general, mutually comprehensible".[1] British linguists distinguish dialect from accent, which refers only to pronunciation. Thus, any educated English speaker can use the vocabulary and grammar of Standard English, but different speakers use their own local words for everyday objects or actions, regional accent, or Received Pronunciation, which within the U.K. is considered an accent distinguished by class rather than by region. American linguists, however, include pronunciation differences as part of the definition of regional or social dialects. The combination of differences in pronunciation and use of local words may make some English dialects almost unintelligible from one region to another.

KB

Problem: Historical diffs cannot be organized or searched by structure

e.g.: Diffs of Chinese Language

(cur) = difference from current version, (prev) = difference from preceding version, **m** = minor edit, → = section edit, ← = automatic edit summary
(latest | earliest) View (newer 50 | older 50) (20 | 50 | 100 | 250 | 500)

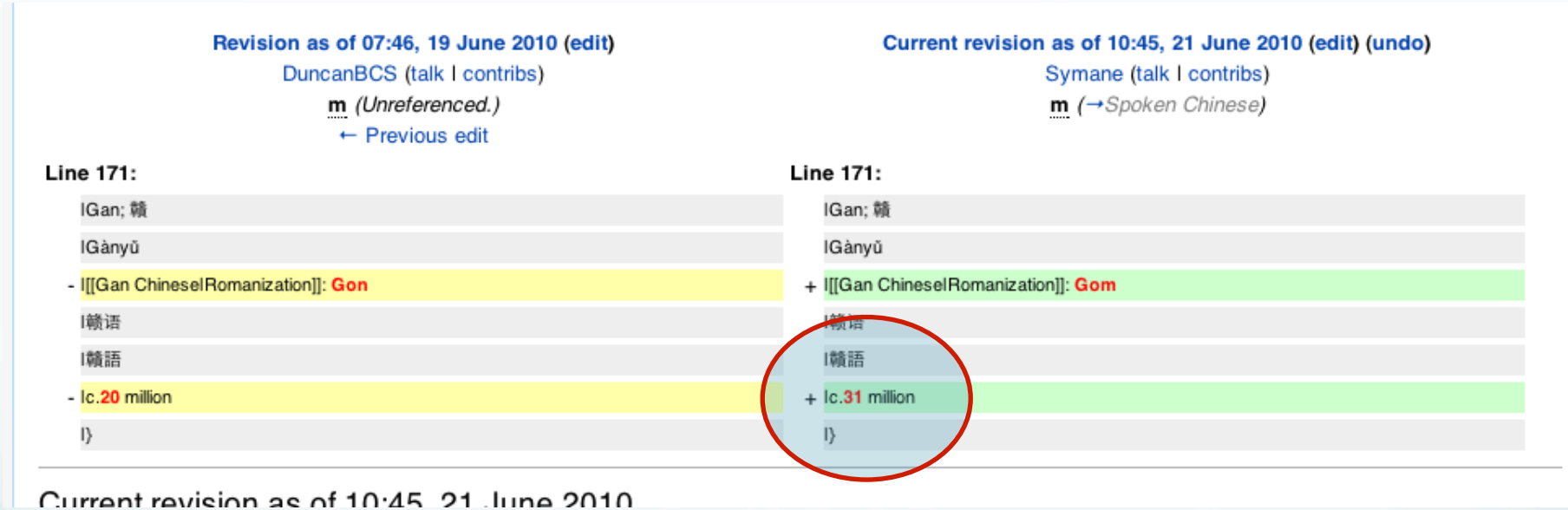
Compare selected revisions

- (cur | prev) 10:45, 21 June 2010 Symane (talk | contribs) **m** (63,712 bytes) *(→Spoken Chinese)* (undo)
- (cur | prev) 07:46, 19 June 2010 DuncanBCS (talk | contribs) **m** (63,712 bytes) *(Unreferenced.)* (undo)
- (cur | prev) 04:55, 19 June 2010 112.213.215.104 (talk) (63,851 bytes) (undo)
- (cur | prev) 09:45, 18 June 2010 Crescent8 (talk | contribs) **m** (63,712 bytes) (undo)
- (cur | prev) 03:25, 16 June 2010 Xqbot (talk | contribs) **m** (63,700 bytes) *(robot Removing: zu:IsiShayina)* (undo)
- (cur | prev) 17:38, 11 June 2010 BilCat (talk | contribs) **m** (63,718 bytes) *(Reverted edits by Mr.Right.Perfect (talk) to last version by TheLeopard)* (undo)
- (cur | prev) 17:14, 11 June 2010 Mr.Right.Perfect (talk | contribs) **m** (63,774 bytes) (undo)
- (cur | prev) 03:28, 7 June 2010 TheLeopard (talk | contribs) (63,718 bytes) *(Grammer)* (undo)
- (cur | prev) 17:15, 6 June 2010 Stevertigo (talk | contribs) **m** (63,721 bytes) (undo)
- (cur | prev) 07:49, 6 June 2010 Kwamikagami (talk | contribs) (63,723 bytes) *(add dialects to info box)* (undo)
- (cur | prev) 12:42, 5 June 2010 Beirne (talk | contribs) (63,167 bytes) *(Reverted 1 edit by 91.178.127.87; The Chinese languages are generally mutually unintelligible. (TW))* (undo)
- (cur | prev) 11:08, 5 June 2010 91.178.127.87 (talk) (63,165 bytes) (undo)
- (cur | prev) 06:52, 5 June 2010 72.190.75.130 (talk) (63,167 bytes) *(→Eduction)* (undo)
- (cur | prev) 19:07, 3 June 2010 ILVTW (talk | contribs) **m** (63,166 bytes) (undo)
- (cur | prev) 18:13, 2 June 2010 Courcelles (talk | contribs) **m** (63,141 bytes) *(Reverted edits by 142.24.53.237 (talk) to last revision by Bjmullan (HG))* (undo)
- (cur | prev) 18:12, 2 June 2010 142.24.53.237 (talk) (empty) *(→Writing)* (undo)
- (cur | prev) 18:12, 2 June 2010 Bjmullan (talk | contribs) **m** (63,141 bytes) *(Reverted 2 edits by 142.24.53.237 identified as vandalism to last revision by Wavelength. (TW))* (undo)
- (cur | prev) 18:11, 2 June 2010 142.24.53.237 (talk) (61,227 bytes) *(→Language and nationality)* (undo) *(Tag: section blanking)*

KB

Problem: Contributors tend to not name their sources.

Example: Where did the contributor get the fact that there are 31 million Gan speakers?



Revision as of 07:46, 19 June 2010 (edit) DuncanBCS (talk | contribs) m *(Unreferenced.)* ← Previous edit	Current revision as of 10:45, 21 June 2010 (edit) (undo) Symane (talk | contribs) m *(→Spoken Chinese)*
Line 171:	**Line 171:**
|Gan; 贛	|Gan; 贛
|Gànyǔ	|Gànyǔ
- |[[Gan Chinese|Romanization]]: **Gon**	+ |[[Gan Chinese|Romanization]]: **Gom**
|贛语	|贛语
|贛語	|贛語
- |c.**20** million	+ |c.**31** million
|}	|}

Current revision as of 10:45, 21 June 2010

KB

Problem: Wikipedia article structure promotes consensus rather than supporting simultaneous differing viewpoints when none is dominant

Example: Taxonomic organization

KB

A working prototype

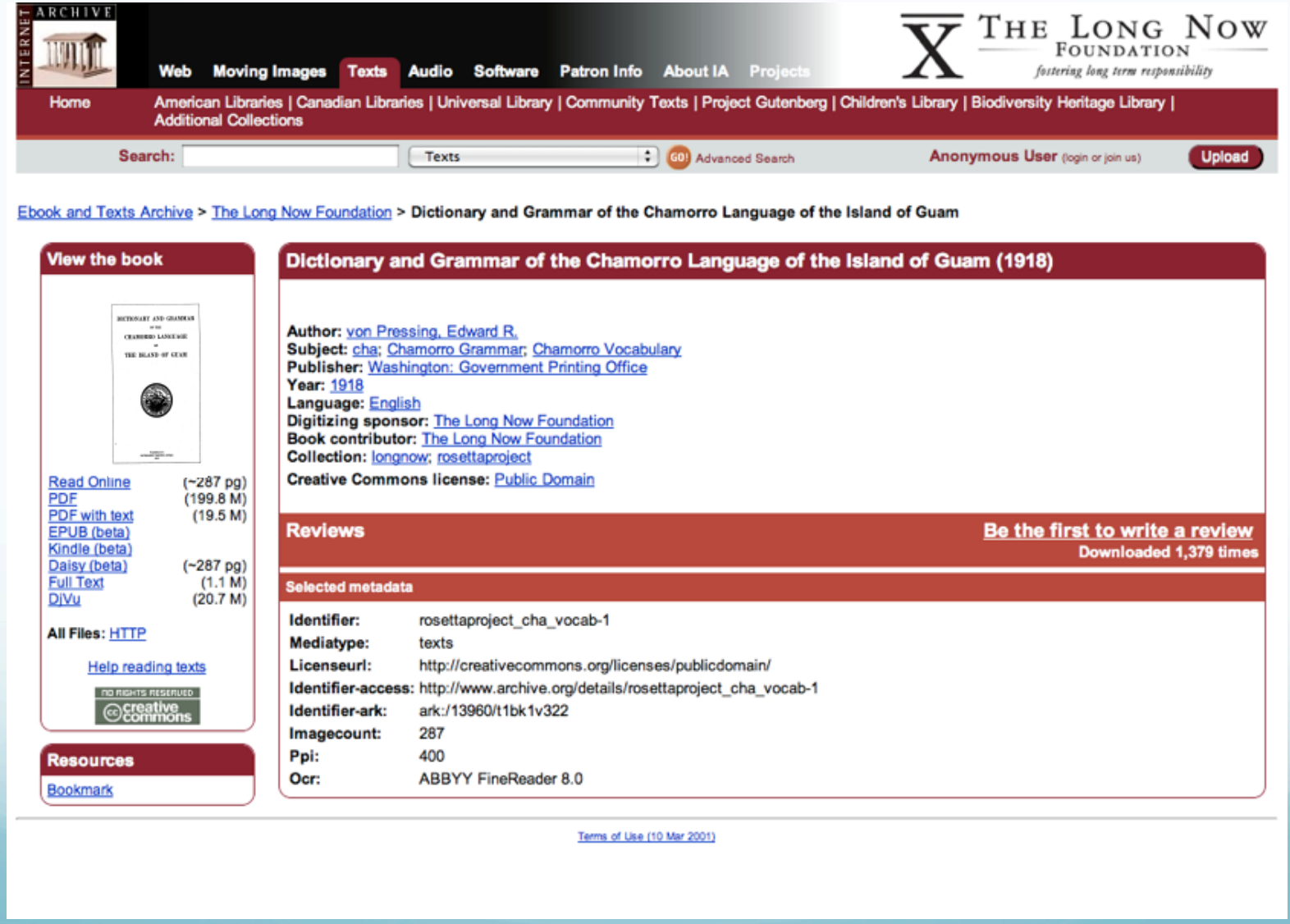
Our example of a possible solution using distributed sources

LW

Rosetta Project
All-language Wiki Prototype

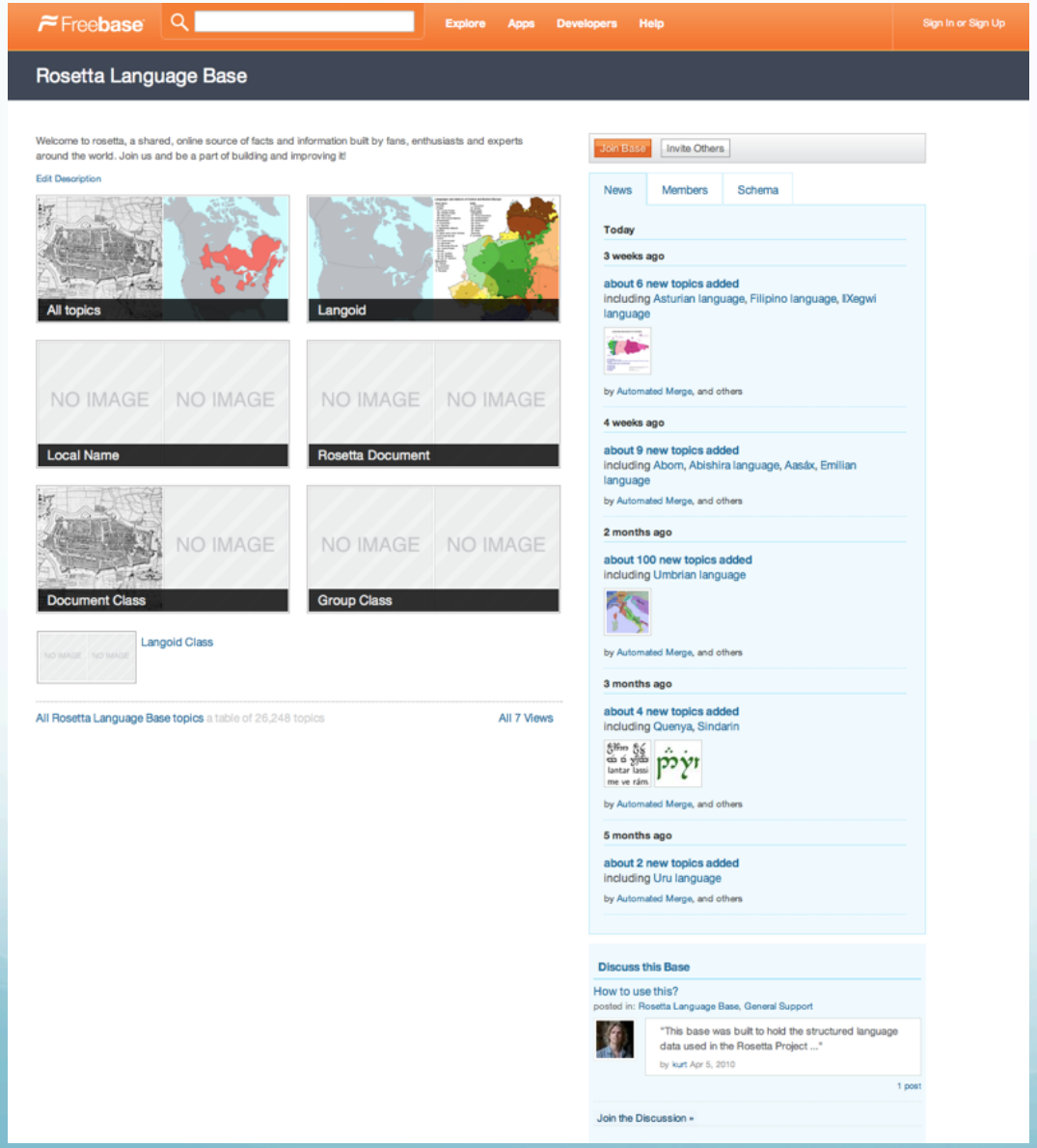
LW

Internet Archive
Rosetta Project Special Collection



LW

All Language Metadata Now in Freebase



- Rosetta Base: over 10,000 languages and linguistic entities linked by language family relationship
- All data is linked to other kinds of data in Freebase
- We have rectified ~1500 Wikipedia pages about human languages to our data set

LW

Rosetta Alpha Wiki

The Rosetta Project

page | discussion | view source | history

Log in

navigation
- Main Page
- Community portal
- Current events
- Recent changes
- Random page
- Help

search
Go Search

toolbox
- What links here
- Related changes
- Upload file
- Special pages
- Printable version
- Permanent link

Swahili Language (swh)

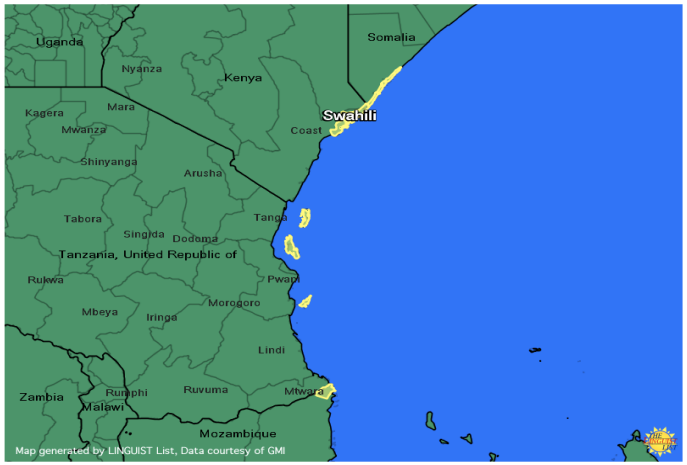
Also Known As: Swahili,Suahili,Kisuahili,Kiswahili,Kiswaheli,Kisuaheli,Arab-swahili

Description:

Swahili (Kiswahili) is spoken by various ethnic groups that inhabit several large stretches of the Indian Ocean coastline from southern Somalia to northern Mozambique, including the Comoros Islands. Although only 5-10 million people speak it as their native language, Swahili is also a lingua franca of much of East Africa and the Democratic Republic of the Congo, is a national or official language of four nations, and is the only language of African origin among the official working languages of the African Union.

Swahili is a Bantu language that serves as an alternative first language to various groups traditionally inhabiting about 1,500 miles of the Southeast African coast. About 35% of the Swahili vocabulary derives from the Arabic language, gained through more than twelve centuries of contact with Arabic-speaking traders. It also has incorporated German, Portuguese, English and French words into its vocabulary through contact during the last five centuries. Swahili has become a second language spoken by tens of millions in three countries, Tanzania, Kenya, and Congo (DRC), where it is an official or national language. The neighboring nation of Uganda made Swahili a required..... *full article at Wikipedia*

Location of Swahili Language Speakers



Overview

Main Country: Tanzania

Spoken In: ***Countries:*** Kenya, Tanzania, Uganda, Rwanda, Burundi, Mozambique, Malawi, Democratic Republic of the Congo, Zambia
Regions: Africa

ISO 639-1 Code: sw

ISO 639-3 Code: swh

Classification Taxonomy

All Languages
Niger-Congo Group
Atlantic-Congo Group
Volta-Congo Group
Benue-Congo Group
Bantoid Group
Southern Bantoid Group
Narrow Bantu Group
Central Narrow Bantu Group
G Group
Swahili (G.40) Group
Swahili Language

Rosetta Document Collection

Detailed Description:
1 (download) (browse)

Genesis Translation:
1 (download) (browse)
2 (download) (browse)

Grammar:
1 (download) (browse)

Miscellaneous:
1 (download) (browse)

Orthography:
1 (download) (browse)

Phonology:
1 (download) (browse)

Universal Declaration Of Human Rights:
1 (download) (browse)

This page was last modified on 13 August 2009, at 20:30. This page has been accessed 18 times. Privacy policy About Testwiki Disclaimers

Powered By MediaWiki

Freebase

Freebase

LW/KB

Language Commons

LW

Please help!

- We plan to build The Language Commons Wiki – an aggregation and discovery portal for information and resources on **all** 6,900 human languages – and we need your help!
- Here are just some of the questions we have:
  - Does this satisfy the needs of researchers, native speakers of rare languages, and students?
  - What should the relationship between the existing Wikipedia and the Language Commons be? (e.g. can it just be a source too?)
  - How do we introduce this new way of editing a wiki article to editors?
  - Who do we need to talk to? Who should we got involved?

KB

Thank you!

Questions / Comments / Ideas:

Laura Welcher
laura@longnow.org