

Welcome to the ColCat Wiki!

The ColCat Wiki is designed as a collaborative research space where color categorization data from a wide variety of languages and dialects can be examined, downloaded, searched, and queried for the purpose of non-commercial research on color cognition and language across language groups. It is a public-access platform created to make available unpublished raw data sets from Dr. Robert E. MacLaury's Color Categorization Archive and other contributed color categorization datasets. Beyond providing public-access to heretofore private color categorization data, the ColCat Wiki also provides tools, data handling utilities, and other resources that can be used by researchers to carry out investigations on a range of topics in the area of color categorization and naming. As part of the wiki's public-release the ColCat Team provides, via an intuitive user-friendly interface, downloads of digitally addressable data sets for a portion of the MacLaury Archive, downloadable raw data image scan files for the entire data in MacLaury Archive, tools for searching and querying and downloading specific portions of the archive as well as searching and querying the existing World Color Survey archive, tools and guides to facilitate manual transcription of yet-transcribed raw data in a standardized format, utilities and processes for creating survey dictionaries and lists of dictionary synonyms, links out for participating in crowdsourced research investigations to help further populate the wiki with transcriptions of digitally-addressable datasets, user forums for engaging in collaborative community dialogue and for troubleshooting research and archive issues, and an array of original research tools, maps of survey locations, linkable bibliographies, supporting documents, relevant reference databases and archives, and more. Please take some time to explore the ColCat wiki, and let us know if you have any questions, comments or suggestions (colcat@calit2.uci.edu).

This UserDocumentation.pdf is a summary of the wiki's features and how to use them, describing the features we have on-board as of public release June 30, 2016. At public-release, downloadable content is provided for 38 surveys, including raw data image scans in .pdf format and transcribed data in digitally addressable format (in two forms).

Acknowledgments

Support for this project was provided in part by Research Awards from *The University of California Pacific Rim Research Program*, 2010-2015 (K. A. Jameson, PI), and *The National Science Foundation* 2014-2017 (#SMA-1416907, K.A. Jameson, PI), and by UC Irvine's *California Institute for Telecommunications and Information Technology* (Calit2).



This work is licensed to the Authors under Creative Commons Attribution-Noncommercial-NoDerivatives Works 4.0 International License. December 31, 2015.

Jameson, K. A., Gago, S., Deshpande, P.S., Benjamin, N.A., Chang, S.M., Tauber, S., Jiao, Y., Harris, I.G., Xiang, Z., Huynh, B.B., Ke, H., Lee, W.J., MacLaury, R.E. (2016). "The Robert E. MacLaury Color Categorization (ColCat) Digital Archive." [http:// colcat.calit2.uci.edu/](http://colcat.calit2.uci.edu/). The California Institute for Telecommunications and Information Technology (Calit2). University of California, Irvine.

User Documentation Contents:

[ColCat Wiki public-release \(June 30, 2016\) UserDocumentation.pdf.](#)

[User Documentation Contents:](#)

[Welcome to the ColCat Wiki!](#)

[Home](#)

[Archive Contents Page](#)

[Mesoamerican Language Datafiles](#)

[Multi-National Language Datafiles](#)

[File Naming Convention for Datasets](#)

[Survey Data Types](#)

[Commenting Utility](#)

[Language Page Contents and Organization](#)

[Organization and Location of Bilingual Participant data in the Robert E. MacLaury archive:](#)

[Data Handling Utilities](#)

[Robert E. MacLaury PhD](#)

[People](#)

[Resources](#)

[Research Tools](#)

[Forums](#)

[What is the Forum for....](#)

[ColCat Wiki public-release \(June 30, 2016\) UserDocumentation.pdf.](#)

Home

<http://colcat.calit2.uci.edu/tiki-index.php>

- A summary describing the digital archive of Robert E. MacLaury's color categorization data from approximately 210 surveys that can be downloaded, searched, and queried.
- Summaries of the Mesoamerican Color Survey and Multinational Color Survey.
- Links to a .pdf of ColCat Wiki Terms and Conditions, link to Download the ColCat Wiki User Documentation, links out to Supporting Documents, that include items such as instructions from MacLaury to field researchers conducting the Color Survey.

Archive Contents Page

<http://colcat.calit2.uci.edu/tiki-index.php?page=Archive%20Contents>

The Archive Contents page show language lists for two separate surveys collected by MacLaury. These are the Mesoamerican survey and the Multinational Survey (see Appendix 1 at end of document). The June 30, 2016 release of the website uses red font typeface to indicate the language surveys that provide digitized data content (These include surveys for Chinantec, Mixtec, Zapotec, American English, Cantonese, Japanese, Korean, Korean-English-Bilinguals, Shanghainese, Baluchi, Hindi, Lani, and Sindhi). At the time of the public-release, all other language surveys listed on this page will have only “raw data image scans” that are also viewable within the browser for users to download and work with (if they are interested in manually transcribing raw data image scans, users should see the standardized Manual Transcription Tools provided on the site's Research Tools page).

Mesoamerican Language Datafiles

This area contains a searchable database of 116 Mesoamerican Language survey .pdf scans published by MacLaury (1997).

Multi-National Language Datafiles

This area contains a searchable database of scanned language survey pdfs of 92 Multi-National Languages.

File Naming Convention for Datasets

Filenames on the site and of downloaded files follow naming conventions that reproduce the organization of files and folders in MacLaury's paper archive. In both scanned raw image data and manually transcribed data files, the "Folder" designation in a downloaded filename denotes an entire survey done by specific investigator(s) at one location. Often raw data image scans of such surveys are distributed across more than one .pdf file (to keep file sizes manageable and to preserve groupings in the paper archive). When more than one image scan .pdf files exists, raw image data are sequentially named, for example the two .pdf files associated with Folder011 are "Folder011.1 and Folder011.2", etc. Filename conventions for .csv data file downloads use this same standardized naming convention.

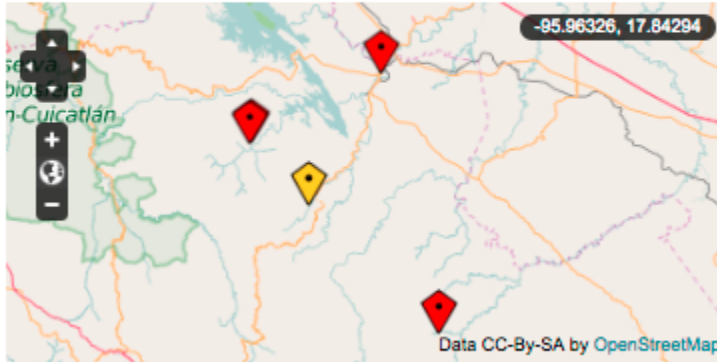
Survey Data Types

- The demographic information contains the following about each informant (available only as part of information bundled in Manual Transcription (.zip) downloads).
 - informant number, age, date, occupation or years in school, place of birth, years of residence in location under investigation, other general comments made by the investigator.
- Naming data. Free-naming response data from the archive (similar to that found in the WCS).
- Focus data. Best-exemplar, or category focus, data from the archive (similar to that found in the WCS).
- Mapping data. Data unique to the MacLaury archive representing a "mapping task" that conveys denotative ranges of color terms elicited from participants in the survey. Informally this may elsewhere be referred to as the "rice mapping" task.

Language Page Contents and Organization

Each language page includes an embedded map indicating the approximate location of where the dataset was collected or locations where the language originates. Below the map, users can download the raw data of each informant from a series of folders. Users can additionally download transcriptions of the raw data transcribed in the following methods: manual or crowdsourced. Language page links for optical character recognition (OCR) results

Language: Chinantec



Download all Chinantec Digitized Data in .csv Format. 

Folder011-Chinantec

Download all Folder011-Chinantec Raw Data Image Scans (.zip)

Download survey folders individually (.pdf):

- [Raw Data Folder 011.1](#)
- [Raw Data Folder 011.2](#)

Download Manual Transcription (.zip) (Transcribed by: Sean Tauber, UC Irvine)

Crowdsourced Transcription

OCR Transcription

Other Resources on Chinantec:

- <http://www.ethnologue.com/language/cco> 
- <http://www.language-archives.org/language/cco> 
- <http://glottolog.org/resource/languoid/id/coma1246> 

on the raw image data scans will be updated as they become available. All the language folders can be downloaded as one .zip or can be individually downloaded as a .pdf. Users can also download the language folders as a .csv. Also provided are links to resources specific to the language represented, including www.ethnologue.com, www.language-archives.org, and www.glottolog.org which provide users additional references for language analysis such as language family, language status, and citations to published work on the language. Additionally, users can make comments at the bottom of the language page.

Downloading data from language pages: Each language page follows a similar organization, and data file downloads aim for a standardized filename format. For example, as shown in the example, for the eleventh survey in the archive, a Chinantec survey, in which the survey was done by specific investigator(s) at one location, is contained in "Folder 011" in the MacLaury

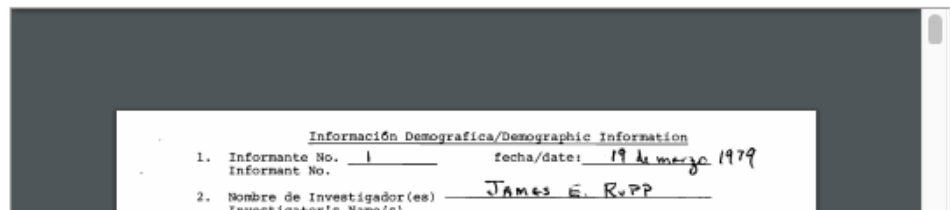
paper archive. Many folders in the survey have multiple .pdfs. Sometimes these represent surveys in a different dialect or different locale of the language. Two separate .pdf files containing raw image data are associated with the Folder011. Following the link “Raw Data Folder 011.1” will take you to the following screen where you can view (in the .pdf preview window) and download (by the link) the individual .pdf file named “Folder011.1-Chinantec_1.pdf.” If you want to first view files before downloading, you can do so at the links provided. However, we do not recommend downloading from the preview window as the filenames will be generic (e.g., from preview all files will be downloaded as “tiki-download_file.pdf” and will not have standardized filenames with vital folder and version tracking information).

Survey: 011.1

Files:

- [Survey011.1_Raw_Data_1_of_1](#)

-
- Preview for Survey011.1_Raw_Data_1_of_1:



Información Demografica/Demographic Information	
1. Informante No. / Informant No.	1 fecha/date: 19 de mayo 1979
2. Nombre de Investigador(es) / Investigator's Name(s)	James E. Rupp

File Types available on Language Pages::

Raw Data Image Scans (.pdf): Contains scanned raw data in .pdf format of surveyed informants responses to the color survey recorded on tables. Other data also in .pdf format include information about the informant's age, gender, place of birth, years of residence in the region and other comments. Individual characters within the pdfs cannot be queried since the pdf acts as a snapshot of the original data.

The order and content of image scan .pdfs exactly replicate that of MacLaury's paper archive folder and page organization. This was considered important to preserve the original structure of the archive, as well as to include all the ethnographic notes, data summary figures and pages created by the investigators, and study correspondence in each folder. Often duplicates of participant data are found in the .pdf files. This was preserved to exactly duplicate the paper archive. Other features of the scans (especially for “mapping” data) include some images where the use of color on a page reflects meaningful data that cannot be disambiguated by a grayscale image scan. In such cases, the pages were scanned at high color resolution,

thereby capturing the important information in participant's mapping data. The ethnographic notes and correspondence are not transcribed as part of the ColCat digitized data, but image scans of these in the .pdf files may be of interest to researchers looking to understand details of experimental sessions conducted.

Name	Date Modified	Size	Kind
Folder011-Chinantec.pdf_1.xlsx	Yesterday, 12:01 PM	141 KB	Spreadsheet
Folder011-Chinantec.pdf_Demographics.xlsx	Yesterday, 11:04 AM	10 KB	Spreadsheet
Folder011-Chinantec.pdf_Dictionary.xlsx	Yesterday, 11:30 AM	10 KB	Spreadsheet
Folder011-Chinantec.pdf_Unique_Response.xlsx	Yesterday, 11:16 AM	10 KB	Spreadsheet

Chinantec Language download

Digitized survey data for Chinantec are available for download on this page. Select the desired task type (*Naming/Foci/Mapping*) and transcription method (currently only *Manual Transcription* data is available). All of the surveys with checkmarks in the right-most column will be downloaded in a single comma-delimited (.csv) file that is utf-8 encoded. You can select/deselect any specific surveys you wish to be included/excluded using the checkboxes. Click the "Download transcription data" link in order to initiate the download.

Further details about the file format of downloaded data can be found in the [ColCat Digitized Data Format Guide](#).

You can also download a list of the unique terms for all of the currently selected surveys by clicking the "Download unique response list" link. This file can be used as the starting point for creating a set of user-defined equivalence classes for terms in the dataset. Instructions for this process can be found [here](#).

<div> Naming Foci Mapping </div> <div>Manual Transcription</div>							
Survey	Transcription Method	Task	Females	Males	Total Participants	Unique Responses	Download
11.000	Manual	Naming	0	1	1	37	<input checked="" type="checkbox"/>
12.000	Manual	Naming	13	11	24	23	<input checked="" type="checkbox"/>

Upload term equivalence map and apply to download:

[Choose File](#) No file chosen

Warning! The data you are about to download contains UTF-8 encoded Unicode characters. **DO NOT** open this file with Microsoft Excel because it does not support this file format. Please see user documentation for further details.

[Download transcription data](#)

[Download unique response list](#)

It is also important to note that while in the raw data image scans, there are many instances of duplicated individual participant data, as well as a tendency for data to be recorded by study investigators in several different formats (for example, (1) a participant's list format of complete naming data (2) a head lexeme table format of the same participant's naming data, and (3) a modifier table format of the same naming data), this kind of duplication is not represented in the digitized manual transcription datasets that are available for analyses. In other words, each digitized data file on this site provides the unique data from an individual participant (usually in the form of naming, mapping and focus data on a single spreadsheet), and these data are not replicated elsewhere in the digital downloads on the archive.

On each language page a link to download "... Digitized Data in .csv Format" can be seen. This link opens a new language download window, as shown above, where different forms of data can be downloaded. The format available at these links aims to follow the WCS data format as much as possible. Note, also the advisory: "*Warning! The data you are about to download contains UTF-8 encoded Unicode characters. DO NOT open this file with Microsoft Excel because it does not support this file format.*" Failure to preserve Unicode character formatting in the file will result in loss of important foreign-language character features in the downloaded data. Please see additional user documentation contained in the "ColCat Digitized Data Format Guide" from the link provided for further details.

Also, note that on each language download page a link to "Unique Response Lists" (or the exhaustive list of all responses observed in a dataset) for all digitized datasets is provided, as shown in the Chinantec example above. Thus, for the Chinantec example the link downloads a file named "*Chinantec_Unique_Responses_for_2_Surveys_Manual_Transcription.csv*" containing the exhaustive list of all responses in that dataset. In order to preserve Unicode characters needed for foreign language transcription these files should only be opened with UTF-8 compatible programs (i.e., opening with not UTF-8 editors or spreadsheets like Microsoft Excel will result in errors or losses of information in the datafile). Please see relevant information contained in the "ColCat Digitized Data Format Guide" for further details.

Finally, as shown in Appendix 1 of MacLaury (1997, p. 397), some data in MacLaury's Mesoamerican database was collected from informants who may also have been surveyed in the WCS. These consist of 23 surveys (as shown in Table I.1, Appendix 1 of MacLaury 1997) and while a portion of these participants have been surveyed by the WCS, the data collected as part of the MCS does not always duplicate the WCS observations, in the sense that it sometimes additionally collects data for the mapping task, which was not part of the WCS procedure. Notes are provided on each language page where such folders are relevant. These include: Folder002-Aguacatec, Folder004-Amuzgo, Folder012-Chinantec, Folder020-Guarijio, Folder022-Huasteco, Folder023-Huave, Folder028-Jicaque, Folder041-Mazahua, Folder043-Mazatec, Folder051-Mixtec, Folder057-Nahuatl, Folder065-Papago, Folder068-Paya, Folder069-Pima, Folder081-Seri, Folder082-Tarahumara(Central), Folder083-Tarahumara(Western), Folder087-Tlapanec, and Folder115-Zapotec.

Manual Transcription Datasets(.pdfs): Data that have been manually transcribed by ColCat developers or by verified ColCat Wiki users with a background in color categorization research and/or linguistics. Screen grabs of the manual transcription data spreadsheet formats are illustrated below. Manually transcribed data strove to maintain complete accuracy of what was present in the paper archive documents.

Sample Naming Data

HEAD NAMING	language	Ss#	age	m/f	page	file	
	50	7	48	f	2	Folder050.2-Mixtec/6.pdf	
START NAMING		A	kwíxj	10	1	2	3
		B	kwíxj	9	#	ndížaa	kwíxj
		C	#	8	náa color žæ'æ	žæ'æ	rósæ ndížaa
		D	ndížaa	7	ndížaa	žæ'æ	kwāā ndížaa
		E	#	6	kwā'á	kwā'á ndížaa	náa jā kwāā tú
		F	#	5	kwā'á	kwā'á	kwā'á
		G	#	4	kwā'á	kwā'á	kwā'á
		H	tūú	3	kwā'á	kwā'á	kwā'á
		I	tūú	2	kwā'á	náa kwā'á tú	kwā'á tú
END NAMING		J	tūú	1	1	2	3

Sample Mapping Data

HEAD MAPPING	language	Ss#	age	m/f	page	file	
	50	7	48	f	17	Folder051.1-Mixtec/1.pdf	
START MAPPING		A	kwíxj ~ žaa	10	1	2	3
		B		9	kwíxj ~ žaa	kwíxj ~ žaa	kwíxj ~ žaa
		C		8	ndížaa	ndížaa	ndížaa
		D		7	ndížaa ~ žæ'æ	ndížaa ~ žæ'æ	ndížaa ~ žæ'æ
		E		6	žæ'æ	žæ'æ	žæ'æ
		F		5	žæ'æ	žæ'æ	žæ'æ
		G		4			
		H		3			
		I		2			
END MAPPING		J		1			
					1	2	3
HEAD FOCUS	language	Ss#	age	m/f	page	file	
	50	7	48	f	2	Folder050.2-Mixtec/6.pdf	

Sample Focus Data

HEAD FOCUS	language	Ss#	age	m/f	page	file		
	50	7	48	f	2	Folder050.2-Mixtec/6.pdf		
START FOCUS		A	kwixi	10	1	2	3	4
		B		9				
		C		8				
		D		7	kwā'a ndizaa	kwā'a ndizaa		
		E		6				kwā'a βæ'aæ
		F		5				
		G		4		kwā'a		
		H		3				kwā'a tū
		I		2			kæfé	
END FOCUS		J	tūu	1				
					1	2	3	4

Manual Transcription Datasets. Each language page shows a link to download the manual transcription data as .zip files, which also include for each survey in a given folder (1.) A transcribed demographic information spreadsheet containing all surveyed participants in the folder, (2.) all the survey's transcribed data for each participant on individual .xlsx spreadsheets, and (3.) a dictionary created from the transcribed survey data. The contents of each of these filetypes are now described:

- Demographic Information:
- Individual data Spreadsheets:
- **Folder Dictionary:** At the time of public-release, folder dictionaries contain the “dictionary” information recorded by investigators on the raw data sheets. Column A contains a list of abbreviations, Column B contains a list of color terms, Column C contains English color translations, Column D has Term Type (“H” for head lexeme or “M” for modifier). These column entries are strictly based on what investigators specified in the raw data .pdfs, no interpretation or extra information was added by the ColCat transcribers. MacLaury's instructions to investigators suggest these classification categories to investigators, and in transcribing the data the ColCat team deemed it important to preserve this in the folder dictionary (See MacLaury's instructions as the first link on: <http://colcat.calit2.uci.edu/tiki-index.php?page=Supporting+Documents>).
- **Unique Response List:** As illustrated with the Chinantec example above, the .zip file for all language manual transcriptions will include a “Unique Term List” that is generated by collecting all the unique response strings that can be **observed in a survey's manual transcription**. Here “unique response string” can be defined as either (1) an abbreviation of one or more alpha-characters, (2) a one word string, (3) a two or more word string, or (4) an abbreviation and a word string. This unique term list is provided as a guide to researchers to permit evaluation of the contents of the transcribed data spreadsheets from a more informed perspective.

Crowdsourced Transcription Datasets (.pdfs): Data that have been transcribed by Amazon's Mechanical Turk in which online participants are paid to manually transcribe the raw data. To be provided in future releases of the ColCat wiki.

Optical Character Recognition Datasets(.pdfs): Data that have been transcribed using software that interprets text and handwriting. To be provided in future releases of the ColCat wiki.

Organization and Location of Bilingual Participant data in the Robert E. MacLaury archive:

In many of the surveys in the REM database, investigator notes clearly indicate some participants are not entirely monolingual speakers of the languages assessed, and instead have some knowledge of another language (e.g., in the case of the mesoamerican survey data many participants have knowledge of Spanish), or they have varying degrees of bilingualism with English (especially in the multinational survey data).

ColCat organization of the data archive has adopted the policy to only segregate multilingual participants' data when they are clearly indicated by investigator notes as bilingual and were assessed in both languages in the survey. With the exception of such assessed "bilinguals", for other participant data in the archive, ColCat data organization presumes that if users download a .zip or spreadsheet dataset for, say, Chinantec language, then (if it is important) it is the responsibility of that user to additionally download the .pdfs for that data to determine which participants had knowledge of other languages. (In addition, the user may want to also refer to the "other languages known to informant" section of the demographic spreadsheet if the language data has been transcribed.) Thus, the decision to handle the data in ways that recognize participants' potential knowledge of other naming systems is entirely left to the user who downloads data and deems it important when performing data analyses.

Two "bilingual" surveys exist in the MacLaury database. Bilingual surveys are defined as those in which a set of participants were assessed in *both* a native language mode ("L1") and a non-native language mode ("L2"). Most frequently the L2 mode was English. In these cases the data from these bilingual participants are separated from the other participants (because they are clearly identified by the investigators as bilingual and participated in the survey in both languages), and those data are always provided as the last folder on the language page corresponding to the language mode in which the data is recorded.

For example, the Korean Language page has the following folder structure:

Language: Korean



[Download all Korean Digitized Data in .csv Format.](#)

Folder184-Korean

[Download all Folder184-Korean Raw Data Image Scans \(.zip\)](#)

Download survey folders individually (.pdf):

- [Raw Data Folder 184.1](#)
- [Raw Data Folder 184.2](#)
 - [Download Manual Transcription \(.zip\)](#) (Transcribed by: Prutha S. Deshpande & Kimberly A. Jameson, UC Irvine)
 - Crowdsourced Transcription
 - OCR Transcription

Korean Data from Korean-English Bilinguals ([Link to English language data for these Bilingual Participants](#))


[Download all Korean-Bilingual Digitized Data in .csv Format.](#)

[Download all Korean-Bilingual Raw Data Image Scans \(.zip\)](#)

- [Raw Data Folder 184.3](#)
 - [Download Manual Transcription \(.zip\)](#) (Transcribed by: Prutha S. Deshpande, UC Irvine)
 - Crowdsourced Transcription
 - OCR Transcription

In which it is clear that the leading folders on the page provide links to download the Korean digitized dataset; and the image scans of folders of those data and links where folders of those .pdf files can be viewed and individually downloaded (Note, however, downloads from the .pdf viewer feature require the user to insert the appropriate filename identifying folder name and number because the default filename is a generic “tiki-download_file.php.” We recommend that downloads of individual surveys are done from the list of Survey links show above the preview windows.); and a link providing .zip of the manually transcribed spreadsheet-organized data for download.

The heading after the above mentioned options is “**Korean Data from Korean-English Bilinguals**”. The contents of the links under this heading are very similar to those described above, except they vary in that they contain only Korean language data from participants who were noted as bilinguals in the survey, and who were additionally run in English language mode for the same tasks. The latter English language data appears in an analogously labeled bilingual folder at the bottom of the English language page.

Note, the leading download link on all language pages containing bilingual data -- for example, Korean: [Download all Korean Digitized Data in .csv Format](#)  -- Do Not contain data from bilinguals via this .csv download. Additional .csv links are supplied for the separated subset of, for example Korean, bilingual respondents.

The list of language pages that include bilingual data folders are the following: Cree and Korean.

Data Handling Utilities

Several on-board data-handling tools are provided that permit search and query of the contents of the database, and identifying the desired data files for download. These include:

- Searching languages by the minimum or maximum number of speakers and by the minimum or maximum terms.
- Filtering transcription types such as manual, crowdsourced, and OCR.
- Searching languages by location.
- Downloading chip keys and language dictionaries.

Robert E. MacLaury PhD

<http://colcat.calit2.uci.edu/tiki-index.php?page=Brief%20Biography%20and%20Selected%20Works>

This page contains a brief biography, selected bibliographies of Robert E. MacLaury and his research along with brief summaries of *The Robert E. MacLaury Color Categorization Archive*, *The Mesoamerican Color Survey*, and the *MacLaury's Multinational Color Survey*.

People

<http://colcat.calit2.uci.edu/tiki-index.php?page=People>

This page contains biographical information for the people listed in the following categories:

1. **Creators**

Members of this group are responsible for developing core features of the ColCat project such as developing the wiki and methods to transcribe surveys through OCR and crowdsourcing.

2. Other Contributors

In addition to creators, other contributors assist the creators in developing core features of the ColCat project.

3. Advisory Board

Comprised of researchers in psychology, cognitive science, linguistics, and computer science from universities around the world, the advisory board consults with the development of ColCat ranging from suggesting user-interface design to clarifying linguistic data notation during transcription.

Resources

<http://colcat.calit2.uci.edu/tiki-index.php?page=Resources>

1. Other Color Categorization Databases

Contains a link to the World Color Survey data archives hosted at Berkeley.

2. Background Reading

Contains some bibliographical information and links of books about the evolution of language, color terms, and color categorization research.

3. ColCat Research Articles & Publications

Contains a selection of bibliographical information and links of publications by ColCat project creators.

4. ColCat Conference Posters, Talks & Media Coverage

Contains a selection bibliographical information and links of posters, magazine articles, and talks mentioning the ColCat project.

Research Tools

<http://colcat.calit2.uci.edu/tiki-index.php?page=Research%20Tools>


1. Manual Transcription Tools contain the following:

Manual transcription tools content includes templates, guides and utilities that can be used by registered users to convert raw data image scan .pdf content into digitized datasets in a format consistent with that already on the archive. The tools include

templates for converting all forms of data, guides for coding conventions and suggestions for accurate transcription, and guides for creating a dictionary and synonym list of empirically observed color terms in a survey. Registered users who transcribe raw image scan data using these tools may submit their results to the ColCat team for review and proofing, and, depending on accuracy, the results may be posted on the site with transcriber credit given to the users who carried out the transcription.

- a. Manual transcription procedure and template
- b. Dictionary creation procedure and template
- c. Samples of completed Naming/Foci/Mapping/Demographic data and Color Term Dictionary

2. Crowdsourcing Transcription Tools:

ColCat's crowdsourcing tools aim to automate the transcription of raw data image scans from the MacLaury archive, or other similar datasets, and convert them into data addressable file types. For this purpose we have implemented custom web application designs that present individuals with online tasks that collect robust and distributed raw data transcriptions. These crowdsourcing methods, coupled with intelligent data aggregation procedures, have been empirically tested as approaches appropriate for accurate and unbiased MacLaury archive transcription efforts (see work listed on the **Resources** page [Link](#) ). Interested users who wish to further contribute to the ColCat archive transcription can participate in our crowdsourcing efforts directly through our web application, or through Amazon Mechanical Turk. Please visit the crowdsourcing website for a more in-depth overview of our crowdsourcing processes.

3. Data Exploration and Downloading Tools:

These tools allow users to download digitized archive data in bulk as comma-delimited (.csv) files for the purpose of data analysis. Users can choose to download the entire available data archive, or choose a subset of the languages in the archive by filtering languages based on criteria such as number of informants or number of color terms. We've made these filtering tools available for both the ColCat data (based on the Robert E. MacLaury Color Survey), and the previously published data from the World Color Survey (WCS) archives.

Data exploration and downloading will be seamless if users spend some time to refer to the documentation provided on the *Resource Tools* page prior to performing any data downloads. We thus recommend, before investigating any data on the site, that users study the information contained in the .pdf files shown below:

- **Guides for Data Tools**
 - [Data Filter & Download Tutorial \(.pdf\)](#)
 - [ColCat Digital Data File Format Guide \(.pdf\)](#)
 - [Downloading Data with User-defined Term Equivalence Classes \(.pdf\)](#)
- **Other Data Resources**
 - [ColCat-WCS Chip Map \(.csv file\)](#) 

Before downloading and using any data from the site it is important to note that failure to preserve Unicode character formatting in downloaded files will result in loss of important foreign-language character features in the downloaded data. For this reason the following advisory is seen in several locations on the download pages: *["Warning! The data you are about to download contains UTF-8 encoded Unicode characters. DO NOT open this file with Microsoft Excel because it does not support this file format."](#)* Please see additional user documentation contained in the "ColCat Digitized Data Format Guide" for further details.

Q&A Forum

<http://colcat.calit2.uci.edu/tiki-index.php?page=Test%20Forum>

The Forum feature of the ColCat Wiki aims to foster collaborative interactions among ColCat users, providing a place to give input and suggestions to Wiki developers, and for reporting Wiki corrections or fixes that are needed as the platform grows. Three discussion channels are provided on-board in the current Wiki public-release, however as this is a user-driven feature of the wiki, anyone can suggest new channels they want started, or changes to the organization and names of existing channels. The two main goals of the Forum is to (1) improve the ColCat site as a research resource, and (2) allow users the opportunity to shape the growth of the ColCat wiki, as well as make inquiries and share information with others for the sake of research development.

Users can start discussions in the following channels:

- 1. Suggestions and Feedback**

Users can post suggestions and feedback for developers and researchers to refine the wiki's structure, user-interface, or contents.

- 2. General Discussion**

Users can have discussions that do not belong to a currently existing forum.

- 3. Language Pages**

Users can post questions, errata, updates and suggestions on language pages.

This work is licensed to the Authors under Creative Commons Attribution-Noncommercial-NoDerivatives Works 4.0 International License. December 31, 2015.



Jameson, K. A., Gago, S., Deshpande, P.S., Benjamin, N.A., Chang, S.M., Tauber, S., Jiao, Y., Harris, I.G., Xiang, Z. Huynh, B.B., Ke, H., Lee, W.J., MacLaury, R.E. (2016). "The Robert E. MacLaury Color Categorization (ColCat) Digital Archive." [http:// colcat.calit2.uci.edu/](http://colcat.calit2.uci.edu/). The California Institute for Telecommunications and Information Technology (Calit2). University of California, Irvine.

Appendix 1 - Table inventory of number of informants per language survey.

Language	Informants	Language	Informants
Acatec	19	Northern Tepehuan	12
Aguacatec	10	Ocuiltec	2
Amuzgo	30	Otomi	21
Cakchiquel	35	Pame	1
Chatino	2	Papago	20
Chichimeca	1	Paya	7
Chinantec	39	Pima	5
Chontal	1	Pocomam	2
Chorti	5	Pocomchi	3
Chuj	12	Popoloca	2
Guarijio	25	Quiche	42
Huasteco	39	Rabinal Achi	5
Huave	9	Sacapultec	5
Huichol	5	Seri	25
Ixil	1	Tarahumara (Central)	9
Jacalteco	10	Tarahumara (Western)	7
Jicaque	10	Tepecano	1
K'ekchi	15	Tepehua	8
Kanjobal	6	Tlapanec	25
Lacandon	6	Tojolabal	3

Mam	26	Totonac	12
Matlatzinca	3	Trique	14
Mazahua	25	Tzeltal	12
Mazatec	26	Tzotzil	3
Mixe	1	Uspantec	11
Mixe of Coatlan	6	Yucatec Maya	1
Mixtec	69	Zapotec	62
Mopan Maya	10	Zoque	1
Nahuatl	28		

Non-MCS Language-Informant Table

Language	Informants	Language	Informants
ASL(English)	1	Korean	44
Afrikaans	1	Kwak'wala	1
American English	38	Lani	10
Apache	32	Lillooet	12
Baluchi	5	Lushootseed	1
Bangala	14	Makah	1
Belarusian	2	Mexican Spanish	1
Canadian English	6	Mikasuki	25
Cantonese	1	Nanja	1
Castellano	1	Ndebele	2
Cherokee	1	Nooksack	1
Chilcotin	5	Northern Sotho	2
Cocupa	2	Panamint	1

Cree	1	Pazande & Bangala	10
Creek	1	Polish	1
English	24	Portuguese	1
Espanol	1	Quechua	16
Finnish	2	Russian	7
French	3	Samish	1
German	1	Sechelt	6
Greek	1	Shanghainese	1
Guarao	2	Shuswap	9
Haisla	0	Sindhi	1
Halkomelem	5	Tectiteco	7
Hindi	1	Tsutu	1
Hochdeutsch	3	Tswana	1
Hopi	1	Tzutujil	1
Hungarian	9	Ukrainian	2
Hupa	3	Venda	1
Icelandic	1	Xhosa	1
Igbo	1	Yakima	1
Italian	1	Yao	1
Japanese	8	Yaqui	1
Karuk	7	Yurok	1
Ki-Bila	3	Zapotec	12
Ki-Lese	3	Zulu	55