

IX<sup>e</sup> SOMMET DE LA FRANCOPHONIE, BEYROUTH 2001

Beyrouth 28-29 septembre 2001



## Indexation en français, Indexation en arabe



# Les apports d'UNICODE, HTML4 et Dublin Core pour le traitement du texte multilingue : le cas des langues arabo-latines

Mokhtar BEN HENDA

CEM-GRESIC, Université de Bordeaux 3-FRANCE  
ISD, Université La Manouba-TUNISIE

## Plan

1. Introduction
2. Problématique
3. Définitions des concepts
4. Modes d'application
5. Conclusion

# 1. Introduction

- L'évolution des systèmes d'information :  
Monoposte, LAN, WAN  
→ Systèmes ouverts et distribués  
↓
- Besoin de normalisation pour l'interopérabilité  
des systèmes informatiques (i.e. Protocoles)
- Besoin de normalisation pour le traitement de  
l'information (codage, structuration,  
indexation, échange)
- ... Défi linguistique pour l'information textuelle

# 1. Introduction

- L'information textuelle :
  - ◆ **Codage** : Représentation (codification)  
du contenu
  - ◆ **Structuration** : Formatage du contenu
  - ◆ **Échange** : Indexation/Recherche &  
diffusion/échange du contenu

## 2. Problématique

- Comment gérer l'aspect linguistique intensif (6703 langues) du contenu informationnel sur le réseau ?
- Comment ont été abordés les deux principes de l'Internationalisation (I18n) et de la Localisation (L10n) dans les systèmes et les ressources d'information ?
- Quels sont les apports de l'I18n pour la langue arabe au sein des systèmes d'information multilingues ?
- Qu'en peut tirer l'Indexation en arabe ?

## 3. Définitions « I18n »

- **C'est quoi l'I18n ?**
  - ◆ L'adaptation de tout produit informatique (informationnel, applicatif ou algorithmique) à un maximum possible de langues
  - ◆ Elle est associée à la L10n (i.e. les produits logiciels de Microsoft)
  - ◆ Elle peut s'appliquer entre autres à travers Unicode, HTML et Dublin Core





## 3. Définitions « Unicode »

### ■ Principes clés d'Unicode

- ◆ Coder des caractères et non des glyphes :
  1. Unicode ne définit pas la forme des glyphes d'un caractère mais plutôt son code
  2. Il codifie chaque variante des caractères comme un caractère indépendant
  3. Universalité linguistique
- ◆ Unifier les langues dans un seul Jeu de caractères universels (JCA=UCS)

## 3. Définitions « HTML »

### ■ HTML Multilingue

- ◆ Étiquetage linguistique « **charset** » « **lang** »

```
<html lang=fr>
<head>
<meta http-equiv="content-type" content="text/html; charset=iso-8859-1">
<title>Multimédia Solutions: moteur de recherche Windex</title>
<meta name="author" content="Multimédia SOLUTIONS">
<meta name="description" content="Indexeur de pages HTML">
<meta name="keywords" content="moteur de recherche, moteur">
<meta name="generator" content="Namo WebEditor v3.0">
</head>
<body>
....
</body>
</html>
```

### 3. Définitions « HTML »

- HTML Multilingue

- ◆ Bidirectionnalité « **dir=RTL** »

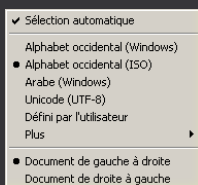
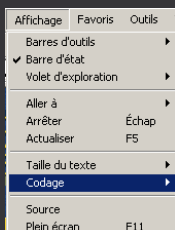
```
<html dir=RTL>
<head>
<meta http-equiv="content-type" content="text/html; charset=utf-8">
....
</head>
<body>
Texte
<p dir=RTL lang=ar>Texte arabe de droite à gauche ...
<p dir=LTR lang=es>Texte espagnol de gauche à droite ...
</body>
</html>
```

### 3. Définitions « HTML »

- Navigateurs multilingues

- ◆ I.Explorer 5.0 +

- ◆ Netscape 4.6 +



Arabe (ASMO 708)
Arabe (DOS)
Arabe (ISO)
Baltes (ISO)
Baltes (Windows)
Alphabet d'Europe centrale (DOS)
Europe centrale (ISO)
Europe centrale (Windows)
Chinois simplifié (GB2312)
Chinois simplifié (HZ)
Chinois traditionnel (Big5)
Cyrillique (DOS)
Cyrillique (ISO)
Cyrillique (KOI8-R)
Cyrillique (KOI8-U)
Cyrillique (Windows)
Grec (ISO)
Grec (Windows)
Hébreu (DOS)
Hébreu (ISO-logique)
Hébreu (ISO-visuel)
Hébreu (Windows)
Japonais (sélection automatique)
Japonais (EUC)
Japonais (Maj-JIS)
Coréen

## Définitions « Dublin Core »

- Unification des modes de description et d'indexation des documents

1. **Titre** : Le nom donné à la ressource.
2. **Auteur ou créateur** : L'entité responsable de la création du contenu de la ressource.
3. **Sujet et mot-clé** : Le sujet du contenu de la ressource.
4. **Description** : Un compte rendu du contenu de la ressource.
5. **Éditeur** : Une entité ayant la responsabilité de rendre la ressource disponible.
6. **Collaborateur** : Une entité ayant la responsabilité de collaborer au contenu de la ressource.
7. **Date** : Une date associée à un événement dans le cycle de vie de la ressource.
8. **Type de ressource** : La nature ou le type de contenu de la ressource.
9. **Format** : Le caractère physique ou numérique de la ressource.
10. **Identificateur de la ressource** : Une référence non ambiguë de la ressource dans un contexte donné.
11. **Source** : Une référence de la ressource d'où est tirée la ressource présente.
12. **Langue** : La langue du contenu intellectuel de la ressource.
13. **Relation** : Une référence à une ressource connexe.
14. **Portée** : L'étendue ou la portée du contenu de la ressource.
15. **Gestion des droits** : Des renseignements au sujet des droits détenus sur une partie de la ressource ou sur son ensemble.

## Définitions « Dublin Core »

- Exemple de DC Spécifique à MetaDubC

```
<HTML LANG=ar-TN>
<HEAD>
<TITLE>Balis
<META HTTP-EQUIV=refresh CONTENT="365" />
<META NAME="generator" CONTENT="HTML à la base" />
<META NAME="description" CONTENT="Balis" />
<META NAME="keywords" CONTENT="Balis" />
<META NAME="author" CONTENT="Mokhtar.Benhenda@isd" />
<META NAME="date" CONTENT="2005-05-12" />
<META NAME="Dublin Core" CONTENT="Dublin Core" />
<META NAME="DC:Title" CONTENT="Balis" />
<META NAME="DC:Description" CONTENT="Balis" />
<!-- Metadata -->
</HEAD>
```

**Owner:** Mokhtar.Benhenda@isd (Email address)

**Expires:** 365 days e.g. +25 days or Sunday, 12-M

**Charset (Recommended if not ISO8859-1)**

**Charset:** ISO-8859-6 (other : )

**Language:** other (other : arabic)

**Dialect (country):** other (other : Tunisia)

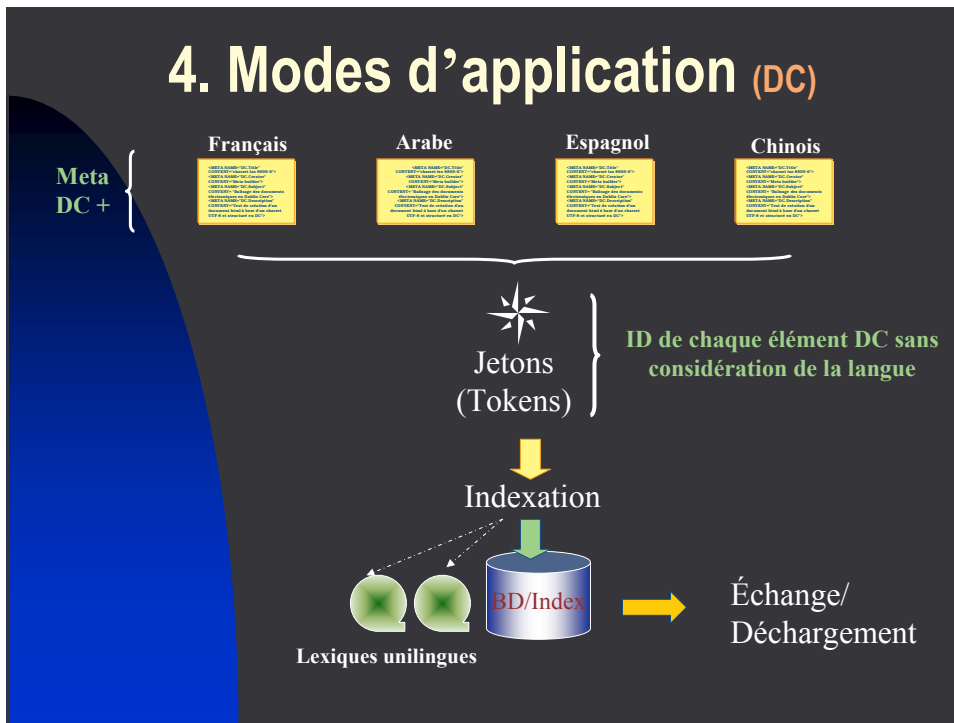
## Définitions « Dublin Core »

- DC Hérite de l'universalité de HTML4 et d'Unicode
- Les versions linguistiques (47 langues)
  - ◆ Source Anglaise et versions linguistiques ou éditions propres ?
  - ◆ Attributs locaux des éléments DC pour des besoins de spécificités
  - ◆ Risque : incohérence dans l'interopérabilité sémantique des champs (interprétations différentes)
  - ◆ Deux courants :
    - ★ Minimalistes : uniformité universelle
    - ★ Structuralistes : spécificités locales

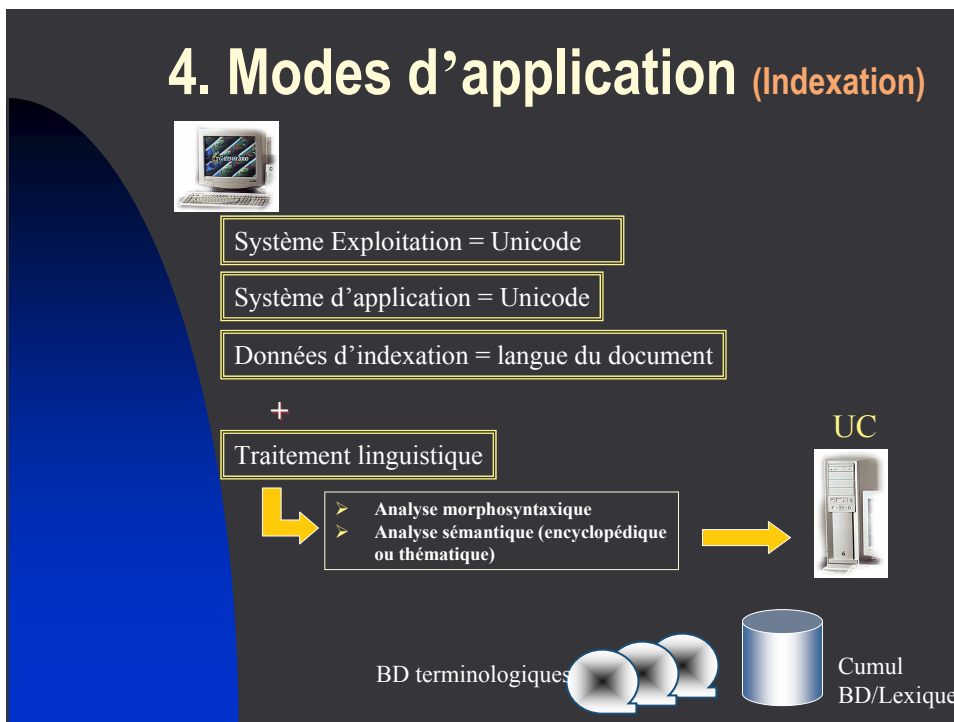
## Définitions « Dublin Core »

- La version arabisée :
  - ◆ Faite par Hichem HADDOUTI (Forwiss, Bavière)
  - ◆ Perspectives :
    - ★ Opter pour des structures DC nationales
    - ★ Optimisation de l'interopérabilité sémantique entre ses structures nationales spécialisées et les partenaires internationaux

## 4. Modes d'application (DC)



## 4. Modes d'application (Indexation)



## 4. Modes d'application (Indexation)

- Et le rôle du professionnel ?
  - ◆ Développement des listes/lexiques
  - ◆ Actualisation des listes
  - ◆ Unification des listes d'indexation
  - ◆ Aide le système à la sélection des termes au moment de l'indexation (validation sémantique)
  - ◆ Échange et la coopération

⇒ L'outil est un appui et non un substitut

## 4. Modes d'application (Recherche)



Système Exploitation = Unicode

Système d'application = Unicode

Données de recherche = langue de l'utilisateur

+

Traitement linguistique



- Analyse morphosyntaxique
- Analyse sémantique (synonymie, homonymie, patronimie ...)
- Lecture des liens sémantiques appropriés



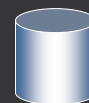
UC



BD terminologiques  
(Saisies / téléchargées)



Thésaurus/  
Lexique



## 4. Modes d'application (Recherche)

- Et le rôle du professionnel ?
    - ◆ L'aide à l'analyse sémantique de la requête
    - ◆ La stratégie de la recherche (Opérateurs)
- ⇒ **L'outil est un appui et non un substitut**

## 5. Conclusion

- Faisabilité : rôle du professionnel



Développeurs de logiciels



Développeurs de contenus



Environnement applicatif



Utilisateur

**Professionnel = Assistant, relais, maître d'œuvres**

- Pousse vers l'évolution et l'ouverture de son système vers les innovations dans le domaine : rôle de veille