NINJAL, 2026.3.14

ELAN/eaftox/InTex/InDIGO Workshop

Michinori Shimoji

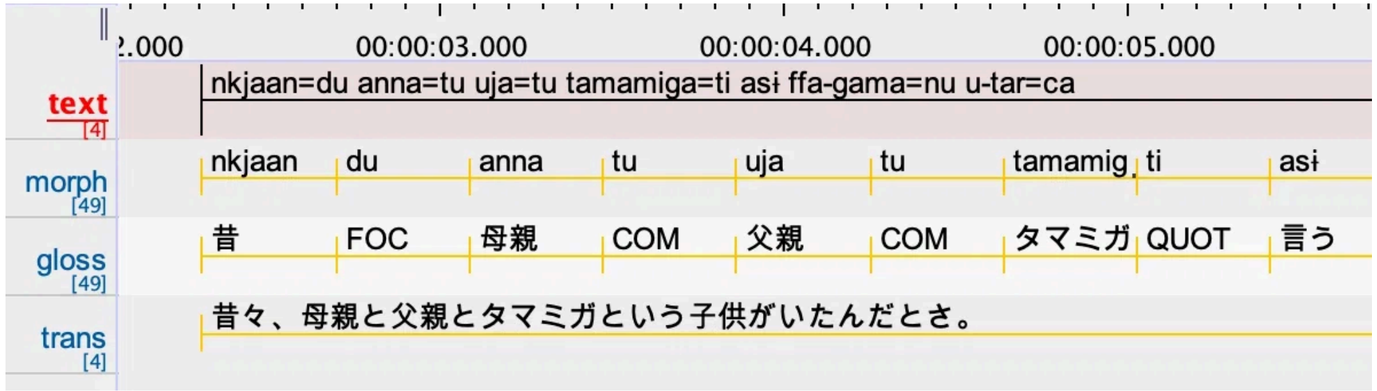
smz@kyudai.jp

ブラウザでご覧になりたい方はこちらから

このワークショップでは

- **データを整える：生データから最低限の注釈データへ**
- **データを守る：注釈済みデータを適切に保存するまで**
- **データを生かす：適切に保存したその先にできること**

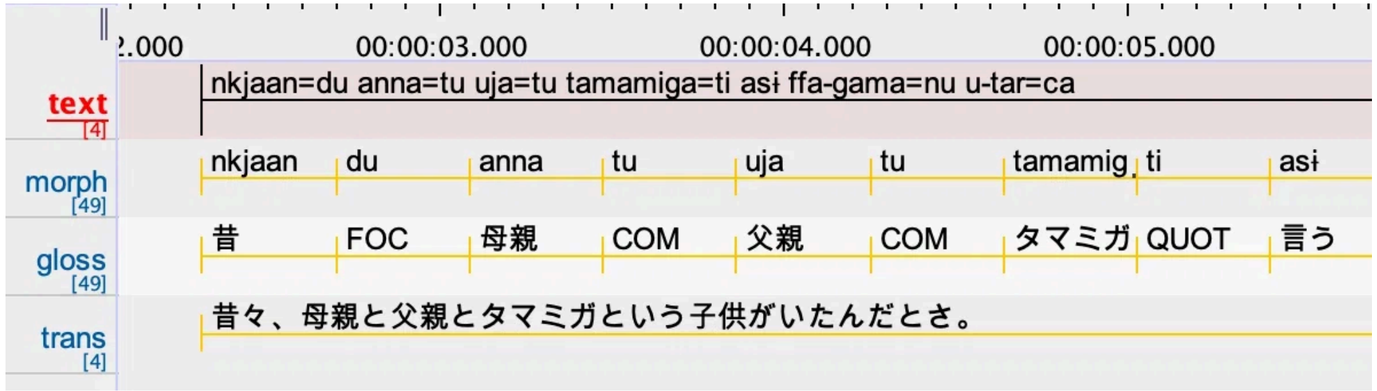
今日使うデーター式は**ここ**からダウンロードできます。



このワークショップでは

- **データを整える：録音に最低限の注釈をつけるまで**
- **データを守る：注釈済み談話データを適切に保存するまで**
- **データを生かす：適切に保存したその先にできること（紹介）**

今日使うデーター式は**ここ**からダウンロードできます。



録音について

- マイクは個人的に**AKG c520**がおすすめ。
- 録音機材はXLR対応で。
  2025年段階では**Zoom H6**/**Marantz PMD661MKIII** などが手頃

Workshop 1:
データを整える

録音に最低限の注釈をつけるまで

ELANで注釈をつけてみよう

- exercise.wav（日本語による1分の録音済み談話, frog story）
- temp3.etf
- temp4.etf
- tempwb.etf

Ex 1 冒頭10文節に対して形態素に分け、グロスをつける練習
Ex 2 形態素融合に対応するには
Ex 3 仮名を使ったとき、注釈をどうするの？
Ex 4 品詞層（POS層）も含める方針：日琉スタンダードにしたい

Workshop 2:
データを守る

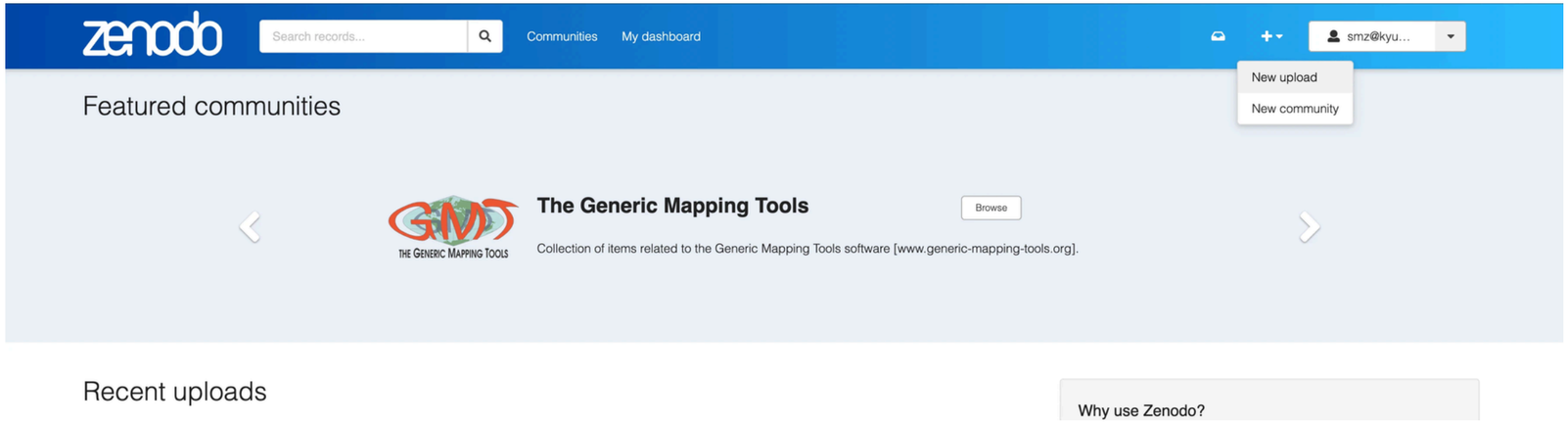
注釈済み談話データを適切に保存するまで

データをどこに保存してますか？

言語ドキュメンテーションの鉄則

- ローカルとクラウドで二重保存
- データの散逸を防ぐ
- データの更新状況の透明化

オープンアクセスの
リポジトリを活用する



3つのステップ

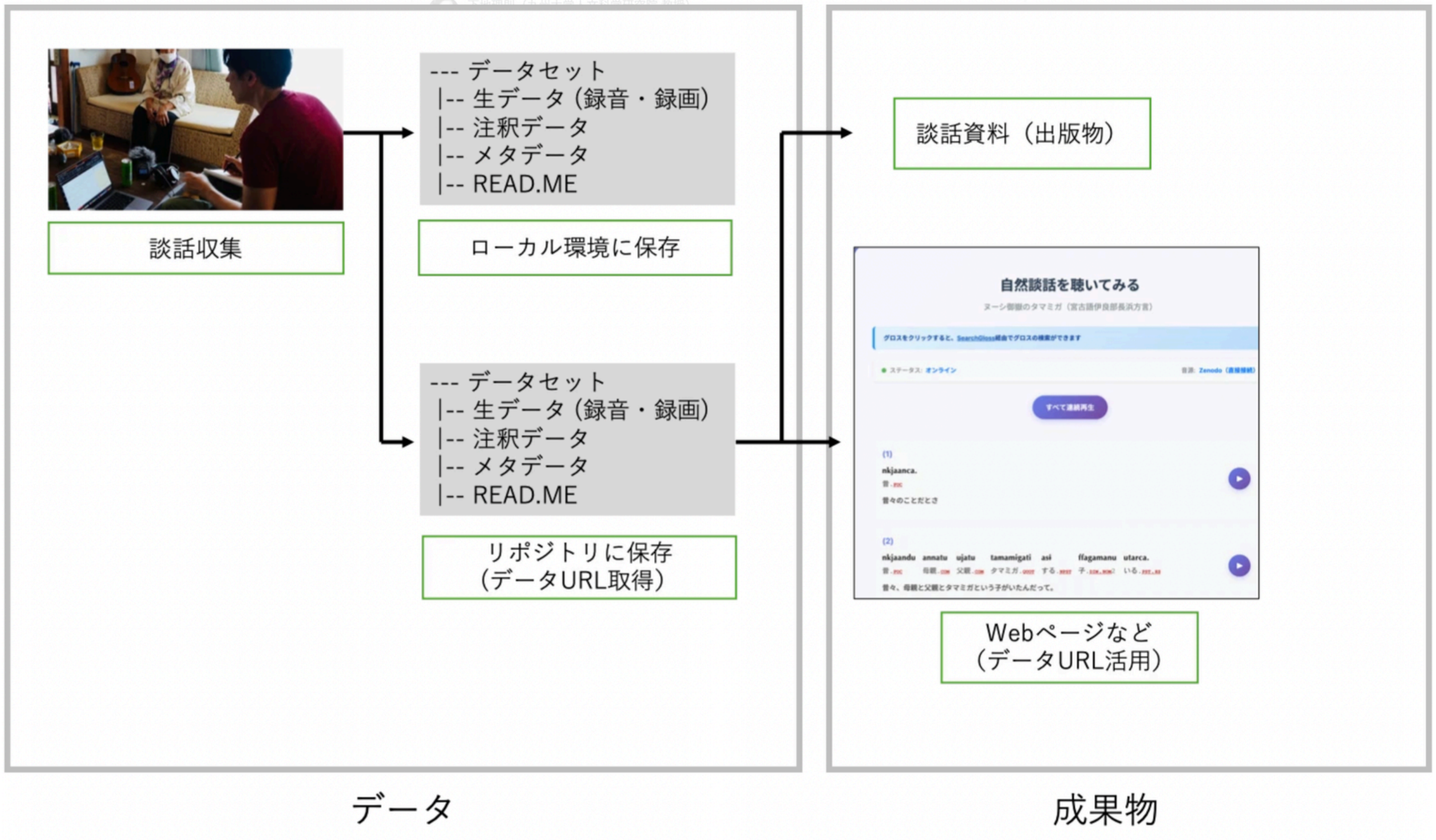
note記事「消滅危機言語の記録保存（特に談話データを例に）」
詳しくは**ここで**読めます

Step 1 絶対にすべきこと：データの永続性。自分のローカル環境（パソコンのHDとかSDカード保存とか）に加えて、**ウェブ上の安定したリポジトリ**にも関連データを全て置く。関連データとは、生データ（録音、録画)、書き起こしデータ（注釈付き)、メタデータ（話者情報、注釈者情報など）

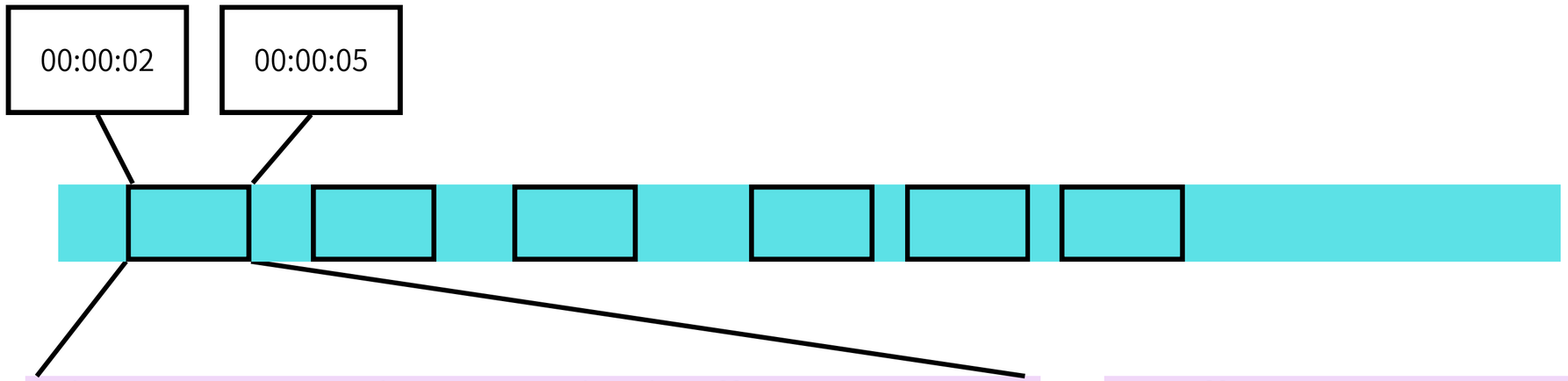
Step 2 望ましいこと：少なくとものちの研究者/コミュニティが利用しやすいような仕組みを整えておく。後の修正が入る場合のバージョン管理の徹底。個々のデータに対するデジタル資料の固有識別番号であるDOI（Digital Object Identifier）の取得など。

Step 3 できると素晴らしいこと：コミュニティが、自らの言語復興や継承言語教育のリソースに使いやすくする。目的特化型のWebページを作り、誰でもわかりやすい方法で談話資料を公開する。談話の例文ごとに音声も聞けるような仕組みなどがあると素晴らしい。

3つのステップ：具体的な実現手順



ELANで生み出される2つのデータ

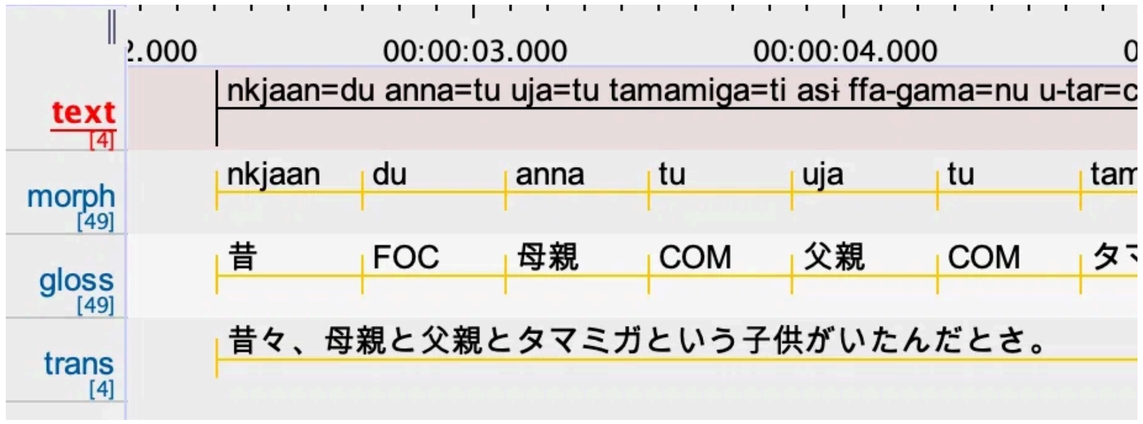


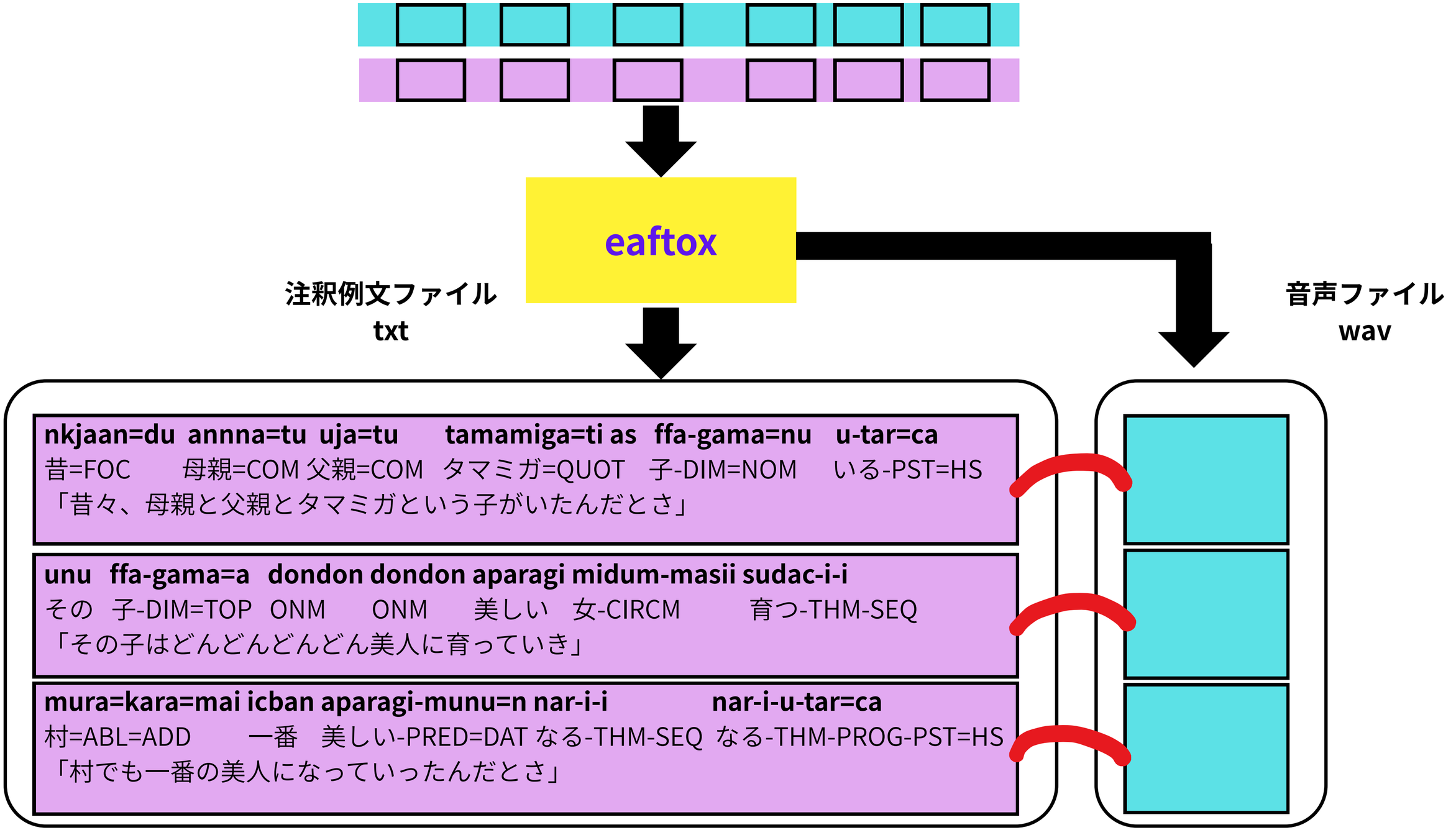
音声データ

nkjaandu annnatu ujatu tamamigati as ffagamanu utarca.

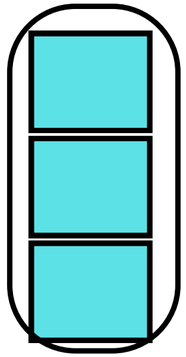
unu ffagamaa dondon dondon...

注釈例文データ




eaftox
注釈例文ファイル
txt
音声ファイル
wav
nkjaan=du annna=tu uja=tu tamamiga=ti as ffa-gama=nu u-tar=ca
昔=FOC 母親=COM 父親=COM タマミガ=QUOT 子-DIM=NOM いる-PST=HS
「昔々、母親と父親とタマミガという子がいたんだとさ」
unu ffa-gama=a dondon dondon aparagi midum-masii sudac-i-i
その 子-DIM=TOP ONM ONM 美しい 女-CIRCM 育つ-THM-SEQ
「その子はどんどんどんどん美人に育っていき」
mura=kara=mai icban aparagi-munu=n nar-i-i nar-i-u-tar=ca
村=ABL=ADD 一番 美しい-PRED=DAT なる-THM-SEQ なる-THM-PROG-PST=HS
「村でも一番の美人になっていったんだとさ」

注釈内容と音声を関連づけて別々に保存



nkjaan=du annna=tu uja=tu tamamiga=ti as ffa-gama=nu u-tar=ca
昔=FOC 母親=COM 父親=COM タマミガ=QUOT 子-DIM=NOM いる-PST=HS
「昔々、母親と父親とタマミガという子がいたんだとさ」

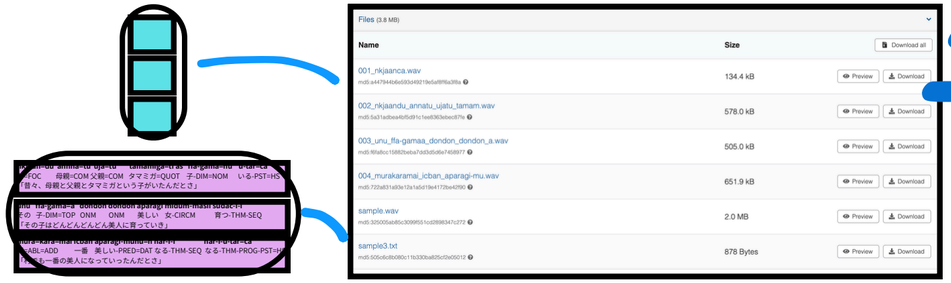
unu ffa-gama=a dondon dondon aparagi midum-masii sudac-i-i
その 子-DIM=TOP ONM ONM 美しい 女-CIRCM 育つ-THM-SEQ
「その子はどんどんどんどん美人に育っていき」

mura=kara=mai icban aparagi-munu=n nar-i-i nar-i-u-tar=ca
村=ABL=ADD 一番 美しい-PRED=DAT なる-THM-SEQ なる-THM-PROG-PST=HS
「村でも一番の美人になっていったんだとさ」

Files (3.8 MB)

Name	Size	Download all
001_nkjaanca.wav md5:a447944b6e593d49219e5af8ff6a3f8a	134.4 kB	Preview Download
002_nkjaandu_annatu_ujatu_tamam.wav md5:5a31adbea4bf5d91c1ee8363ebec87fe	578.0 kB	Preview Download
003_unu_ffa-gamaa_dondon_dondon_a.wav md5:f6fa8cc15882beba7dd3d5d6e7458977	505.0 kB	Preview Download
004_murakaramai_icban_aparagi-mu.wav md5:722a831a93e12a1a5d19e4172be42f90	651.9 kB	Preview Download
sample.wav md5:325005ab85c3099f551cd2898347c272	2.0 MB	Preview Download
sample3.txt md5:505c6c8b080c11b330ba825cf2e05012	878 Bytes	Preview Download

個別保存するメリット絶大（後述）



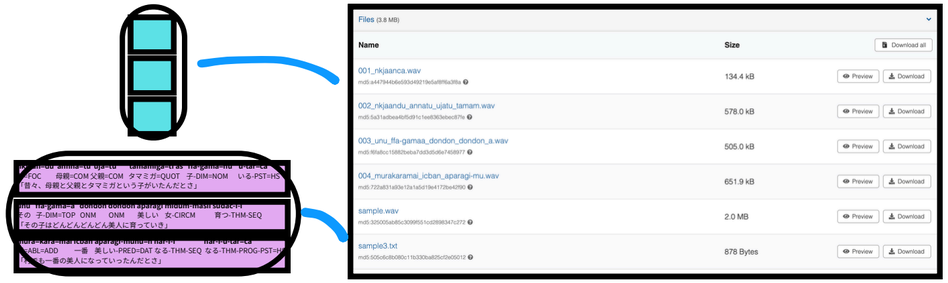
Files (3.8 MB)

Name	Size	
001_nkjaanca.wav	134.4 kB	Preview / Download
002_nkjaandu_annatu_ujatu_tamam.wav	578.0 kB	Preview / Download
003_unu_ffa-gamaa_dondon_dondon_a.wav	505.0 kB	Preview / Download
004_murakaramai_icban_aparagi-mu.wav	651.9 kB	Preview / Download
sample.wav	2.0 MB	Preview / Download
sample3.txt	878 Bytes	Preview / Download

Zenodo

文ごとに再生可能な
Webページ

ワークフロー



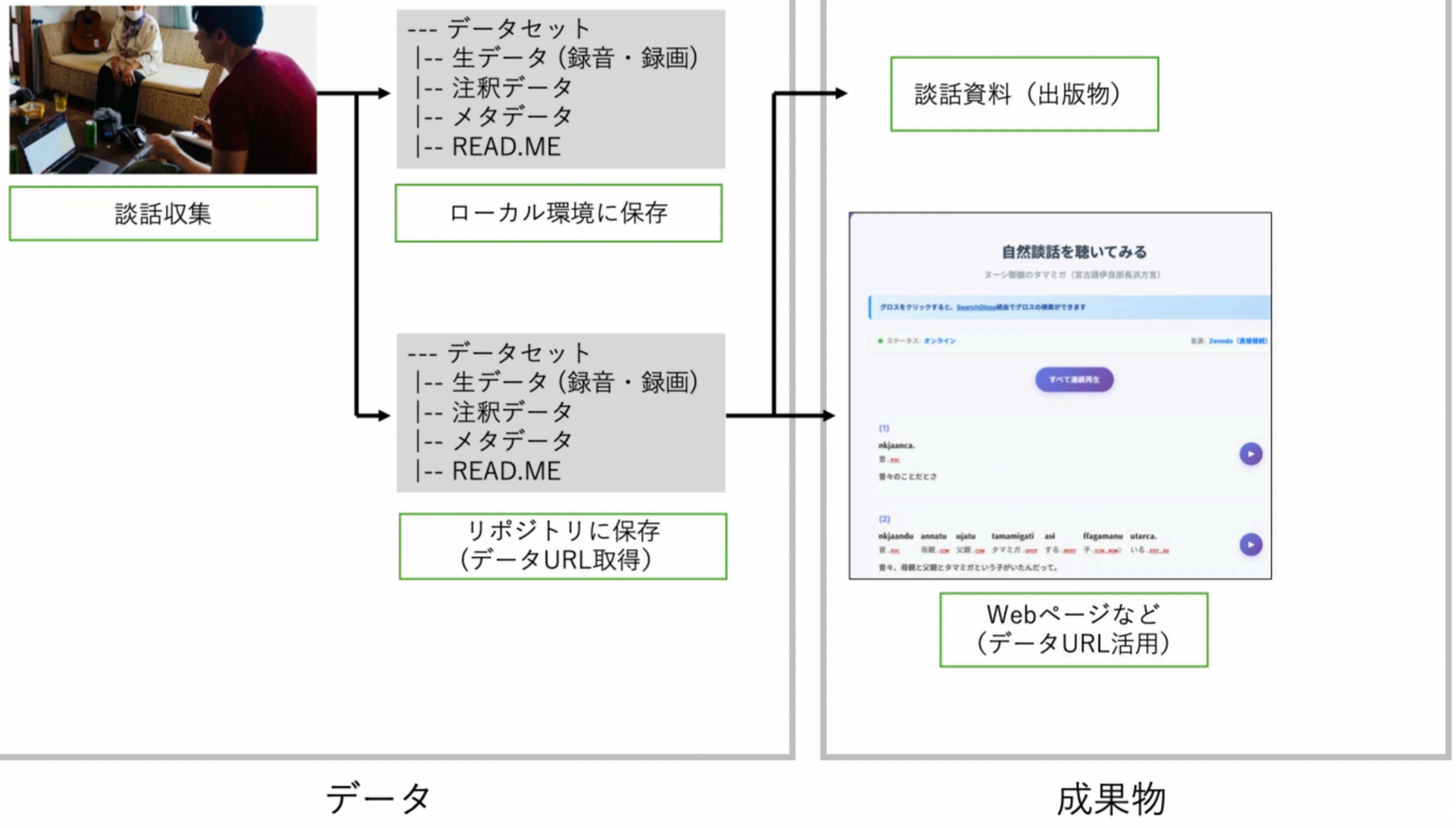
Files (3.8 MB)

Name	Size	Download all
001_nkjaanca.wav md5:a4479f4d06e593d49219e5a8ff6a3f8a	134.4 kB	Preview Download
002_nkjaandu_annatu_ujatu_tamam.wav md5:5a31adbea4bf5d91c1ee8636ebec87fe	578.0 kB	Preview Download
003_unu_ffa-gamaa_dondon_dondon_a.wav md5:6fda8cc1588f2beba7dd3d5d6e7458977	505.0 kB	Preview Download
004_murakaramai_icban_aparagi-mu.wav md5:720a831e83e12a1a5d19e417fbe42f90	651.9 kB	Preview Download
sample.wav md5:325005ab865c3099f551cd2898347c272	2.0 MB	Preview Download
sample3.txt md5:505c6c8b080c11b330ba825cf2e05012	878 Bytes	Preview Download

eaftox

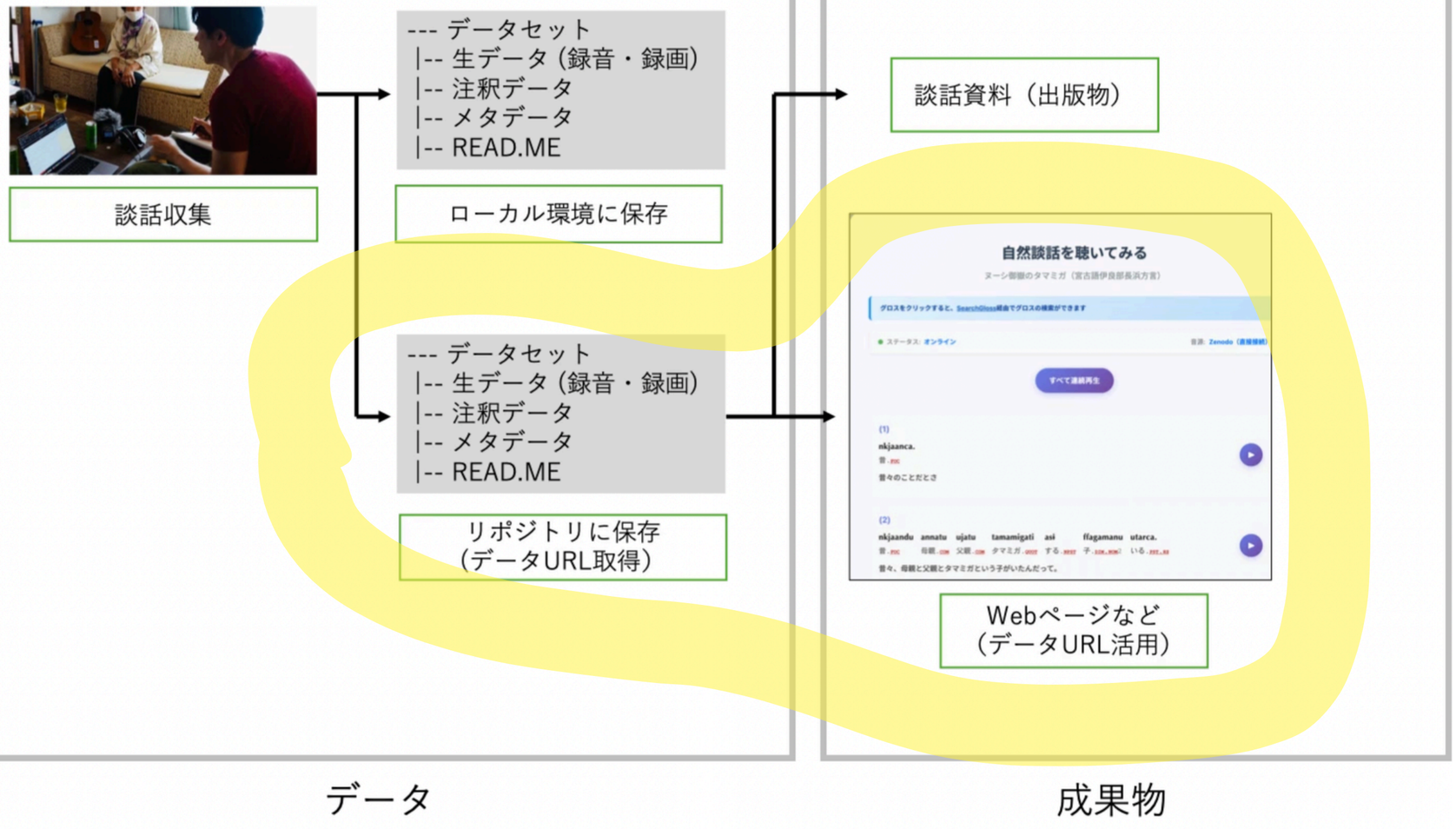
Zenodo

Zenodoにデータをアップしてみよう

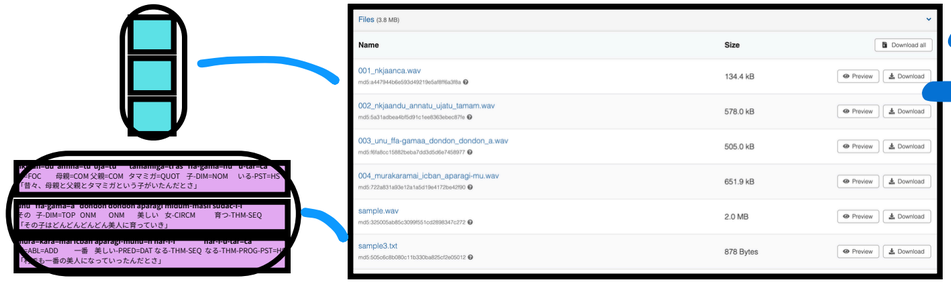


Workshop 3:
データを生かす
適切に保存したその先にできること

Workshop 3でやること



InTEx: zenodo経由でWebページ作成

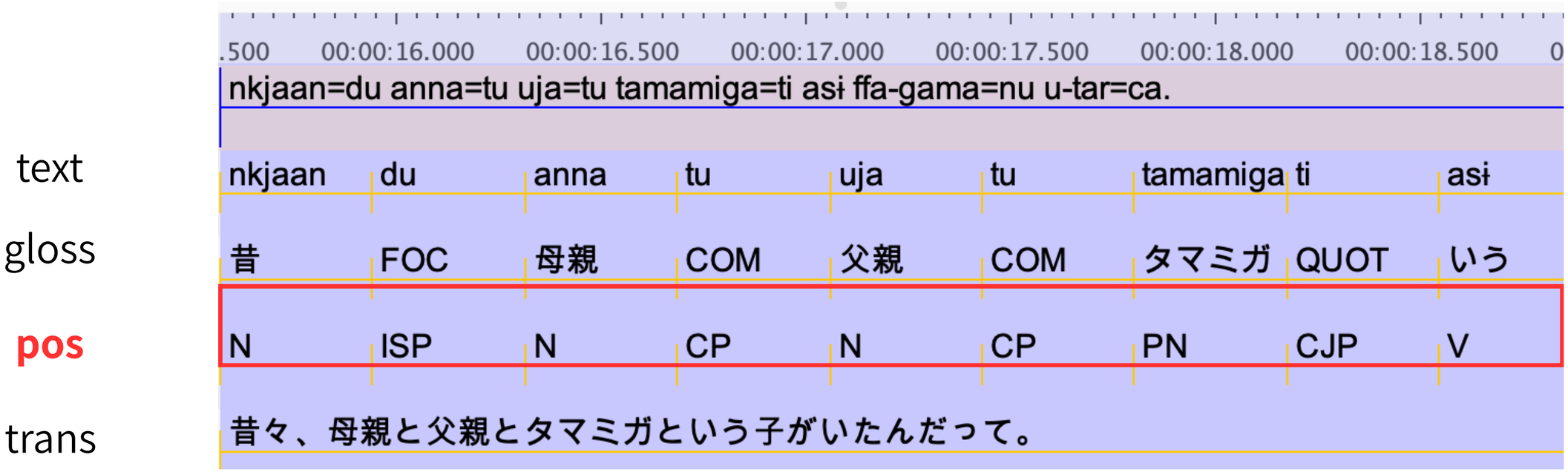


eaftox

Zenodo

InTEx

pos層を整備しておくと可能性が広がる



pos層を整備しておくと可能性が広がる

- ELANのデータから簡易電子辞書を作る際に便利　→link
- ELANのデータからPOS一覧を作り、簡易文法スケッチを作成する足掛かりになる
  （ELAN → json → Pythonスクリプトで生成）　→link
- 機械学習でグロス自動化を実装する上で、精度の高い教師データになる（下地2026)　→link

おわりに

このワークショップの振り返り

WS1 データを整える：録音に最低限の注釈をつけるまで
WS2 データを守る：注釈済み談話データを適切に保存するまで
WS3 データを生かす：適切に保存したその先にできること

- WS2, 3の解説動画は**ここから**ダウンロードできます（九大の下地ゼミ院演習録画）
- eaftoxの解説記事と動画は**こちらから**（note記事）
- InTExの解説記事と動画は**こちらから**（note記事）
- 消滅危機言語のドキュメンテーションの概説記事は**こちらから**（note記事）
- グロス付けのツール全般は**こちらから**（下地理則研究室HP内）