Manejo de Corpus 101:
Haciendo documentación de lenguas que es lista a archivar

Heidi Johnson
El Archivo de los Idiomas Indígenas de LatinoAmérica (AILLA)
Universidad de Texas en Austin

Definiciones

- **Archivo:** un repositorio fidedigno creado y mantenido por una institución con un **compromiso a permanencia demostrado** y un compromiso a la preservación sobre lo largo de los recursos archivados.
- **Corpus de documentación de lenguas:** la colección de materiales documental creada por investigadores y hablantes naturales.

Lo que se debe archivar - I

- Grabaciones, ambas audio y video:
  - eventos públicos: ceremonias, oratorias, bailables…
  - narativos: históricos, traditionales, mitos, …
  - instrucciones: como construir una casa, como tejer un petate, como pescar, ...
  - literatura: oral o escrito – cualquiera obra creativa
  - conversaciones: si no sean tan personales

Lo que se debe archivar - II

- Materiales secundarios (derivados):
  - transcripciones, traducciones, & anotaciones, comentarios
  - notas del campo, listas para elicitación, ortografías
  - juegos de datos, bases de datos, hojas de cálculo
  - esbozos, e.g. gramáticos, etnografías
- Fotografías
- Artículos que no son publicados por otros medios.

Lo que se debe archivar - III

- Materiales para enseñanza y aprendizaje:
  - lecciones elemental
  - calendarios, carteles, etc.
  - diccionarios ilustrados, enciclopedias
  - diseños de curriculum
  - cualquiera que otra gente encontrarían útil o de inspiración en sus propios programas.

Lo que NO se debe archivar

- Cualquiera cosa que puede hacer daño o dar vergüenza á los hablantes, e.g.:
  - Pamela Munro's entrevistas con Zapotecos en L.A. sobre cruzar la frontera sin documentos.
  - Chismes que no han maturado suficientemente (chismes antiguos son historia)
- Obras sagradas con usos muy restringidos.

rchivo de los Idiomas Indígena de

Cuando se debe archivar?

- Lo más pronto que sea posible:
  - a prevenir daño accidental damage o pérdida;
  - a recobrar formatos útiles de presentación.
- Se puede restringir acceso a obras en curso.
- Se puede añadir transcripciones, etc. más tarde.

Porqué se debe archivar? I

- a preservar grabaciones de lenguas en peligro/de minoridad para las generaciones que vienen.
- a facilitar el re-uso de materiales para:
  - programas para el mantenimiento y revitalización de lenguas;
  - estudios de typología, historia, etc;
  - cualquiera clase de estudio – lingüístico, antropológico, psicológico – que Ud. no hace.

Porqué se debe archivar? II

- a fomentar el desarollamiento de literaturas, ambas orales y escritas, para lenguas en peligro.
- a hacer saber que documentación existe para cuales lenguajes.
- a aumenter su CV aún antes de que esté listo a publicar resultos.

Archivar puede ser una forma de publicar

- Aunque sean restringidos los recursos, los metadatos son publicados.
- Liste Recursos Archivados en su CV para ganar reconocimiento para su trabajo.
- Reconozca á los creadores del obra:

Sánchez Morales, Germán. (1994). "Satornino y los soldados." [audio] Heidi Johnson, (Researcher.) [online] ZOH001R010. Access=public. http://www.ailla.utexas.org: Archive of the Indigenous Languages of Latin America.

Como construir un corpus listo para archivar - I

- Regla #1: Marque cada cosita que produce con una COHERENCIA DESPIEDADA. Si no se que es, no lo puedo archivar.
- Regla #2: Contacte sus archivistas amistosos y pidanos que le ayudamos.
- Regla #3: Pruebe su sistema antes de salir: aparatos, modo de catalogar, etiquetas.

Como construir un corpus listo para archivar - II

- Define un sistema al respecto a los derechos y desarolle una práctica consistente para obtener el consentimiento, e.g., formularios y/o declaraciones grabadas.
- Siempre pida permiso para todo:
  - grabación
  - archivación
  - extracción, publicación, etc.
- Aprenda como hablar con sus consultantes sobre derechos: visite a la Escuela de Prácticas Mejoras en

http://emeld.org/school/classroom/ethics/index.html

rchivo de los Idiomas Indígena de

Etiquetar I : grabaciones

- Audio - grabe una "cabecera" con informaciones básicas, en una lengua de contacto – inglés, español, portugués…

"Hoy, el 28 de octubre 2005, estamos en San Miguel Chimalapa, Oaxaca, en la casa del Sr. Germán Sánchez Morales. Sr. Sánchez nos va a contar la historia de Juan Flojo en zoque."

- Video – hagalo en el estilo Hollywood: use una tabla con la info escrita.

Etiquetar II: media y archivos

- Decida el tema fundamental para organizar su sistema de etiquetas:
  - media, e.g. CDs, cuadernos
  - nombres o iniciales de consultantes
  - lenguas/dialectos
  - nombres o iniciales de lingüístas
  - géneros, e.g. lexicas, narativos, …

Etiquetar III: ítemes relacionados

Materiales de la documentación de lenguas tipicamente viene en *juegos* relacionados:

- grabación de una historia + texto interlinear + traducción repasada + comentario
- entrevista + fotografías
- sesión de elicitación grabada + notas del campo
- grabaciones simultanéas de audio y video

Etiquetar IV: clases de relaciones

- derivación: una transcripción se deriva de una grabación
- serie: una grabación larga que ocupa varias media (cds solamente guardan 650 mb =~ 60 mins)
- parte-entero: grabaciones en video y audio hechas simultaneamente del mismo evento
- asociación: (borrosa) fotografías, comentarios

Etiquetar V: AILLA IDs

- ZOH001R040I001.mp3
  - ZOH = código del idioma
  - 001 = número del depósito (el primero)
  - R040 = recurso (juego) 40 en ese depósito
  - I001 = primer ítem (archivo) en ese recurso
  - .mp3 = que clase de archivo
- Apoya nuestra sistema administrativa: archivamos muchas lenguas, hacemos un depósito a la vez…

Etiquetar VI: objeto de media es fundamental

Facilita localización de cosas en el campo. Extensiones de los archivos identifican la clase de cada ítem.

- cd1t1.wav - cd 1, pista 1
- cd1t1.db - base de datos de shoebox (interlinear)
- cd1t1.doc - doc de word c/notas sobre cd1t1
- ds19.xls - juego de datos (raices de verbos)
- ds5.db - juego de datos en shoebox (deícticos)
- nb1 - cuaderno de campo (objeto de papél)

Catálogo del corpus / Metadata I

- Información catálogo para recursos digitales se llama *metadata*.
- Metadata apoya:
  - guardando ítemes relacionados juntos
  - protección de materiales sensibles
  - buscando para la cosa que quiere
  - el uso de recursos por mucha genta
  - citación apropriada de recursos archivados

Metadata II : Info mínima

- Nombres completos de todos los creadores: Ud. y los hablantes.
- Lenguaje: sea específico.
- Fecha de creación: YYYY-MM-DD.
- Lugar de creación: sea específico.
- Restricciones del acceso, instrucciones particulares sobre usos futuros.
- Palabra clave del género, e.g. narativo.

Metadata III : Info adicional

- Del proyecto: nombre, director, fundador, etc.
- Papeles de las participantes (e.g. hablador, investigador), edad, sexo, info de contacto
- De los media: aparatos, formatos, calidad…
- Del contenidos: descripciones del contexto de grabación, del subjecto – lo mas detalle, lo mejor.
- Citaciones: publicaciones pertinentes

Metadata IV

Dos esquemas (interoperables) son recomendados. Escoje uno para su base y extiéndalo a su gusto.

- OLAC – Open Language Archives Community – http://www.language-archives.org
- IMDI – International Standards for Language Engineering Metadata Initiative – http://www.mpi.nl/IMDI

Herramientos para el manejo de corpus

IMDI Browser & IMDI Data entry
(http://www.mpi.nl/IMDI)

AILLA's Shoebox 2.0 & 5.0 templates
(http://www.ailla.utexas.org/site/download_md_forms_sp.html)

- Cualquier base de datos, hoja de cálculo, templato de Word doc...
- Una carpeta de hojas sueltas con un formulario copiado.

Sitios útiles

- AILLA: http://www.ailla.utexas.org/
- DELAMAN: http://www.delaman.org/
- IMDI: http://www.mpi.nl/ISLE
- OLAC: http://www.language_archives.org
- EMELD: http://emeld.org
- Escribame: ailla@ailla.utexas.org

rchivo de los Idiomas Indígena de