# Méta-données et ressources linguistiques

# Contexte général

- Uniformisation des formats d'échanges
  - XML est un acquis (cadre: W3C; www.w3.org)
- Intégration des données et des méta-données associées
  - Expérience pionnière de la TEI (Text Encoding Initiative; www.tei-c.org)
  - Vision ouverte du W3C avec RDF/RDFS/OWL
- Évolution des pratiques documentaires
  - E.g. littérature grise en ligne (idem thèses, publications)
    - Normalisation des formats (TEI?)
    - Évolution des besoins en méta-données (e.g. versions)

# Définition

- Méta-donnée : toute donnée décrivant une autre donnée
  - Identification (titre, auteur etc.)
  - Administration (droits)
  - Localisation (objet physique, URL)
  - Utilisation (caractéristiques physique, format de fichier)

# Vision historique

- Les méta-données : à la base des pratiques documentaires traditionnelles
  - Unimarc
- Extension de ces pratiques au web
  - Dublin Core
- Mise en place de mécanismes spécifiques d’accès à ces méta-données
  - Indexation en aveugle (moteurs de recherche)
  - Mécanismes de moissonnage (e.g. OAI => OLAC)

# Un jeu réduit de méta-données

## Le Dublin Core

(DCMS: Dublin Core Metadata Set)

https://www.dublincore.org/

# Utilisation simplifiée: moissonnage automatique de sites

- DCMES (Dublin Core Metadata Element Set) : un jeu simplifié de 15 éléments de donnée pouvant être attachés à tout document présent sur le web (ISO 15836)
- DCMI Metadata Terms : 55 (avec les 15 du DCMES)
- Usage typique : balises HTML META dans <head>

<!DOCTYPE html PUBLIC "-// W3C // DTD XHTML 1.0 Strict // EN " "http://www.w3.org/TR/xhtml1 /DTD/xhtml1-strict.dtd" >
< html xmlns = " http:// www.w3. org/1999/xhtml" xml:lang="fr" lang="fr" >
<! -- Generated from data / hea-home.php , ../../ smarty /{ head . tpl } -- >
< head >
<META NAME = “Keywords” CONTENT = “metadata, Dublin Core, TEI”>
<META NAME = “Description” CONTENT = “Discussion du concept de métadonnée, des différents formats, et des utilisation pour les ressources linguistiques.”>
</ head >

# Eléments du Dublin Core

- Title
- Author
- Subject
- Description
- Publisher
- Contributor
- Date
- Type
- Format
- Identifier
- Source
- Language
- Relation
- Coverage
- Rights

# Exemple d’utilisation

<meta
NAME=“DC:identifier”
CONTENT = “http://www.cas.usf.edu/lis/lis6511/”>
<meta
NAME=“DC:author”
CONTENT=“Vicki L. Gregory”>
<meta
NAME=“DC:subject”
CONTENT=“collection development, selection, weeding, preservation, intellectual freedom”>

# Exemple (suite)

<meta
NAME="DC:description"
CONTENT = A survey course dealing with all aspects of collection development and collection maintenance issues.">
<meta
NAME="DC:date"
Content="January 5, 1999">
<meta
NAME="DC:language"
CONTENT="English">
<meta
NAME="DC.:format"
CONTENT="HTML">

# Intégration méta-données + texte

La *Text Encoding Initiative* (TEI)

# TEI Vision historique

- Premières recommandations en 1994
- Un consortium gérant les évolutions des directives
  www.tei-c.org
- Des DTD spécifiques pour des genres de documents comme les manuscrits, les transcriptions de l'oral, les dictionnaires, etc.
- Très utilisée aujourd'hui en humanité numériques
  Corpus 14 – histoire du français écrit peu lettré (correspondance de la grande Guerre)
  https://www.univ-montp3.fr/corpus14/

# Infrastructure TEI

- Format XML : validation des descriptions TEI
- Recommandations TEI P5 très documentées avec de nombreux exemples (en 8 langues dont Fr)

http://www.tei-c.org/release/doc/tei-p5-doc/fr/html/REF-ELEMENTS.html

- TEI ODD (One Document Does it all) permet de générer à la fois le manuel d'encodage XML et les schémas de validation des différents formats
- Espace de nommage des balises XML

# TEI - Documentation des textes

Fondamental : documenter les textes électroniques

- identification et suivi
  - cf. catalogage des documents électroniques (ex. Silfide)
  - cf. échange des documents électroniques

Comment : entête TEI (TeiHeader)

- aspects bibliographiques du document source (auteur, éditeur, édition, etc.)
- aspects propres au document électronique (aspects bibliographiques, codage, historique des révisions etc.)

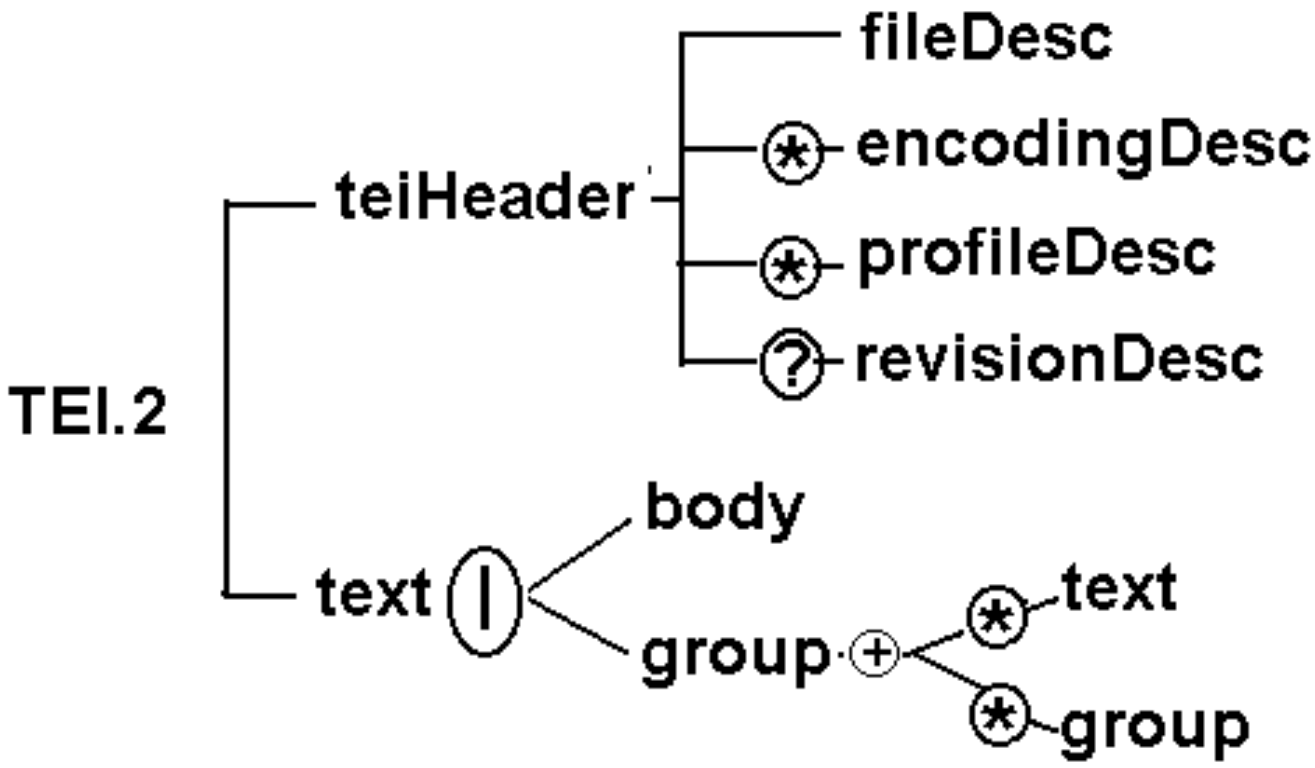
# Structure d’un document TEI

```
<TEI.2>
<teiHeader> <!-- ... --> </teiHeader>
<text>
<front>
<!-- front matter of copy text goes here. -->
</front>
<body>
<!-- body of text goes here. -->
</body>
<back>
<!-- back matter of text, if any, here. -->
</back>
</text>
</TEI.2>
```

# L’en-tête TEI

- description bibliographique normalisée
  - du document electronique (titre, responsables, maison d’édition, source....)
  - de son encodage  (éléments presents, codes internes...)
  - de sa classification (sujets,  genres...)
  - de son état de revision
- facilite la découverte des ressources sur réseau et dans les bases de données

# Structure d'un document TEI




A Group element must contain at least one text or group element

# En-tête TEI: structure générale

```
<teiHeader>
        <fileDesc>
            <titleStmt>
                <title>…</title>
            </titleStmt>
            <sourceDesc>
                …
            </sourceDesc>
        </fileDesc>
        <encodingDesc>
            …
        </encodingDesc>
        <profileDesc>
            …
        </profileDesc>
        <revisionDesc>
            …
        </revisionDesc>
</teiHeader>
```

# Exemple d'un entete TEI

```
<teiHeader>
   <fileDesc>
        <titleStmt>
               <title>Thomas Paine: Common sense, a machine-
   readable transcript</title>
               <respStmt>
                      <resp>compiled by</resp>
                      <name>Jon K Adams</name>
               </respStmt>
        </titleStmt>
        <publicationStmt>
               <distributor>Oxford Text Archive</distributor>
               </publicationStmt>
        <sourceDesc>
               <bibl>The complete writings of Thomas
   Paine,collected and edited by Phillip S. Foner (New York,
   Citadel Press, 1945) </bibl>
        </sourceDesc>
   </fileDesc>
<teiHeader>
```

# D’une initiative à l’autre…

| | |
|---|---|
| <title type=“main”> | DC.title.main |
| <author> | DC.creator.name |
| <publicationStmt> | DC.publisher.name |
| <sourceDesc> | DC.source |
| <classDecl> | DC.subject.schema |

…en attendant les répertoires de méta-données.

# En-tête TEI: structure générale

```
<teiHeader>
        <fileDesc>
            <titleStmt>
                <title>…</title>
            </titleStmt>
            <sourceDesc>
                …
            </sourceDesc>
        </fileDesc>
        <encodingDesc>
            …
        </encodingDesc>
        <profileDesc>
            …
        </profileDesc>
        <revisionDesc>
            …
        </revisionDesc>
</teiHeader>
```

# En-tête TEI : description des contenus

- Déclaration de balisage dans <encodingDesc>

```
<tagsDecl>
  <tagUsage gi="div" occurs="26"> Utilisé pour
  marquer les séparations alphabétiques du
  dictionnaire. </tagUsage>
  <tagUsage gi="entry" occurs="14526"/>
  <tagUsage gi="orth" occurs="22638"/>
  <tagUsage gi="sense" occurs="8304"/>
  …
</tagsDecl>
```

# Le site de la TEI

http://www.tei-c.org

Le tutoriel fait par ses pères
1ère version 1996
Lou Burnard et C. M. Sperberg-McQueen
Tutoriel