Manejo de Corpus 101

Heidi Johnson

El Archivo de los Idiomas Indígenas de LatinoAmérica (AILLA)

Universidad de Texas en Austin

Definiciones

- **Archivo de preservación:** 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.
- **Archivo local o centro de lenguas:** un repositorio donde crean o admiten recursos y donde proveen acceso a recursos a una comunidad particular. No se preocupan de preservación sobre lo largo.
- **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, discursos…
  - narrativas: históricas, traditionales, mitos…
  - sesiones de elicitación
  - instrucciones: como construir una casa, como tejer un petate, como pescar, ...
  - literatura: oral o escrita – obras creativas
  - conversaciones: si no son tan personales

Lo que se debe archivar - II

- Textos y más (manuscrito o digital):
  - 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
- Tésis, artículos no-publicados, memorias…

Lo que se debe archivar - III

- Materiales para enseñanza y aprendizaje:
  - lecciones elemental
  - calendarios, carteles, etc.
  - diccionarios ilustrados, enciclopedias
  - diseños de curriculum
  - cualquiera cosa 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.
  - pero estas pueden ser archivadas con una fecha límite de 50 – 100 años.
- Obras sagradas con usos muy restringidos.
  - estas pueden ser archivadas si el acceso es controlado por miembros de la comunidad que tienen autoridad apropriada.

Cuando se debe archivar?

- Lo más pronto que sea posible:
  - a prevenir daño accidental 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

- preservar grabaciones de lenguas en peligro/de minoridad para las generaciones que vienen.
- 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

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

Archivar puede ser una forma de publicar

- Aunque los recursos sean restringidos, 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 sabemos que es, no lo podemos 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.
- Habla de quien puede / no puede leer/oir/ver las obras.
- Aprenda como hablar con sus consultantes sobre derechos: visite a la Escuela de Prácticas Mejoras en

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

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 de 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. O parate enfrente de la camera y platica la cabecera.
- Metadatos grabados así no se pierden nunca.

Etiquetar II: media y archivos

- Decida el tema fundamental para organizar su sistema de etiquetas:
  - media, e.g. CD1, cuaderno A
  - nombres o iniciales de consultantes
  - lenguas/dialectos (use el código de 3 letras)
  - 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: nombres enormes

Facilita relacionar archivos digitales en un disco duro. (Cortesía de Tony Woodbury.)

- SJQ-2007_09_11-Txt_ec-acw-1.wav
- SJQ-2007_09_11-Txt_ec-acw-2.wav
- SJQ-2007_09_11-Txt_ec-acw.eaf
- SJQ-2007_09_11-Txt_ec-acw.jpg

Nombres enormes II

- SJQ: Chatino de San Juan Quiahije (lugar y lengua en un golpe)
- 2007_09_11: fecha de grabación en formato ISO AAAA-MM-DD
- Txt: clase de discurso en este proyecto.
- ec: narrador/consultante
- acw: investigador
- 1: número en una serie
- .wav: tipo de archivo

Nombres enormes III

Ejemplo de Jonathan Amith:

digital_3:12:704_SP_Oa_testimonial-vida-camote-merchant.wav

Digital_3: sobre en el disco duro

12: número del disco compacto

704: julio 2004 [supongo: YYYY-MM-DD mejor]

SP: Silvestre Pantaleón (narrador)

Oa: Oapan (Guerrero: lugar y dialecto)

lo demás es un título breve

Nombres enormes IV

- Ventajas:
  - incluyen los metadatos esenciales
  - facilitan ordinación, identificación
  - fácil mantener cosas relacionadas juntas, por medio del extensión del archivo (.wav, .eaf)
- Desventajas:
  - largo para teclar! (pero cortas y pegas)
  - más difícil relacionar cosas no-digitales, como cuadernos (pero no tanto)

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 otra gente
  - citación apropriada de recursos archivados

Metadata II : Info mínima

- Nombres completos de todos los creadores: Ud. y los hablantes.
- Lenguaje: sea específico y/o use el código.
- Fecha de creación: AAAA-MM-DD.
- Lugar de creación: sea específico.
- Restricciones del acceso, instrucciones particulares sobre usos futuros.
- Palabra clave del género, e.g. narrativa.

Metadata III : Info adicional

- Depositante: info de contacto
- Proyecto: nombre, director, fundador, etc.
- Papeles de las participantes (e.g. hablador, investigador), edad, sexo, papel comunitario
- Los media: aparatos, formatos, fuentes
- Contenidos: descripciones del contexto de grabación, del subjecto – lo mas detalle, lo mejor.
- Citaciones: publicaciones pertinentes

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/school/index.html
- Escribame: ailla@ailla.utexas.org