Restructuring access: The creation of an online California Language Archive

Amy Campbell, Andrew Garrett, Hannah Haynie, Justin Spence, Ronald Sprouse, and John Sylak (University of California, Berkeley)

ICLDC 2, Honolulu
12 February 2011
Outline

Language archives at the University of California, Berkeley
  Survey of California and Other Indian Languages
  Berkeley Language Center
  Hearst Museum & Bancroft Library

Assembling the California Language Archive
  Goals
  Documents
  Sound recordings
  Data migration

The California Language Archive: Structure and use
  Database structure and use
  Geographical metadata and interface
Language archives at Berkeley

Four campus archives hold documentation of indigenous languages of the Americas, with a special focus on California:

- Survey of California and Other Indian Languages: paper documents deposited from 1950 to today
- Berkeley Language Center: sound recordings deposited from 1950 to today
- Phoebe A. Hearst Museum of Anthropology: sound recordings deposited between 1901 and 1950
- Bancroft Library: paper documents created from about 1800

California is linguistically the densest area of the Americas, with 80–90 indigenous languages belonging to 20 language families . . .
Survey of California and Other Indian Languages

An archive of paper documents deposited since about 1950:

- c. 270 linear feet of material from 296 languages
- Contents: field notebooks, unpublished manuscripts, lexical files, letters and notes, microfilms, etc.
- Organization: 186 collections, > 2,450 distinct catalog items
- Digitization status: currently c. 58,000 scanned images from 303 items in 48 collections
- Online catalog: [http://linguistics.berkeley.edu/survey/catalog/](http://linguistics.berkeley.edu/survey/catalog/)
Representative collection:
Abraham M. Halpern (1914–1985)

- Halpern (Ph.D. Chicago 1947) studied with A. L. Kroeber and Harry Hoijer, then had a career in international relations
- collection has 80 items, mainly concerning Pomoan languages
- includes 23 field notebooks dated 1939–1940 and 1983
- also lexical file slips, linguistic notes, unpublished manuscripts, correspondence, etc.
Berkeley Language Center

An archive of sound recordings deposited since about 1950:

- 1,730 hours of recordings from 136 languages (cataloged to date)
- Media: wire recordings, tape reels, cassette tapes, DATs, CDs, videotapes, DVDs, digital files
- Organization: 221 collections, 7842 audio segments
- Digitization status: all cataloged material is digitized (but there is a large backlog)
- Online catalog: http://www.mip.berkeley.edu/blc/la/
Representative collection: Sylvia M. Broadbent, Southern Sierra Miwok

- Broadbent (Ph.D. Berkeley 1960) recorded three languages between 1955 and 1957, including SSM

- Contents
  - 20 songs
  - 18 elicitation sessions
  - 14 narratives
  - 3 conversations

- Duration: 6 hours, 22 minutes

- Related material: field notebooks in the Survey of California and Other Indian Languages

Collection includes sound recordings created beginning in 1902, mainly before 1940:

- c. 400 hours of recordings from c. 65 languages
- Organization:
  - Wax cylinders: 2713 cylinders = c. 2200 texts in 113 collections
  - Wire recordings and tapes: c. 160 recordings
- Digitization status: all content is digitized (no public access)
- No online catalog of language content
Representative collection: E. W. Gifford’s 1926 Mono recordings

- Gifford (1887-1959) recorded three Mono speakers in August 1926
- Contents (on 59 wax cylinders)
  - 7 songs
  - 5 narratives
  - 1 ‘explanation’ (of a dance)
- Duration: 1 hour, 22 minutes
Bancroft Library

Representative collection:
T. T. Waterman’s field notebooks

- TTW (Ph.D. Columbia 1913)
  - at Berkeley 1907–1918, 1920–1921
  - d. 1936 while Territorial Archivist of Hawai’i

- Linguistic fieldwork
  - 6 Yahi notebooks
  - 16 Yurok notebooks
  - also Kumeyaay, Luiseño, Makah, Tubatulabal

- Related material: wax cylinder recordings in the Hearst Museum

Waterman’s 1914 transcript of Ishi’s wood duck story, p. 1 (Yahi language; Bancroft Library CU-23.1 #35.1)
## Language archives at Berkeley

<table>
<thead>
<tr>
<th>Date</th>
<th>Content</th>
<th>Repository</th>
<th>Extent</th>
<th>Digitized</th>
</tr>
</thead>
<tbody>
<tr>
<td>post-1950</td>
<td>sound</td>
<td>BLC</td>
<td>1,730 hours</td>
<td>all</td>
</tr>
<tr>
<td></td>
<td>documents</td>
<td>SCOIL</td>
<td>2,451 items</td>
<td>in progress</td>
</tr>
<tr>
<td>pre-1950</td>
<td>sound</td>
<td>Hearst</td>
<td>c. 400 hours</td>
<td>all</td>
</tr>
<tr>
<td></td>
<td>documents</td>
<td>Bancroft</td>
<td>extensive!</td>
<td>none</td>
</tr>
</tbody>
</table>

The BLC, SCOIL, and Bancroft collections are open and continue to expand; the Hearst audio collection is closed.

Currently in the CLA: BLC and SCOIL collections. Next (this spring): Hearst sound recordings
California Language Archive objectives

An integrated online portal:

- easy access
- convenient digital content
  - scanned documents
  - sound recordings
- consistent content description
- consistent access principles (including restricted access cases)

As a regional archive with dense coverage of a language area, we have a duty to create intuitive access for indigenous communities (speakers, learners, heritage users) as well as academic researchers.
Survey of California and Other Indian Languages

Database history

- 1970s: Paper catalog of archive contents
- 1980s: Electronic catalog in Panorama, migrated to Filemaker and reorganized c. 2001
  - Better organized
  - Digital content
  - Still inaccessible
- 2007: MySQL
  - Accessible
  - Backed up regularly
- Access to digital content (scanned images) since 2009
Survey of California and Other Indian Languages

Search parameters

- Search by contributor or language
- Filter: Images only

http://linguistics.berkeley.edu/survey/catalog/
Survey of California and Other Indian Languages

Search result:
Southeastern Pomo
Berkeley Language Center

Database history

- **1970s:** Legacy media remastered to reels

![Contents Sheet](contents-sheet.png)

- **1990s:** Permissions records updated; metadata → database
- **Early 2000s:** Collection digitized; catalog interface online
Berkeley Language Center

Search parameters

- Search by contributor or language
- Or by collection title, identifier, language family, content type, or keyword

http://www.mip.berkeley.edu/blc/la/
Language archives at Berkeley

Assembling the CLA

CLA: Structure and use

Berkeley Language Center

Search result:

Pomoan languages

2. Central Pomo / collected by Abraham Halpem; spoken by Esther Ward (1984)
3. Eastern Pomo / collected by Sally McLendon
4. Kashaya Pomo / collected by Robert Oswalt; spoken by David Antone et al. (1957)
5. Northeastern Pomo / collected by Sally McLendon (1958)
7. Northern Pomo / collected by Mary Catherine O'Connor; Spoken by Edna Guerrero and Annie Lake
8. Southeastern Pomo / collected by Catherine A. Callaghan; spoken by John Knight (Lake Miwok speaker) (1956)
10. Southern Pomo / collected by Roy Sniard; spoken by Maggie Woho, Laura Fish Somersal (1966)
Data migration

Very different data models: SCOIL documents (vs. BLC audio)
Data migration

Very different data models: (SCOIL documents vs.) BLC audio
Data migration

Data migration issues

- Merging two data models
  - Understanding both models
  - Mapping content from one model onto another
  - Example: Audio segment duration

- Handling new types of content
  - Creating new tables (audio segments, places)
  - Restructuring existing tables
  - Example: Contributions
Data model: Content

Collections

- **Collections** are assembled and deposited by individuals: usually academic researchers, but also Native community members
- Many Collections have 1-5 Items; others have hundreds

Items

- **Items** are describable objects in Collections
- Examples: field notebook; box of tape reels; box of file slips; draft dictionary; unpublished grammar

Digital content (associated with Items)
Data model: Content

Digital content (associated with Items)

- Images
  - Scanned document images: notebooks, file slips, ...
  - A field notebook might be associated with 20 images, a file slip box with 1,500

- Audio segments
  - Legacy sound recordings were divided into shorter meaningful segments
  - New material respects depositor’s organization
  - Some Items associated with 5 segments, others with 200
Data model: Ontologies

Languages

<table>
<thead>
<tr>
<th>id</th>
<th>name</th>
<th>parent_id</th>
<th>alternative_name_of</th>
<th>ISO 639-3</th>
</tr>
</thead>
<tbody>
<tr>
<td>238</td>
<td>Pomo</td>
<td>234</td>
<td>NULL</td>
<td>NULL</td>
</tr>
<tr>
<td>269</td>
<td>Southeastern Pomo</td>
<td>238</td>
<td>NULL</td>
<td>pom</td>
</tr>
<tr>
<td>270</td>
<td>Lower Lake Pomo</td>
<td>NULL</td>
<td>269</td>
<td></td>
</tr>
</tbody>
</table>

People

<table>
<thead>
<tr>
<th>id</th>
<th>lastname</th>
<th>firstname</th>
<th>alternative_name_of</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>130</td>
<td>O’Connor</td>
<td>Mary Catherine</td>
<td>NULL</td>
<td></td>
</tr>
<tr>
<td>10245</td>
<td>O’Connor</td>
<td>Catherine</td>
<td>130</td>
<td></td>
</tr>
<tr>
<td>10246</td>
<td>O’Connor</td>
<td>Cathy</td>
<td>130</td>
<td></td>
</tr>
<tr>
<td>10247</td>
<td>Connor</td>
<td>Mary Catherine</td>
<td>130</td>
<td></td>
</tr>
</tbody>
</table>
Web interface: Search → Collections → Items

Search metadata fields with one input (multiple fields searched)

CALIFORNIA LANGUAGE ARCHIVE
A catalog and online archive of the indigenous languages of California, western North America, and the Americas

Explore collections

Search collections

- All collections  - Digital only  - Audio only

Collections included:
Berkeley Language Center
Survey of California and Other Indian Languages

© 2010 The Regents of the University of California. The California Language Archive is a joint initiative of the Berkeley Language Center and the Survey of California and Other Indian Languages.
Search result: Collections for **Kashaya**, digital content only
Web interface: Search → Collections → Items

Search result:
Items for Kashaya

1. Item identifier: Barrett.014 [description]
   - Title: [Pomo and Yuki notes and vocabulary]
   - Contributor: S. A. Barrett (researcher)
   - Languages: Northern Pomo, Kashaya, Yuki
   - Extent: 1 folder
   - Date: 1948
   - Description: Original notes in pencil and ink; carbon copy of a text (story about Coyote). Includes phonetic inventory and vocabulary.
   - Collection: Samuel Alfred Barrett Papers
   - Repository: Survey of California and Other Indian Languages
   - Access conditions: unrestricted

2. Item identifier: Haas.083 [description [51 images]]
   - Title: Readings in American Indian linguistics
   - Contributor: Mary Haas (researcher)
   - Languages: Blackfoot, Chiricahua Apache, Creek, Eastern Ojibwa, Eastern Pomo, Hill Patwin, Hopi, Karuk, Kashaya, Koasati, Konkow, Nahuatl, Navajo, Nez Perce, Quileute, Shasta, Takelma, Tunic, Wintu, Wyot, Nuu-chah-nulth, Tubatulabal, Yavwiman
   - Extent: 5 folders
   - Date: [undated]
   - Description: Mimeographs of parts 2-4 of "Readings in American Indian Linguistics", dated 1963, accompanied by three extra copies of part 3. Also contains mimeographs of thumbnail sketches by Haas and her students on miscellaneous California languages.
   - Collection: Repository: Access conditions: unrestricted

3. Item identifier: Halpern.042.001 [description]
   - Title: [Kashaya and Central Pomo texts]
   - Contributors: Mary James (consultant), Abraham M. Halpern (researcher)
   - Languages: Kashaya, Central Pomo
   - Extent: 1 folder
   - Date: [undated]
   - Description: Texts, typed with handwritten notes.
   - Collection: Abraham M. Halpern Papers on Pomoan Languages
   - Repository: Survey of California and Other Indian Languages
   - Access conditions: unrestricted

CLA: Structure and use

CALIFORNIA LANGUAGE ARCHIVE
A catalog and online archive of the indigenous languages of California, western North America, and the Americas

Search within your results:

Your search: Kashaya

ISO 639-3: kju

Alternative names: Southwestern Pomo

- Hokin
- Chimariko
- Chumashan
- Esselen
- Karuk
- Palaihnihan
- Pomo
  - Central Pomo
  - Eastern Pomo
  - Kashaya
    - Northeastern Pomo
    - Northern Pomo
    - Southeastern Pomo
    - Southern Pomo
- Salinan
- Seri
- Shastan
- Washo
- Yana
- Yuman-Cochimi

California Language Archive

Campbell, Garrett, Haynie, Spence, Sprouse, Sylak
Collection details

CALIFORNIA LANGUAGE ARCHIVE
A catalog and online archive of the Indigenous languages of California, western North America, and the Americas

Search result:
A collection

Abraham M. Halpern Papers on Pomoan Languages

Collection identifier: SCL Haplpm

Additional contributors: John Kelsey (consultant), Nancy McCoy (consultant), M.E.S. (consultant), Jenny Piko (consultant), Elsie Allen (consultant), Mary James (consultant), Mack Williams (consultant), Lowe Anderson (consultant), Robert L. Oswald (compiler), A.T. (consultant), Abraham M. Halpern (researcher), G.A. (consultant), Thomas Leon (consultant), Santiago McDaniel (consultant), Katherine Halpern (researcher), Clifford Salvador (consultant), Daley Lowell Lorenzo (consultant), Esther Ward (consultant), Edna Campbell (consultant), Julius Meshinsky (researcher), George Patch (consultant), Gretchen S. Hillard (depositor), Julia Mamufo (consultant), Annie Burke (consultant), Grant Jako (consultant), E.S. (consultant), Joe Augustine (consultant), Steve Parrish (consultant)


Extent: 11.91 linear feet (24 boxes, 2 binders, and 1 envelope)

Scope and content: Most of the material in this collection documents Halpern's ethnographic and linguistic field research on Pomoan languages; also included are ethnographic and linguistic materials related to his work on Patwin and Quechan (Yuma). The Pomoan material includes original notebooks from field trips conducted in 1936, 1939-1940, and the early 1980s, plus derived materials such as vocabulary files, slips, manuscript articles, text transcriptions, and genealogies. Halpern's language consultants included the following people: Steve Parrish, Jenny Piko, and Esther Ward (Central Pomo); Joe Augustine (Eastern Pomo); Mary James and Julia Mamufo (Kashaya); Santiago McDaniel (Northern Pomo); Lowe Anderson, Edna Campbell, Nancy McCoy, and Mack Williams (Northern Pomo); John Kelsey, Effie Kelsey, Thomas Leon, George Patch, and Clifford Salvador (Southeastern Pomo); Elsie Allen and Annie Burke (Southern Pomo); Daley Lowell Lorenzo (Hill Patwin).

Repository: Survey of California and Other Indian Languages

Associated materials: Audio recordings associated with the Papers are in the Berkeley Language Center, Berkeley, California (LA 285, LA 283, LA 284, LA 250). Other collections of materials related to Halpern's Pomoan, Yuman, and Wintuan research are held by the Bancroft Library (Berkeley, California) and the American Philosophical Society (Philadelphia, Pennsylvania).

Access conditions: unrestricted

© 2010 The Regents of the University of California. The California Language Archive is a joint initiative of the Berkeley Language Center and the Survey of California and Other Indian Languages.
Image display

Search result: An image (A. M. Halpern, Pomo field notebook 1)
### Search result:
#### Audio segments

**CALIFORNIA LANGUAGE ARCHIVE**
A catalog and online archive of the indigenous languages of California, western North America, and the Americas

<table>
<thead>
<tr>
<th>Description</th>
<th>Audio (122)</th>
</tr>
</thead>
</table>

**Item identifier:** LA 98.1 [description | 122 audio segments]

- **Title:** Kashaya Pomo audio recordings
- **Contributors:** Elizabeth Sheridan (consultant), Nettie Manuel (consultant), Julia Picola Mamuyo (consultant), William Graves (consultant), Herman James (consultant), Louisa Smith (consultant), Violet Fern Chapel (consultant), Ethel Lucas (consultant), Sid Pamish (consultant), Jesse Frank (consultant), Gladys James (consultant), Johnny Smith (consultant), Isabelle Johnson (consultant), Robert Oswalt (researcher), Essie Pamish (researcher), consultant

- **Language:** Kashaya
- **Duration:** 33 hours, 34 minutes
- **Date:** 1957-1964
- **Description:** Linguistic field recordings; linguistic data; stories; ethnographic data; songs; Additional ethnographic or ethno-historical texts, conversations, reminiscences, untitled texts. Some in English.
- **Collection:** The Robert Osvald collection of Kashaya Pomo field recordings
- **Repository:** Berkeley Language Center
- **Access conditions:** local access only [at University of California, Berkeley]

1. **Title:** Prayer in English followed by a monologue in Kashaya
   - **Segment number:** 001_1
   - **Date recorded:** Jul. 3, 1958
   - **Duration:** 51 minutes
   - **Contributors:** Essie Pamish (consultant), Robert Oswalt (researcher)
   - **Download MP3:** 001_1

2. **Title:** Untitled monologue in Kashaya
   - **Segment number:** 002_1
   - **Date recorded:** Jul. 3, 1958
   - **Duration:** 59 minutes
   - **Contributors:** Robert Oswalt (researcher), Essie Pamish (consultant)
   - **Download MP3:** 002_1
   - **Play mp3**
Language ontology display

Numic
ISO 639-3: [No ISO code]
- Uto-Aztecan
  - Northern Uto-Aztecan
    - Hopi
  - Numic
    - Central Numic
      - Comanche
    - Panamint
    - Shoshone
  - Southern Numic
    - Chemehuevi-Ute-Southern Paiute
      - Kawaiisu
  - Western Numic
    - Mono
      - Northern Paiute
    - Takic
    - Tubatulabal
  - Southern Uto-Aztecan

- Uto-Aztecan
  - Northern Uto-Aztecan
    - Hopi
  - Numic
    - Central Numic
      - Comanche
    - Panamint
    - Shoshone
  - Southern Numic
    - Chemehuevi-Ute-Southern Paiute
      - Kawaiisu
  - Western Numic
    - Mono
      - Northern Paiute
    - Takic
      - Cupan
        - Cahuilla
        - Cupeño
        - Juaneño
        - Luiseño
      - Gabrielsono
      - Serran
      - Tataviam
    - Tubatulabal
  - Southern Uto-Aztecan
Places in the CLA

Biographical data about contributors [e.g. 1 Annie Lake, Northern Pomo consultant for Robert L. Oswalt]:

- Born at [2 El Roble]
- Lived at [3 Upper Coyote Valley]
- From
- Mother from
- Father from [4 Ukiah Rancheria]

At least 1 important place per person is selected for map display [above, Upper Coyote Valley].

Robert L. Oswalt, Northern Pomo field notebook (1967?), p. 1
Places in the CLA

Geographic information about documentation activities [e.g. Oswalt Kashaya field notebook (1989)]:

- Fieldwork location [1 Petaluma]

Item places are useful where biographic information about linguistic consultants is not recorded.

Conceptual database schema for geography

Place
- latitude
- longitude
- name
- ...

(Item) Sulphur Creek

ItemPlace
- has_many belongs_to

LifeFact
- date
- type, e.g., 'born at', 'lived at'

Item
- title
- date
- ...

Abraham M. Halpern
Pomo field notebook 14

Person
- firstname
- lastname
- ...

Annie Burke

California Language Archive

Campbell, Garrett, Haynie, Spence, Sprouse, Sylak
Maps in the CLA user interface

Maps allow users to interact with the archives in new ways:

▶ Entry into collections through a visual interface, rather than text searches
▶ Visualization of geographic coverage of holdings and search results
▶ Geographic navigation of archival materials
▶ Filtering of map display by audio vs. text materials, and by availability of online content
Maps in the CLA user interface

Navigation to find content

CALIFORNIA LANGUAGE ARCHIVE
A catalog and online archive of the indigenous languages of California, western North America, and the Americas

Archival Documents and Manuscripts
- Content Available Online
- No Online Content

Audio Recordings
- Content Available Online
- No Online Content

[Southern Pomo field notebook] (Oswalt.001.004)
Collection: Oswalt
Contributors: Robert L. Oswalt; Olive Fulwider; Oswalt family
Language: Southern Pomo
Extent: 1 notebook
Date: 1992-08-20
Description: Vocabulary (personal names), notes from listening to 1965 Elizabeth Dollar stories.
Associated Materials: Audio recordings associated with the Papers are in the Berkeley Language Center, Berkeley, California (LA 98).
Repository: Survey of California and Other Indian Languages

© 2010 The Regents of the University of California. The California Language Archive is a joint initiative of the Berkeley Language Center and the Survey of California and Other Indian Languages.
Maps in the CLA user interface

Location map for Hill Patwin Items

CALIFORNIA LANGUAGE ARCHIVE
A catalog and online archive of the indigenous languages of California, western North America, and the Americas

Collections (6)  Items (22)

21. Item identifier: LA 105.1 (description | 12 audio segments)
   Title: Hill Patwin audio recordings
   Contributors: Oscar McDermott (consultant), Kenneth Wayne Whistler (researcher)
   Language: Hill Patwin
   Duration: 2 hours, 55 minutes
   Date: 1975
   Description: Linguistic field recordings; songs; reminiscences.
   Collection: The Kenneth W. Whistler collection of Hill Patwin field recordings
   Repository: Berkeley Language Center
   Access conditions: unrestricted

22. Item identifier: LA 20.1 (description | 16 audio segments)
   Title: Hill Patwin audio recordings
   Contributors: Nora Lovell (consultant), Elizabeth Bright (researcher)
   Language: Hill Patwin
   Duration: 1 hour, 18 minutes
   Date: 1992
   Description: Linguistic field recordings; linguistic data; stories; songs. Some English glosses.
   Collection: The Elizabeth Bright collection of Hill Patwin field recordings
   Repository: Berkeley Language Center
   Access conditions: unrestricted

© 2010 The Regents of the University of California, The California Language Archive is a joint initiative of the Berkeley Language Center and the Survey of California and Other Indian Languages.
Acknowledgements

Thanks to:

- the National Science Foundation and National Endowment for the Humanities, whose grant PD-50005-07 funds our work, through the Documenting Endangered Languages program;
- Lisa Conathan and Leanne Hinton for archival and conceptual guidance; and
- generations of scholars and Native community members who entrust their careful work and cultural heritage to our archives.

Corresponding author: Andrew Garrett (garrett@berkeley.edu)

California Language Archive: http://cla.berkeley.edu/ (in March)