Inter-disciplinary Curation and Use of Language Data
Experiences from the DOBES Programme

Sebastian Drude, Peter Wittenburg, Daan Broeder
The Language Archive - Max Planck Institute for Psycholinguistics
Nijmegen, The Netherlands
DOBES: Dokumentation bedrohter Sprachen (documentation of endangered languages)

Linguistic diversity is disappearing dramatically

Since the late 90ies: “Language Documentation”: building lasting collections of recordings of language use
Some DOBES Facts

• More than 50 teams working independently
• Primary agreement: a copy of all data goes into the DOBES archive at the MPI-PL (Nijmegen)
• Result so far: ca. 15 TB of online accessible data
• Teams are interdisciplinarily composed
• Many different data types – highly interrelated at various levels
• DOBES is a fairly coherent part of a 80 TB large structured repository at the MPI-PL
• There are other initiatives and archives (e.g., HRELP at SOAS in London)
• DOBES material is about an important part of our cultural heritage

• Some purposes of documenting these languages:
  – Help maintaining language diversity
  – Preserve material for future generations

  • There is still much language diversity worldwide – so let’s create a “language bank” (like a seed bank)
  • The (descendants of) speakers themselves (will) have much interest
  • Language revitalization based on language use
The Societal Challenge 2

• Some purposes of documenting these languages:
  – Provide a comprehensive basis for research on big questions:
    • How flexible is the human language capacity?
    • What are the patterns and limits of variation? Are there language universals?
    • How did our languages evolve? (➔ understanding future development)
  – We don’t know what future generations will do with the material

• How to do preservation, and how can we offer and maintain access?
Make many “safe” copies of bit-streams and spread them (well known)

- Currently 6 full copies (physical level)
- MPG gives an institutional guarantee of 50 years for 2 of our copies
- Working on safe replication at logical level with iRODS, based on policy rules
- Selective copies to an increasing number of ‘regional archives’ worldwide

The goal: “access archives” – why?

Fundamental change:

Analogue era: “don’t touch”
Digital era: “touch frequently”
• We are bound to rely on software, which is changing
• We need to make sure that object integrity is maintained (PID – DOI, Handles – , checksum, ...)
• Digital archives are a living bodies: additions, updates, changes, extensions, new relations within and to other resources, etc. (“live archives”)
• Access “archives” can be funded from research budgets if they are used in current and future research
• For the DOBES archive and TLA in general: all bit-stream preservation costs can be neglected as long as the procedures are automatic
Curation Challenge 1

- Achieving and maintaining interpretability is much more costly (see Beagrie results)
- UNESCO: 80% of lang. & cult. recordings endangered
  - digitization is at least real-time – much will be lost?
- Important: context and provenance information (metadata)
- Question: immediate or later data conversion
  - Example: curating a wonderful 5000 entry lexicon into properly structured XML cost about 0.5 person years
  - Later data curation is multiple times more expensive (also see Beagrie results)
  - But do we have time and funds now to curate all resources we get? → NO
  - Do we need to take them as well anyways? → YES
How to achieve a coherent and consistent archive?

- Extensive checks when ingesting new data:
  - metadata
  - formats/schemas
  - relations?
  - content? (→ own library, or in future JHOVE2)

- Given the previous slide we have two parts in the archive
  A “coherent part” and a “unverified part”

- DOBES is mostly part of the coherent part

- Migrating the “unchecked part” may become very expensive, since it can not be done automatically
Migrating the coherent part can be done widely automatic, but:

- Testing is required as transformations may not be lossless
- Important that provenance information is updated

What about “out-phased” / legacy formats?

- Tapes, cassettes etc.: maintaining old equipment is expensive – some will survive, but we have too little resources to manage transformation of all material
- Digital formats could be maintained – in theory, but in praxis it might become quite complex
Economic Aspects 1

• Our data has a value since it is part of researchers’ data daily workflow
• Need to add new data to maintain attractiveness
• Costs at bit-stream level w/o. specific issues is close to 0
• Cost of digitization is “real time”, but economy of scale factor possible
• Costs of curation are not specifiable
• Whatever can be done automatic is inexpensive
• A coherent and consistent archive needs a clear economy of scale
• Current archive costs per year (without curation):

  – 1 FTE archive manager, 0.5 FTE system manager, stud assistens (economy of scale) 120 k€
  – Costs for own storage system (up to PetaBytes): 80 k€
  – Costs for 4 external copies: ~10 k€
  – 1 FTE archive software maintenance: 60 k€ \(\Sigma 270 \text{ k€}\)
  – Optionally 1 FTE access software maintenance: 60 k€
  – Optionally digitization equipment, hardware 10 k€ \(\Sigma 340 \text{ k€}\)
Summary: Technical Aspects

• The “machinery” has been working for several years in a robust way
• As much as possible is automated
• We offer “open archiving” to all researchers with serious language data
• “Unverified part” of the archive remains a point of concern
Societal Aspects

• Research organizations have a duty to maintain accessibility to their data sets
  • Best solution is to maintain an archive relevant for research
  • There may come a moment in time when our language data need to be moved
  • An organization like ANDS may be a choice
• Trust is of key importance (for depositors & users)
  • Therefore we make a clear statement: right of archiving only, respect of personal rights
  • Certification according to RAC or DSA is very important (OAIS)
Inter-disciplinary Curation and Use of Language Data Experiences from the DOBES Programme

Sebastian Drude, Peter Wittenburg, Daan Broeder
The Language Archive - Max Planck Institute for Psycholinguistics
Nijmegen, The Netherlands