AHN - La TEI pour des données de l'oral : un fichier ODD pour tout spécifier ?

icar
UMR 5191 - CNRS / Université Lyon 2
Interactions, Corpus, Apprentissages, Représentations

Carole Etienne, Nov 2016

Consortiums IRCOM et CORLI

- Infrastructure de recherche **IRCOM** puis **CORLI (Huma-Num)**
  - IRCOM : IR **C**orpus **O**raux et **M**ultimodaux
    - données orales
    - 6 groupes de travail : corpus en tant qu'objet scientifique, **interopérabilité**, multilinguisme et plurilinguisme, multimodalité, problèmes juridiques, archivage
    - identifier les pratiques et les ressources et proposer des standards

      >>> **un format commun modulaire pour l'oral**

  - CORLI : **COR**pus, **L**angues et **I**nteractions regroupe IRCOM et Corpus écrits
    - groupe de travail **interexplo** interopérabilité et exploration de corpus
    - mise en commun des ressources pour établir des **bonnes pratiques**
    - **format écrit et oral >>> vers un format commun**

Les corpus oraux

- Un corpus oral
  - donnée primaire = un enregistrement audio ou vidéo
  - des métadonnées sur le corpus, la situation enregistrée, l'enregistrement physique, les locuteurs, les droits d'accès, les annotations
  - la représentation écrite est donc une donnée **secondaire** recouvrant **le verbal** (glose, traduction, phonèmes, …) mais aussi le **non-verbal** (productions vocales, éléments prosodiques, gestes, regards…)

    ➔ des annotations, accompagnées d'une convention
  - ces annotations pouvant être **alignées à du texte** ou **alignées au signal, à un intervalle de temps**
    - utilisation de **logiciels d'alignement** avec des formats propriétaires : transcriber, praat, clan, elan, anvil, eXmaralda, … des structures soit en liste soit en partition
    - pas d'annotation ***directe*** en xml, pas d'utilisation de logiciels comme oXygen, xml reste un format d'échange ou d'export

Les corpus oraux et la TEI

- La **Text** Encoding Initiative pour représenter les corpus oraux
  - un **seul fichier** pour les métadonnées et la représentation textuelle des annotations ➔ on ne "perd" pas les métadonnées pour les consulter dans les analyses et les outils de visualisation des résultats
  - un standard **déjà utilisé** dans des projets de l'oral Alipe, Clapi ou Colaje
  - une initiative **européenne** ISO-TEI : proposition de spécifications communes pour faire évoluer la TEI discutées en novembre 2015 avec le Tei Council
  - un article IRCOM-interopérabilité dans jTEI "**Using the TEI as a pivot format for oral and multimodal language corpora"**
  - travaux de recherche sur les écrits non planifiés et l'oral
  - **granularité** des métadonnées comme des annotations
  - sémantique et nombre de balises >>> fichier **ODD**

Les corpus oraux en TEI

- Un formulaire pour les métadonnées
  - tentative d'utilisation du mode auteur d'oXygen pour les métadonnées
  - des solutions par type de données : quelle unité pour le corpus, une situation à contextualiser, un enregistrement vidéo complexe, des locuteurs à décrire, des guides d'annotations, des droits particuliers
  - un export automatique en Tei et dans les formats d'archivage Dublin Core/Olac, CMDI (Component MetaData Infrastructure) Clarin
- Un export automatique des annotations à partir des versions des logiciels d'annotations
  - le fichier TEI devient le format pivot d'un logiciel à l'autre
  - attention aux différences de structuration et de métadonnées, aux lignes d'annotation

Les choix en matière de représentation TEI

- Premier constat : nombre de balises et mauvaise interprétation de la sémantique
  - **erreurs de sémantique**
  - niveaux de granularité
  - variantes
  - éléments (desc, span) ou propriétés (type, n) **génériques**
- Standardisation pour sélectionner, visualiser
  - vocabulaire contrôlé
  - **restreindre à un sous-ensemble d'éléments … et de propriétés**
  - exemplifier

>>> optimiser le temps passé, en particulier **usage occasionnel**

>>> anticiper la mise en commun des données

>>> diffuser

Quelques exemples dans les métadonnées

- Le type de situation orale

```
<profileDesc>
        <textDesc>
            <channel mode="s" ana="#face_a_face #en_public"/>
            <domaine type="professionnel" ana="#conversation"/>
            <interaction type="inapplicable" active="2" passive="1"/>
            <preparedness ana="#prepare"/>
            <purpose ana="#politique"/>
        ...
```

- Les locuteurs

```
<profileDesc>
        <particDesc>
            <listPerson>
            <person xml:id="P1" age="#adulte/jeune/enfant/retraite" sex="1">
                <langKnowledge>
                    <langKnown tag="fr" level="first"/>
                </langKnowledge>
                <occupation ana="#avant_scolarité/en_scolarite/en activité/sans emploi/retraité"/>
        ...
```

Âge : vocabulaire contrôlé

Quelques exemples dans les métadonnées

- L'enregistrement manque d'éléments sur l'anonymisation et la qualité

desc + type

```
<sourceDesc>
  <recordingStmt>
    <recording>
      <media type="audio" mimeType="audio/mp3" dur-iso="P36M" url="//xxx.mp3"
      <desc type="anonymisation" ana="#anonymisation_ana"/>
      <desc type="qualite" ana="#qualite_ana"/>
    </recording>
    <recording>
    <media type="video" mimeType="video/mp4" dur-iso="P20M" url="//xxx.mp4">
    ....
    </recording>
    ...
```

Quelques exemples dans les annotations

- Les timelines pour aligner
  - Définition

```
<timeline unit="s" origin="#T0">
                <when xml:id="Tn-1" absolute="00:00:05.26"/>
                <when xml:id="Tn"  />
                <when xml:id="Tn+1" absolute="00:00:06.00"/>
                …
```

  - Utilisation exemple d'annotations non verbales et verbal es

```
<annotationBlock who="#A" start"#T1" end="#T3">
            <u> <w>bon</w>
                 <w>donc</w>
                <anchor synch="#T2"/>
                <w>on</w>
                 <w>a</w>
            </u>
 <spanGrp type="tempo"> <span from="#T1" to="#T2">lengthening</span></spanGrp>
 <incident start="#T2" end="#T3"> <desc> déplacement de chaises </desc> </incident>
<annotationBlock/>
```

Quelques exemples dans les annotations

- Les différentes couches d'annotations

```
<annotationBlock start="#T1" end="#T2" who="MainLine" xml:id="a1">
    <u>
        <seg>this is it!</seg>
    </u>
     <spanGrp type="Words">
       <span from="#T1" to="#T3" xml:id="a2">this
            <spanGrp type="Phones">
                <span from="#T1" to="#T4" xml:id="a5">ð</span>
                <span from="#T4" to="#T5" xml:id="a6">ɪ
                    <spanGrp type="PhonesInformation">
                        <span target="#a6" xml:id="a8">nucleus</span>
                    </spanGrp>
                </span>
                <span from="#T3" to="#T5" xml:id="a7">s</span>
            </spanGrp>
       </span>
       <span from="#T3" to="#T6" xml:id="a3">is</span>
       <span from="#T3" to="#T6" xml:id="a3">it</span>
  </spanGrp>
</annotationBlock>
```

Un fichier ODD pour tout faire ?

- Centraliser le sous-ensemble d'éléments et de propriétés
- Centraliser les éléments optionnels et obligatoires
- Décrire le vocabulaire contrôlé
- Donner des exemples réutilisables en TEI
- Générer un schéma pour les fichiers TEI : **valider/invalider** les exports des corpus oraux déjà structurés de manière automatique
- Générer un formulaire de saisie des métadonnées !

Un fichier ODD pour tout faire ?

- Centraliser le sous-ensemble d'éléments et de propriétés
  - éléments : moduleRef + include
  - propriétés : attList et attDef
- Centraliser les éléments optionnels et obligatoires
  - éléments : elementSpec , content et one / zero / oneOrMore / zeroOrMore
  - propriétés : attDef et usage="req"
- Décrire le vocabulaire contrôlé
  - valList et valItem avec un desc par langue
- Donner des exemples réutilisables en TEI
  - exemplum
- *Générer un schéma pour les fichiers TEI : **valider/invalider** les exports des corpus oraux déjà structurés de manière automatique*
- *Générer un formulaire de saisie des métadonnées !*

Un fichier ODD pour tout faire ?

- Centraliser le sous-ensemble d'éléments et de propriétés

```
<moduleRef  key="spoken"
include="recordingStmt recording transcriptionDesc annotationBlock u vocal
            kinesic incident pause" />
```

```
<element name="idno"  module="header">
    <attList>
            <attDef ident="type" usage="req" value="title" mode="change"/>
    </attList>
</element>
```

Un fichier ODD pour tout faire ?

- Centraliser les éléments optionnels et obligatoires

```
<zeroOrMore>
      <element name="respStmt"  module="core">
              <content>
                 <oneOrMore>
                        <element name="resp" module="core"> … </element>
                        <element name="name" module="core" default="alphanum">… /element>
                 </oneOrMore>
              </content>
       </element>
</zeroOrMore>
```

```
<attList>
              <attDef ident="type" usage="req" value="title" mode="change"/>
</attList>
```

Un fichier ODD pour tout faire ?

- Décrire le vocabulaire contrôlé

```
<attList>
        <attDef ident="type" usage="req" value="quality" mode="change"/>
        <attDef ident="subtype" usage="req"  mode="change">
                <valList>
                    <valItem ident="less_noisy">
                            <desc xml:lang="fr">inaudible ou bruité moins de 5%</desc>
                            <desc xml:lang="en">inaudible or noisy less than 5%</desc>
                    </valItem>
                    <valItem ident="noisy">
                            <desc xml:lang="fr">inaudible ou bruité plus de 5%</desc>
                            <desc xml:lang="en">inaudible or noisy more than 5%</desc>
                    </valItem>
                    <valItem ident="soundproof_room">
                            <desc xml:lang="fr">chambre sourde</desc>
                            <desc xml:lang="en">soundproof</desc>
                    </valItem>
                    <valItem ident="OTHER">
                            <desc xml:lang="fr">Autre</desc>
                            <desc xml:lang="en">Other</desc>
                    </valItem>
                </valList>
        </attDef>
</attList>
```

Un fichier ODD pour tout faire ?

- Donner des exemples réutilisables en TEI

```
<exemplum versionDate="2016-10-04" xml:lang="fr">
  <sourceDesc>
    <recordingStmt>
      <recording>
        <p> une balise media par enregistrement</p>
            <media type="audio" mimeType="audio/wav" dur-iso="P36M" url="xxx">
                    <p> un signal audio anonymisé</p>
                    <desc type="anonymisation" subtype="oui"/>
                    <desc type="qualite" subtype="peu_bruite"/>
            </media>
      </recording>
  <recording>
            <media type="video" mimeType="video/mp4" dur-iso="P36M" url="xxx">
                     <p> un signal video anonymisé</p>
                    <desc type="anonymisation" subtype="oui"/>
                    <desc type="qualite" subtype="peu_bruite"/>
            </media>
  </recording>
  <recording>
             <media type="audio" mimeType="audio/wav" dur-iso="P36M" url="xxx">
                    <p> un signal audio non anonymisé</p>
                    <desc type="anonymisation" subtype="non"/>
                    <desc type="qualite" subtype="peu_bruite"/>
             </media>
  </recording>
</recordingStmt>
</sourceDesc>
</exemplum>
```

Conclusion …

- Les difficultés
  - le temps passé dans les premiers projets oraux utilisant la TEI
  - des données hétérogènes à regrouper pour différents besoins : des sélections, des outils automatiques, de la visualisation
- Les solutions
  - Un fichier ODD très détaillé
  - vocabulaire contrôlé
  - ODD à figure de schéma mais quels outils pour exploiter ce fichier ODD , Roma et Byzantium ne gèrant pas les ODD détaillés?
- L'interopérabilité renvoie en boomerang tous les problèmes non résolus en amont !