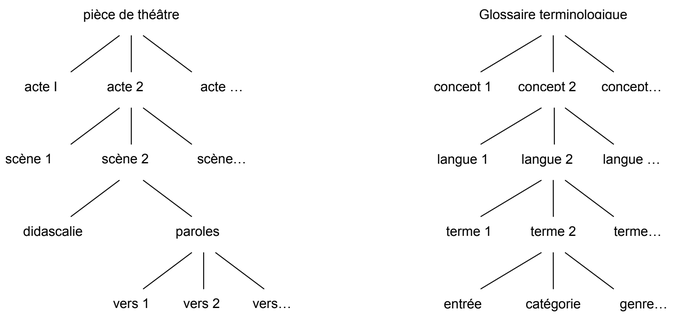
UN DOCUMENT = UN CONTENU STRUCTURÉ



XML = REPRÉSENTATION ARBORESCENTE DE LA STRUCTURE D'UN DOCUMENT

Les données sont identifiées grâce à un emboîtement de balises d'ouverture et de fermeture

```
<pièce>
    <acte>
        <scène>
            <didascalie>…</didascalie>
            <paroles>
                <vers>…</vers>
            </paroles>
        </scène>
    </acte>
</pièce>
```

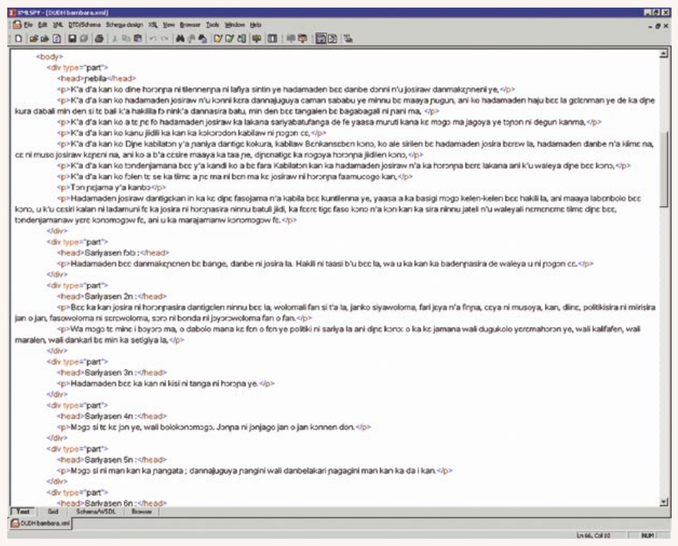
```
<glossaire>
    <concept>
        <langue>
            <terme>
                <entrée>…</entrée>
                <catégorie>…</catégorie>
                <genre>…</genre>
            </terme>
        </langue>
    </concept>
</glossaire >
```

Le modèle de données peut-être spécifié par un formalisme : la description du type de document (DTD)

```
<?xml version="1.0" encoding="UTF-8"?>
<!ELEMENT pièce (acte+)>
<!ELEMENT acte (scène+)>
<!ELEMENT scène (didascalie*, paroles+)>
<!ELEMENT paroles (vers+)>
<!ELEMENT didascalie  (#PCDATA)>
<!ELEMENT vers (#PCDATA)>
```

```
<?xml version="1.0" encoding="UTF-8"?>
<!ELEMENT glossaire (concept+)>
<!ELEMENT concept (langue+)>
<!ELEMENT langue (terme+)>
<!ELEMENT terme (entrée, catégorie, genre)>
<!ELEMENT entrée (#PCDATA)>
<!ELEMENT catégorie (#PCDATA)>
<!ELEMENT genre (#PCDATA)>
```

XML EST COMPATIBLE UNICODE



Encodage en Unicode dans le logiciel XML Spy

XML : UNIVERSALITÉ, RIGUEUR, PÉRENNITÉ ET FACILITÉ D'ÉCHANGE

- Représentation universelle des caractères.
- Informatisation rigoureuse, mais aisée à lire pour l'être humain.
- Pérennité des données, qui – sauvegardées au format « texte seulement » – ne dépendent pas d'un logiciel particulier.
- Facilité des échanges et des transformations de données grâce aux feuilles de style XSL.

XML : UN LANGAGE DE BALISAGE UTILISÉ PAR DES NORMES D'ÉCHANGE DE DONNÉES LINGUISTIQUES

XCES : Corpus Encoding Standard for XML (www.xml-ces.org)

- Adaptation à XML de la norme SGML Corpus Encoding Standard (CES) qui a résulté des projets européens Multext et Eagles.

TEI : Text Encoding Initiative (www.tei-c.org)

- Norme d'échange présentant une large panoplie de balises pour la représentation des données textuelles et linguistiques.
- La version P4 (P4) utilise XML.

TMF : Terminological Markup Framework (www.loria.fr/projets/TMF)

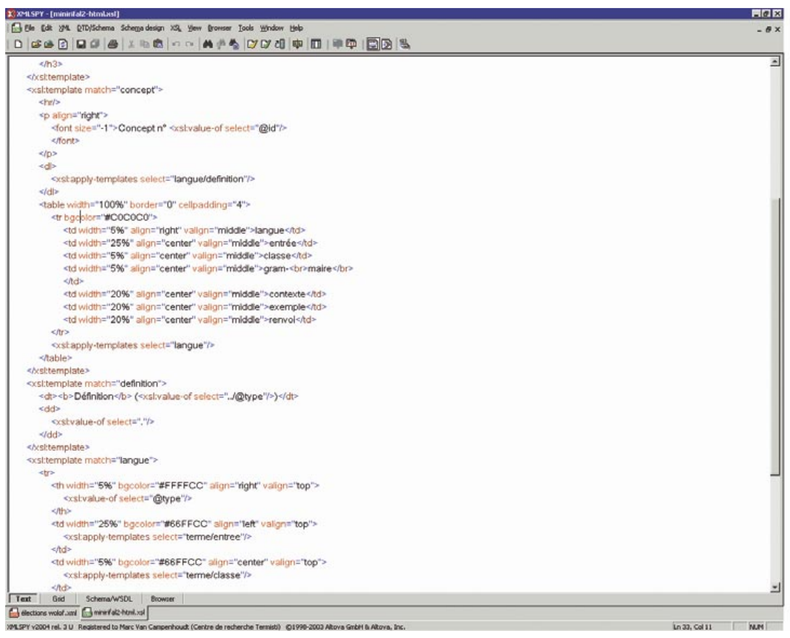
- Norme ISO (12 642) proposant un méta-modèle pour la structuration et l'échange de données terminologiques.

Et beaucoup d'autres :

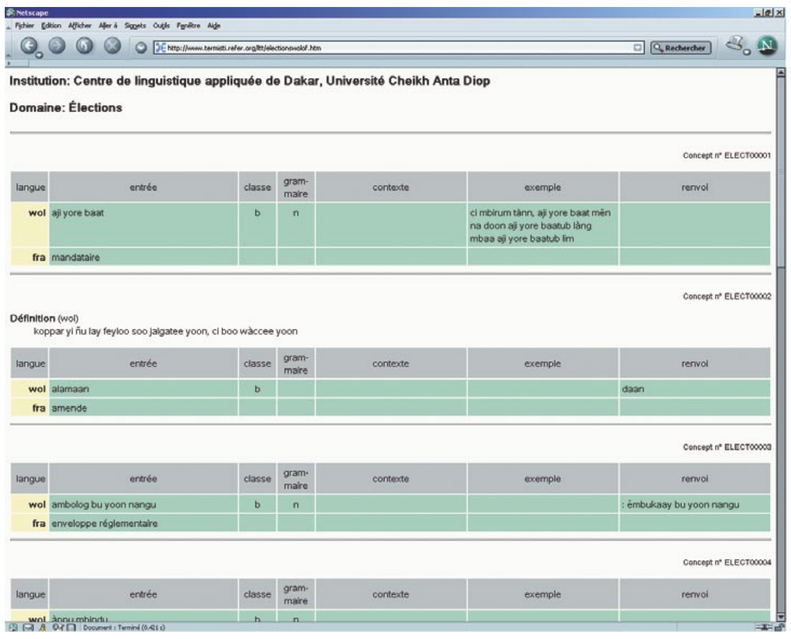
- TMX (Translation Memory Exchange) : www.lisa.org/tmx
- TBX (Termbase Exchange Format) : www.lisa.org/tbx
- TT (Timed-Text) : www.w3.org/AudioVideo/TT
- OLIF (Open Lexicon Interchange Format) : www.olif.net
- SMIL (Synchronized Multimedia Integration Language) : www.w3.org/AudioVideo
- etc.

XSL : TRANSFORMATION DES DONNÉES POUR UNE APPLICATION PRÉCISE (INTERNET, TRAITEMENT DE TEXTE, BASE DE DONNÉES…)

À partir d'un fichier XML bien formé et valide, une feuille de style permet de définir les transformations à effectuer



Après transformation, les données sont présentées au format HTML



EXPÉRIMENTATION DE NORMES DE BALISAGE EN LANGUES PARTENAIRES

Thierno Cisse
Chérif Mbodj
Marc Van Campenhoudt
Mohamédoune Wane

www.termisti.refer.org/ltt/ltt.htm

QUELQUES LOGICIELS 100 % COMPATIBLES UNICODE

Bureautique : *Office 2000sv.* et *XP, Open Office, Star Office*

HTML : *Web Expert 6*

XML : *XML Spy*

Linguistique descriptive : *Toolbox* (ex-*Shoebox*)

PAO : *Adobe InDesign*

Expérimentation de normes de balisage en langues partenaires

Action de recherche en réseau du Réseau Lexicologie, terminologie et traduction de l'Agence universitaire de la francophonie

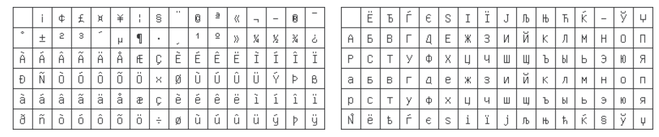
Centre de linguistique appliquée de Dakar (CLAD) et Département de linguistique, Université Cheikh Anta Diop de Dakar.

Département des langues nationales et de linguistique , Université de Nouakchott.

Centre de recherche TERMISTI, Institut supérieur de traducteurs et interprètes, Haute École de Bruxelles.

ANCIENNES POLICES ISO 8859-XX

1 table = 256 caractères ne couvrant que quelques écritures



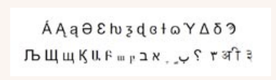
Chaque table ne permet d'écrire qu'un nombre restreint de langues

- Mise en circulation de polices modifiant de manière anarchique les tables pour intégrer les caractères manquants
- Complication de tout échange de données : aucune certitude d'afficher le bon caractère.

UNICODE (ISO-CEI 10 646)

Une table unique inclut toutes les écritures du monde

95 221 caractères codés dans la version 3.2 (870 000 demeurent disponibles…)



Avantages :

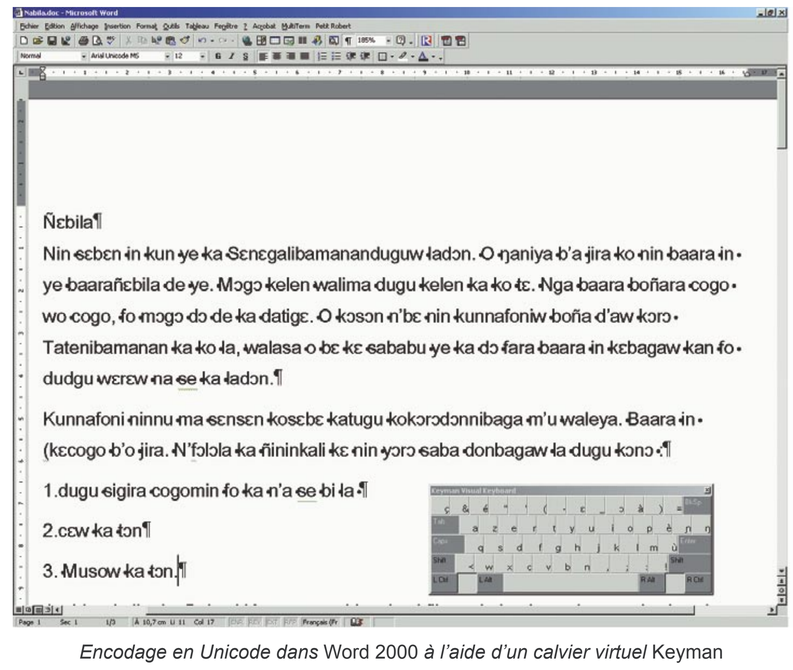
- Un même document (fichier) peut allègrement mélanger les écritures.
- Chaque caractère est codé de manière universelle.
- Certitude d'un affichage correct.

Seuls inconvénients actuels :

- Suppose une maîtrise minimale des réglages d'un ordinateur personnel.
- Fonctionne mal sur les anciens systèmes.
- Tous les logiciels ne sont pas encore compatibles Unicode, notamment les outils de TAO.
- Les besoins des langues partenaires pourraient être mieux satisfaits.

ENCODAGE DES LANGUES PARTENAIRES À L'AIDE D'UN CLAVIER VIRTUEL

S'il n'existe pas de clavier standard pour la langue, on crée un clavier virtuel :



Encodage en Unicode dans Word 2000 *à l'aide d'un calvier virtuel* Keyman

Exigences minimales :

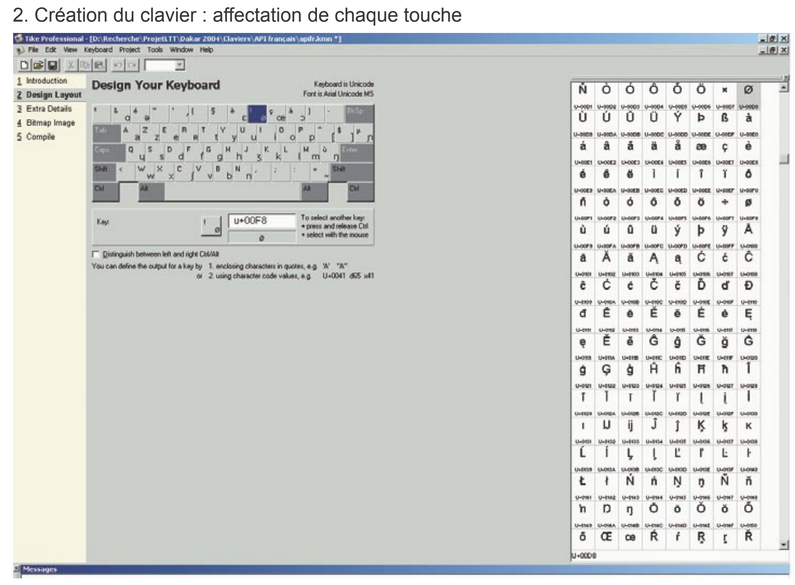
- Compatibilité avec les logiciels Unicode
- Réaffectation des touches du clavier physique.
- Possibilité de visualiser le clavier à l'écran.
- Possibilité de générer une documentation.

Étapes de la création d'un clavier virtuel (logiciel *Keyman*)

1. Inventaire descriptif des caractères Unicode nécessaires

ɲ	Ɲ
nom : lettre minuscule latine N hameçon à gauche bloc : extensions IPA notation Unicode : U+0272 entité : ɲ	nom : lettre majuscule latine N hameçon à gauche bloc : latin étendu B notation Unicode : U+019D entité : Ɲ

2. Création du clavier : affectation de chaque touche



3. Création de la documentation