Archiving challenges

Jacquelijn Ringersma
Max Planck Institute for Psycholinguistics

March 2010

Archiving challenges: content

What is a digital archive?

Parties involved in digital archiving

Archiving challenges

- organization of data
- coherence and persistency
- access and safety

Language archiving software

Different users, different needs

What is a digital archive?

What is a digital archive

Your stuff is buried here and gone forever

Some history in documentary linguistics

Primary data collection: recordings, photo, notes
Traditional way of knowledge sharing: through books and publications

On the shelf

Preserved

What is a digital archive

The digital ERA:

It is ‘easy’ to create and distribute copies of digital material

It is ‘easy’ to give access to this material

The original data carrier has lost importance

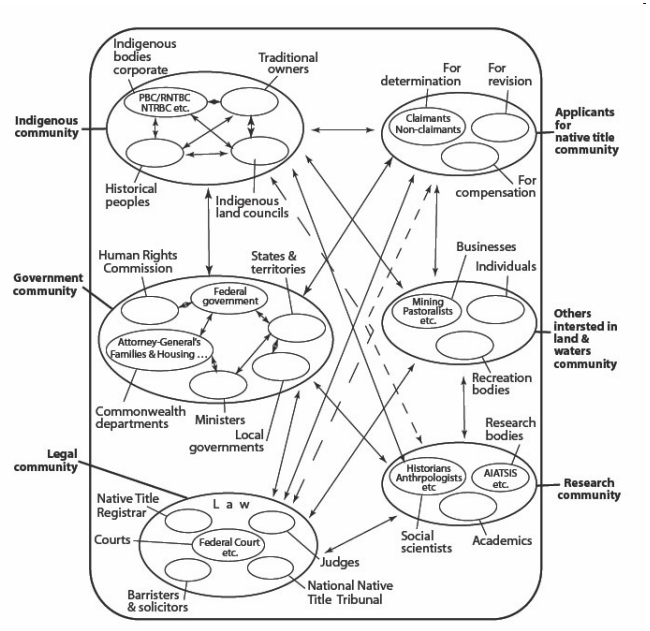
What is a digital archive?

a trusted repository created and maintained by an ***institution*** with a ***demonstrated commitment to permanence*** and the long-term preservation of archived resources.

New requirements?

Objects in the archive are subject to change, extension
Users may want to add information about the objects in the archive
Objects need to be accessed, but also save

Long term preservation – short time access



Parties involved in digital archiving

Parties involved

Depositor

- Speakers, linguists, anthropologists
- Anyone who wants the language documentation materials that they produce to survive and remain useful for generations to come.

In other words:

Parties involved

Users?

- Speakers, linguists, anthropologists
- Anyone who wants to use the material

Teachers, journalists, general public

Heterogeneous group

Heterogeneous knowledge on technical infrastructures

Parties involved

Archiving instance:

Example: Max Planck Institute for Psycholinguistics

Archive managers: 3

Archive developers: 2

System manager: 1

Archiving software development: 4

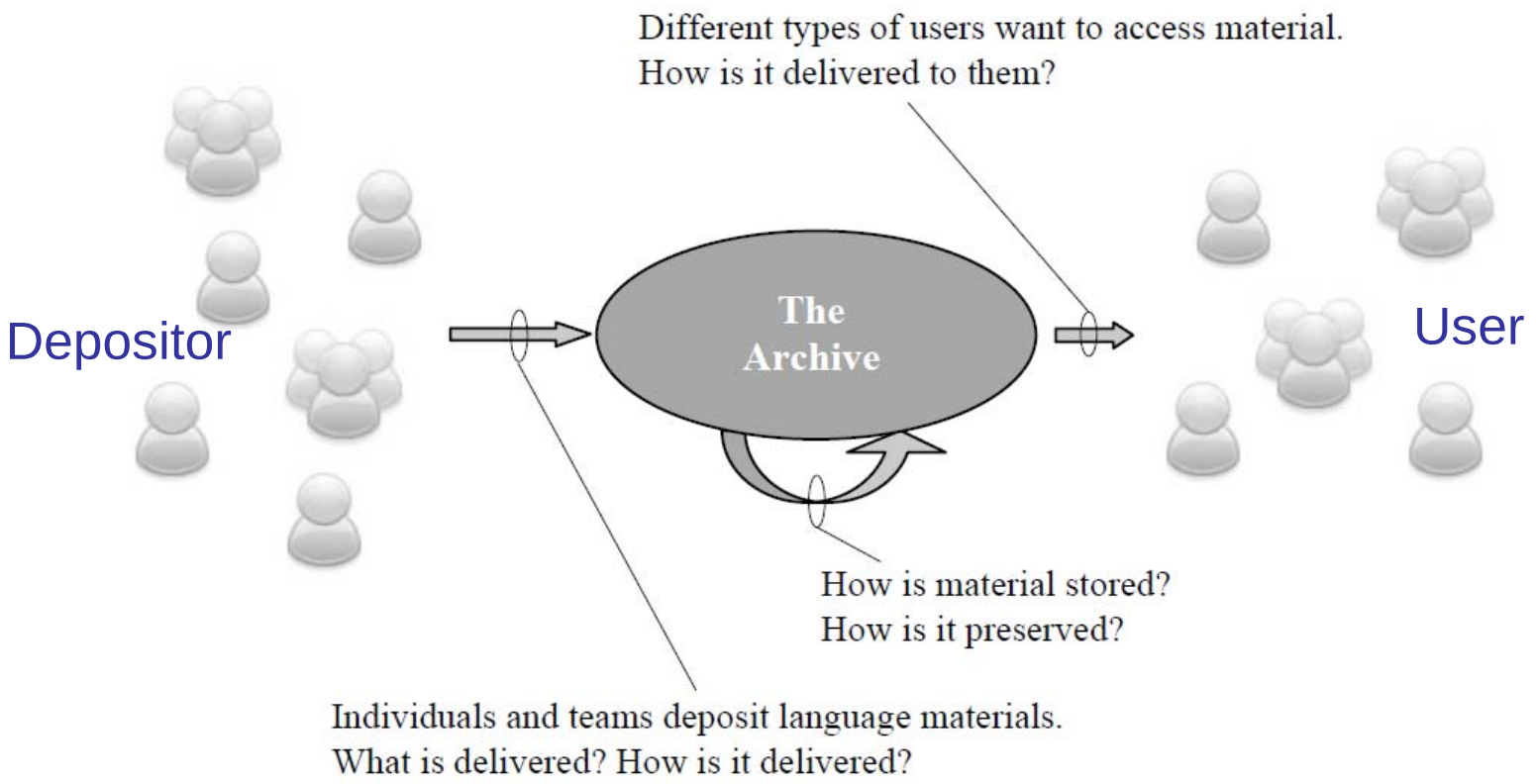
Enrichment software development: 4

Archive for language data:

40 Terabyte of data

400.000 archived objects

Parties involved



Archiving challenges

Archiving challenges

(1) Organizational aspects
Organize data following clear principles

(2) Long term persistency requirement
Create a coherent and consistent archive
Adhere to open standards
Store data in an accessible and persistent form

(3) Security issues
Give access to data to different users
But protect data against unauthorized access

Archiving challenges

(1) Organizational aspects

Resources need to be described (metadata)

Resources need to be structured (corpus structure)

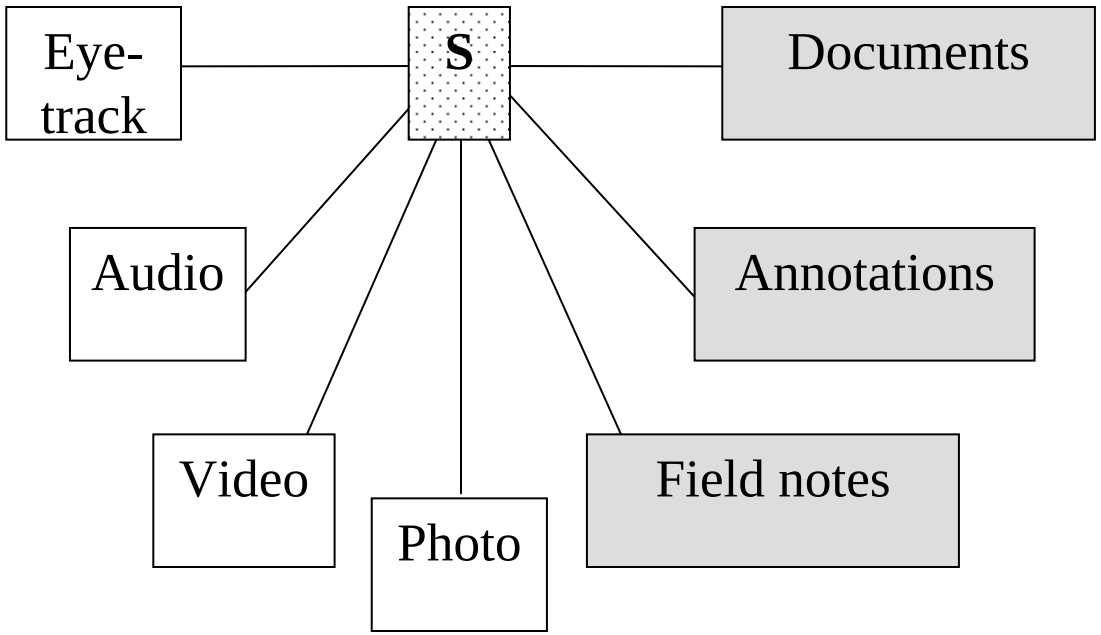
Archiving challenges

Language resources that make up corpora:

- (Digital) video or audio recordings, photographs
- Digitisations of images used as stimuli
- Transcription files
- One or more analysis files
- Field notes and experiment descriptions

Special to Language Resources

- In the linguistic domain often *clustered* resources
- Clustered because they refer to or result from the *same linguistic event/performance*.
- In our MPI archive terminology: **session resource bundle**



IMDI metadata set (1)

Aim: describe a session or resource bundle

- in a *structured* way
- with a sufficiently rich metadata set
- using domain specific names

IMDI = ISLE Metadata Initiative

ISLE = International Standards of Language Engineering

IMDI metadata set (2)

Categories

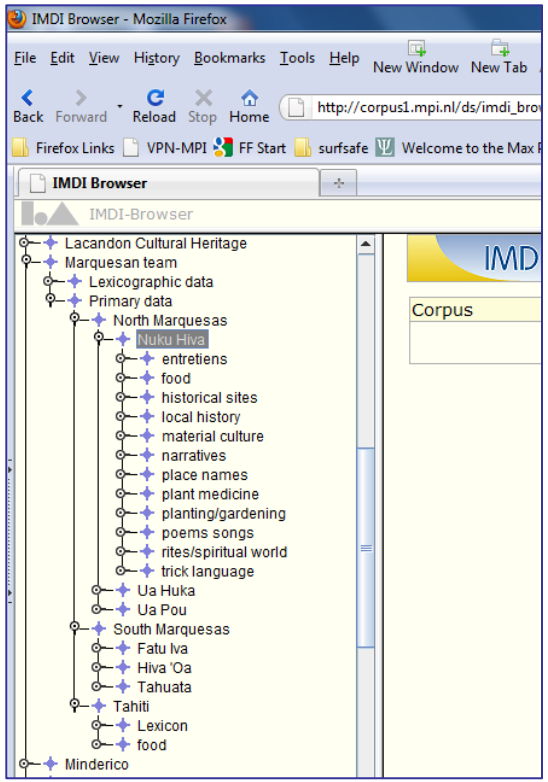
Administrative – (date, tool, version …)
General – (project, location …)
Content – (language, genre, modality …)
Actors/Participants – (biographic/contact information)
Resources – (URL, type, format, accessibility …)

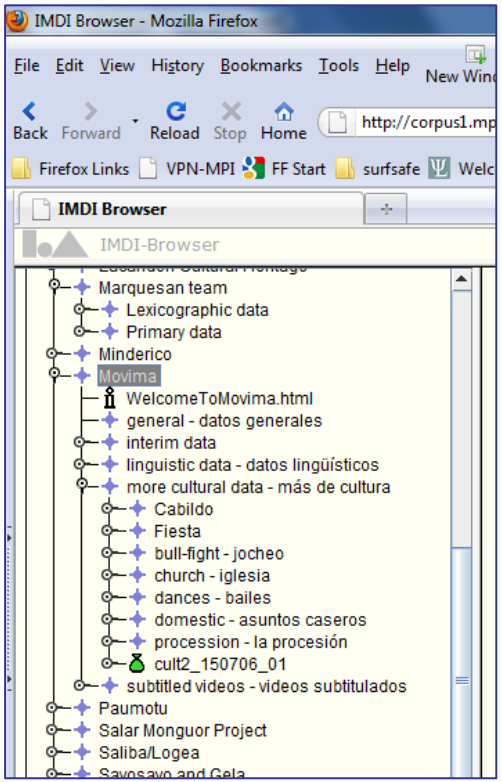
Archiving challenges

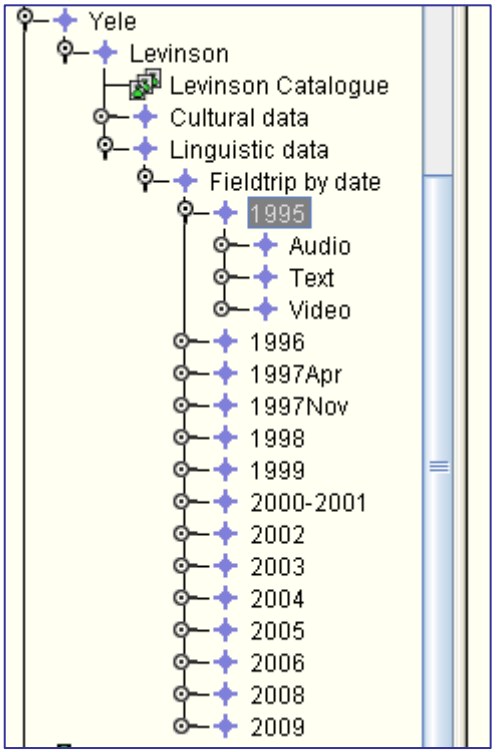
(1) Organizational aspects

Resources need to be described → IMDI metadata

Resources need to be structured (archive structure)







Archiving challenges

(2) Long term persistency requirement

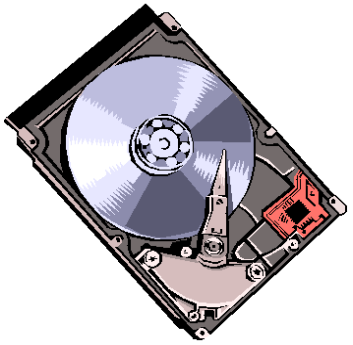
Is there a danger that we loose digital data?

YES,

UNESCO: 80% of our recordings is endangered

How much of your data and files on the notebook is organized, backed-up?

How long can media and formats be accessed?



Archiving challenges

(2) Long term persistency requirement

Open format strategy, advantages:

Easy to use (users)

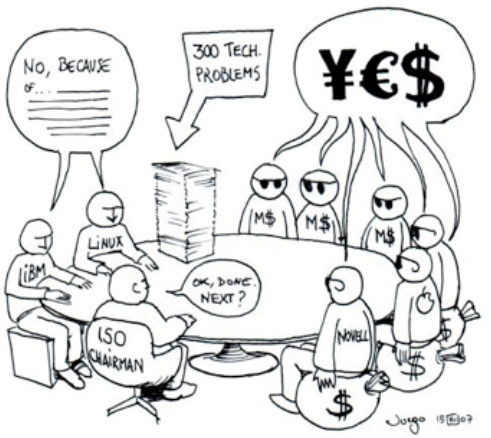
Easier to maintain (archivist)

Increased change of preservation (depositor)

↓

Forced to use certain formats

Archive role: conversion task



Archiving challenges

(2) Long term persistency requirement

Archive must be clear about accepted formats

Assist and train depositor in the use of these

Archivable:
Audio: wav,m4a
Video: mpeg1, mpge2, mp4
Text: ELAN, toolbox, XML, pdf

Non-archivable:
Ms-office files (or other proprietary formats)

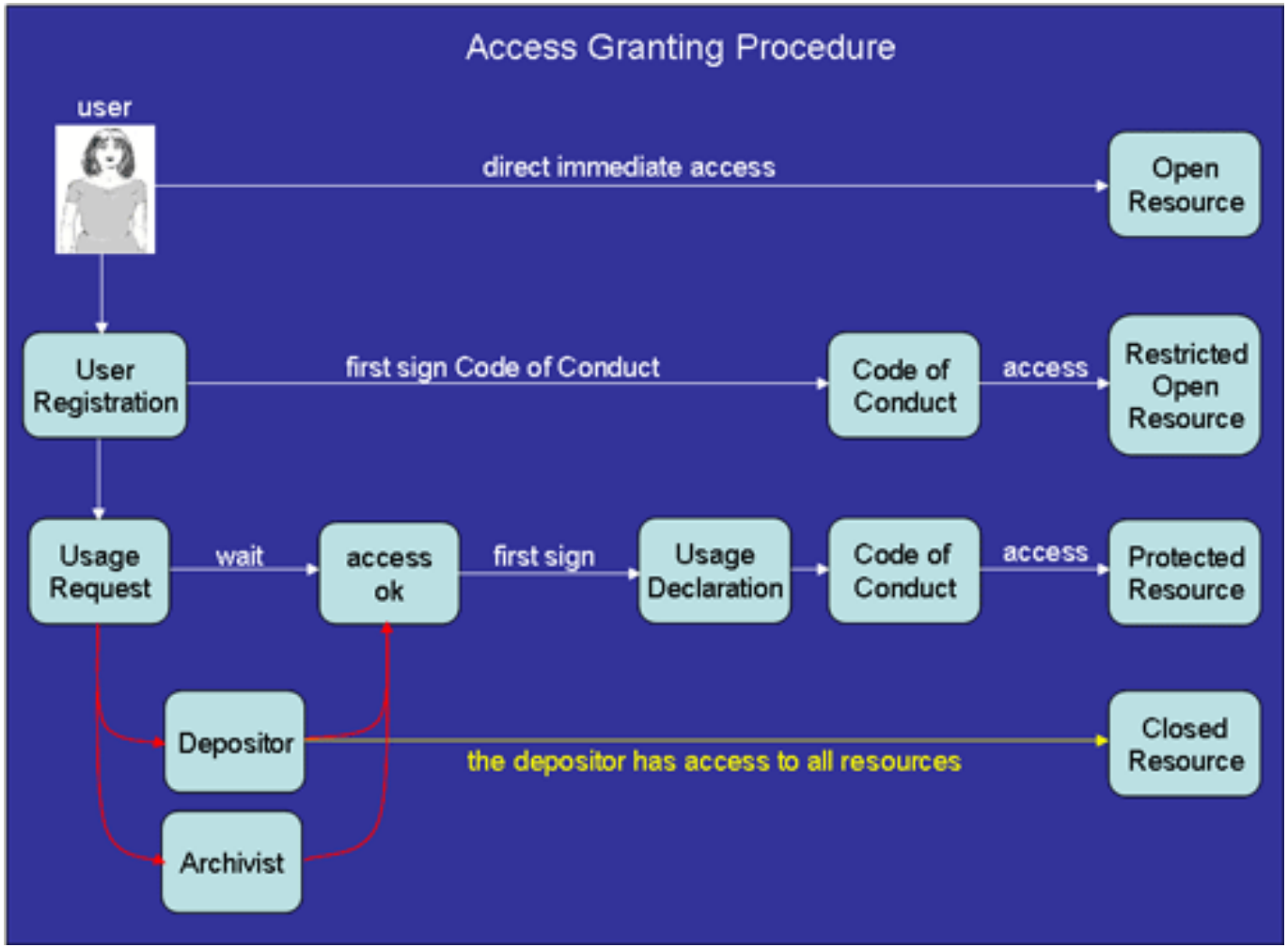
Archiving challenges

(3) Security issues
Give access to data to different users
But protect data against unauthorized access

MPI archive:

1. Metadata is accessible for everyone (OAI)
2. Resources are not always open for everyone
    - Sensitive information on actors (children, speech community etc.)
    - Ongoing research work with publication aims
3. Different access rights for different users

Archiving challenges



Archiving challenges

Ethical Rules (Code of conduct)

• Everyone will respect the Intellectual and Cultural Property Rights of the individual consultants and their communities. Wishes with respect to protecting the privacy of individuals will be respected.

• No one is allowed to offend the religious feelings of the consultants or communities. No missionary or political activities are allowed to be carried out under the umbrella of the documentation task.

• No one is allowed to use the recorded and analyzed data for commercial purposes without permission from the speech community.

• All parties must respect the national laws and the rules of the authorized organizations in the different countries.

• All parties will support the general frameworks as defined by the UN, UNESCO and WIPO where they are applicable.

Archiving challenges

Ethical Rules

• The consultants and communities will be informed openly and seriously about the goals of the DOBES program

• The parties will record and archive the data according to professional standards, but taking in consideration the limitations of the field work.

• The archivist will take appropriate steps to ensure long lifetime of the data.

• The users of DOBES data will respect the intention of the recordings and acknowledge the work that the archivist and especially the collectors have invested. In general this is to be documented by making references in publications and by mentioning the names of the consultants if this is wanted.

http://www.mpi.nl/DOBES/ethical_legal_aspects/DOBES-coc-v2.pdf

Language archiving software

Language archiving software

Clear principles: Data organization and access infrastructure

IMDI Metadata editor/Arbil

Browser and search

Coherent, consistent and persistent: Data management

LAMUS Checks the content of the files, and file type check

Assigns a persistent identifier to the uploaded file

Allows the creation of corpus structures

Web based, easy to use

Safe access: Data access rights and protection

AMS All metadata in the archive is open

All resource access can be controlled by AMS (web based)

Users remain the owners and stay in control of the access

Setting of licenses and code of conducts

Finding resources:
Different users, different needs

Different users, different needs

researcher/depositor

researchers from other domains

speech communities

general public

Different users, different needs

IMDI browser

Metadata search

TROVA

researcher/depositor

IMDI browser

Google earth overlay

Researcher from other domains

Google earth overlay

Web portals

speech communities

Google earth overlay

Facetted browser

general public

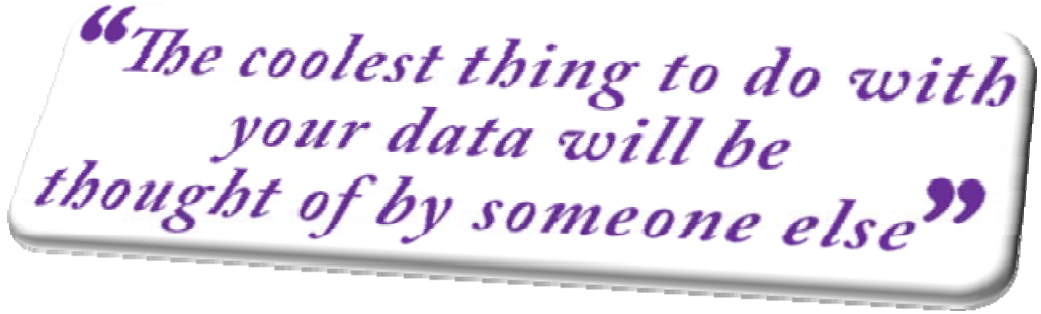
www.clarin.eu

Last slide:

Depositor fear:

Other linguists will take advantage of your hard work and take away your good ideas

The people who care about your work are the members of the speech communities – and they care about it in a different way than you do



Discussion:

Is archiving worth the effort?

Should archives prescribe formats to depositors?

Should all (or at least most) resources be open to access

(who does the data belong to?)