

Cour 02

Les standards des humanités numériques

Simon Gabay

Université de Neuchâtel

30 septembre 2019



Les standards des humanités numériques: langages et vocabulaires

Langages informatiques

Langages

Cour 02
Les
standards
des
humanités
numériques

Simon Gabay

Langage



Vocabulaire

Il existe plusieurs langages de programmation

- ▶ langages de programmation
- ▶ langages de définition de données
- ▶ langages de requête
- ▶ langages de balisage

Langage de programmation

- ▶ Permet de formuler des algorithmes ¹ et produire des programmes informatiques qui les appliquent
- ▶ C'est ce qui fait fonctionner l'ordinateur et les logiciels de votre ordinateur
- ▶ Exemples de langage de programmation: C, C++, R, Python, JavaScript
- ▶ Ainsi MacOS ou Linux sont des systèmes UNIX, qui est écrit en C Windows aussi est écrit en C.

¹Un algorithme est une suite finie et non ambiguë d'opérations ou d'instructions permettant de résoudre un problème ou d'obtenir un résultat  

Un document

Cour 02
Les
standards
des
humanités
numériques

Simon Gabay

Langage

Vocabulaire

```
function enEdition(){
    /* Ne rien faire mode edit + preload */
    if( encodeURIComponent(document.location).search(/%26preload%3D/) != -1 ) re
turn;
    // /&preload=/

    if ( !wgPageName.match(/Discussion.*\s/Traduction/) ) return;
    var diff = new Array();
    var status; var pecTraduction; var pecRelecture;
    var avancementTraduction; var avancementRelecture;

    /* ***** Parser ***** */
    var params = document.location.search.substr(1, document.location.search.len
gth).split('&');
    var i = 0;
    var tmp; var name;
    while ( i < params.length )
    {
        tmp = params[i].split('=');
        name = tmp[0];
        switch( name ) {
            case 'status':
                status = tmp[1];
                break;
            case 'pecTraduction':
```

Exemple de code en JavaScript (source: Wikimedia commons)



Définition

Langage de définition de données (*data definition language*, DDL)

- ▶ manipuler les structures de données d'une base de données, et non les données elles-mêmes
- ▶ Dans un tableur (par ex., excel), cela reviendrait à définir le nombre de colonnes et de lignes, ainsi que le le domaine des données ².
- ▶ Exemple: SQL

²valeurs que peut prendre une donnée : nombre, chaîne de caractères, date, booléen.

Définition

Cour 02
Les
standards
des
humanités
numériques

Simon Gabay

Langage

Vocabulaire

```
4 CREATE DATABASE filmographie;
5 USE filmographie;
6
7 -----
8 -- On crée une table de films
9 -----
10 CREATE TABLE films (
11     id          INT(11)          NOT NULL AUTO_INCREMENT,
12     titre       VARCHAR(50)      NOT NULL,
13     sortie      DATE             NOT NULL,
14     PRIMARY KEY (id)
15 );
16
17 -----
18 -- On crée une table de réalisateurs
19 -----
20 CREATE TABLE realisateurs (
21     id          INT(11)          NOT NULL AUTO_INCREMENT,
22     nom         VARCHAR(30)      NOT NULL,
23     film_id     INT(11)          NOT NULL,
24     INDEX (film_id)
25     PRIMARY KEY (id)
26 );
```

Création d'une base de données relationnelle en SQL.

Définition

Cour 02
Les
standards
des
humanités
numériques

Simon Gabay

Langage

Vocabulaire

Films

id	titre	sortie

Réalisateurs

id	nom	filmId

Manipulation

Cour 02
Les
standards
des
humanités
numériques

Simon Gabay

Langage

Vocabulaire

Langage de manipulation de données (*data manipulation language*, DML)

- ▶ permettent de réaliser les traitements sur les données
- ▶ Dans un tableur (par ex., excel), cela reviendrait à remplir le tableau et aller chercher le contenu dans une table
- ▶ Exemple: SQL

Manipulation

Cour 02
Les
standards
des
humanités
numériques

Simon Gabay

Langage

Vocabulaire

```
28 -----
29 -- On remplit la table des films
30 -----
31 INSERT INTO films (titre, sortie)
32   VALUES ('STAR WARS', '1977')
33 INSERT INTO films (titre, sortie)
34   VALUES ('INDIANA JONES', '1981')
35 INSERT INTO films (titre, sortie)
36   VALUES ('TITANIC', '1997')
37
38 -----
39 -- On remplit la table des réalisateurs
40 -----
41 INSERT INTO films (nom, film_id)
42   VALUES ('GEORGE LUCAS', 1)
43 INSERT INTO films (nom, film_id)
44   VALUES ('STEVEN SPIELBERG', 2)
45 INSERT INTO films (nom, film_id)
46   VALUES ('JAMES CAMERON', 3)
47
48 -----
49 -- On fait une requête
50 -----
51 SELECT nom FROM realisateurs
```

Remplissage et recherche dans une base de données relationnelle en SQL.

Définition

Cour 02
Les
standards
des
humanités
numériques

Simon Gabay

Langage

Vocabulaire

Films

id	titre	sortie
1	STAR WARS	1977
2	INDIANA JONES	1981
3	TITANIC	1997

Réalisateurs

id	nom	filmId
1	GEORGE LUCAS	1
2	STEVEN SPIELBERG	2
3	JAMES CAMERON	3

Définition

Cour 02
Les
standards
des
humanités
numériques

Simon Gabay

Langage

Vocabulaire

Films

id	titre	sortie
1	STAR WARS	1977
2	INDIANA JONES	1981
3	TITANIC	1997

Réalisateurs

id	nom	filmId
1	GEORGE LUCAS	1
2	STEVEN SPIELBERG	2
3	JAMES CAMERON	3

Exercice

Définition

```
SELECT titre FROM films
```

Films

id	titre	sortie
1	STAR WARS	1977
2	INDIANA JONES	1981
3	TITANIC	1997

Réalisateurs

id	nom	filmId
1	GEORGE LUCAS	1
2	STEVEN SPIELBERG	2
3	JAMES CAMERON	3

Définition

```
SELECT titre FROM films
```

Films		
id	titre	sortie
1	STAR WARS	1977
2	INDIANA JONES	1981
3	TITANIC	1997

Réalisateurs		
id	nom	filmId
1	GEORGE LUCAS	1
2	STEVEN SPIELBERG	2
3	JAMES CAMERON	3

Langage de de balisage (*Markup language*)

- ▶ Spécialisés dans l'enrichissement d'information textuelle. Ils utilisent des balises, unités syntaxiques délimitant une séquence de caractères ou marquant une position précise à l'intérieur d'un flux de caractères
- ▶ Exemple: HTML, XML ou **LaTeX!** (comme pour ce cours)

Manipulation

Cour 02
Les
standards
des
humanités
numériques

Simon Gabay

Langage

Vocabulaire

```
1 ▾ <doc>
2   <partie>Filmographie</partie>
3   <sous-partie>Films</sous-partie>
4   <contenu>STAR WARS, INDIANA JONES, TITANIC</contenu>
5   <sous-partie>Réalisateurs</sous-partie>
6   <contenu>GEORGE LUCAS, STEVEN SPIELBERG, JAMES CAMERON</contenu>
7 </doc>
```

Exemple de texte balisé en XML

Manipulation

Cour 02
Les
standards
des
humanités
numériques

Simon Gabay

Langage

Vocabulaire

```
1  <!DOCTYPE html>
2  ▼ <html>
3  ▼   <body>
4      <h1>Filmographie</h1>
5      <h2>Films</h2>
6      <p>STAR WARS, INDIANA JONES, TITANIC</p>
7      <h2>Realisateurs</h2>
8      <p>GEORGE LUCAS, STEVEN SPIELBERG, JAMES CAMERON</p>
9   </body>
10 </html>
```

Le même exemple de texte balisé en HTML.

On utilise un vocabulaire précis: `body`, `h1`, `h2`, `p`

Manipulation

Cour 02
Les
standards
des
humanités
numériques

Simon Gabay

Langage

Vocabulaire



Filmographie

Films

STAR WARS, INDIANA JONES, TITANIC

Realisateurs

GEORGE LUCAS, STEVEN SPIELBERG, JAMES CAMERON

Le code HTML dans le navigateur

Vocabulaire: l'exemple de la TEI

Règles principales

Le langage de balisage fonctionne de manière simple

```
<élément attribut="valeur">donnée</élément>
```

- ▶ Un `<élément>` est entre chevrons
- ▶ Une `<balise>` doit être fermé `</balise>`
- ▶ Une `<balise1>` ne doit `<balise2>` pas être croisé `</balise1>` avec un autre `</balise2>`
- ▶ Une `<balise/>` peut être auto-fermante
- ▶ Un `<élément>` peut porter un `@attribut` (noté avec un `@`)
- ▶ l'`@attribut` a une `"valeur"` (entre guillemets)

Sémantique et procédural

Cour 02
Les
standards
des
humanités
numériques

Simon Gabay

Langage

Vocabulaire

On emploie *a priori* les italiques pour les locutions et termes empruntés à d'autres langues.

Procédural

On emploie `<italique>a priori</italique>` les italiques pour les locutions et termes empruntés à d'autres langues.

semantique

On emploie`<locutionEtrangère>a priori</locutionEtrangère>` les italiques...

semantique II

On emploie`<latin>a priori</latin>` les italiques...

Une question fondamentale

Cour 02
Les
standards
des
humanités
numériques

Simon Gabay

Langage

Vocabulaire

Comment choisir le nom des `<éléments>` et des
`@attributs`?

Balisage

Cour 02
Les
standards
des
humanités
numériques

Simon Gabay

Langage

Vocabulaire

Langage de de balisage (*Markup language*)

- ▶ TEI pour *Text Encoding Initiative*
- ▶ Elle est créé en 1987 (donc avant internet)
- ▶ La TEI est pilotée par un consortium qui maintient et développe des recommandations pour l'encodage des textes
- ▶ Ces recommandations sont en constante évolution
- ▶ Elles sont disponibles en ligne
<http://www.tei-c.org/guidelines/>

D'autres vocabulaires que la TEI

- ▶ EAD pour les archivistes
- ▶ Dublin Core (DC) pour les bibliothécaires
- ▶ Ces vocabulaires peuvent être exprimés avec d'autres langages (RDF-DC).
- ▶ Pour cette raison, on parle de XML-TEI, (ainsi il a existé un SGML-TEI).

La solution en TEI

Cour 02
Les
standards
des
humanités
numériques

Simon Gabay

Langage

Vocabulaire

La solution en TEI

On emploie`<foreign xml:lang="la">`a priori`</foreign>` les
italiques...

Manipulation

Cour 02
Les
standards
des
humanités
numériques

Simon Gabay

Langage

Vocabulaire

```
1  <?xml version="1.0" encoding="UTF-8"?>
2  <?xml-model href="http://www.tei-c.org/release/xml/tei/custom/schema/relaxng/tei_all.rng"
3      type="application/xml" schematypens="http://relaxng.org/ns/structure/1.0"?>
4  <TEI xmlns="http://www.tei-c.org/ns/1.0">
5    <teiHeader>
6      <fileDesc>
7        <titleStmt>
8          <title>Exercice pour le cours d'humanités numériques</title>
9        </titleStmt>
10       <publicationStmt>
11         <p>Université de Neuchâtel</p>
12       </publicationStmt>
13       <sourceDesc>
14         <p>Cours original</p>
15       </sourceDesc>
16     </fileDesc>
17   </teiHeader>
18   <text>
19     <body>
20       <head>Filmographie</head>
21       <div>
22         <head>Films</head>
23         <p>STAR WARS, INDIANA JONES, TITANIC</p>
24       </div>
25       <div>
26         <head>Réalisateurs</head>
27         <p>GEORGE LUCAS, STEVEN SPIELBERG, JAMES CAMERON</p>
28       </div>
29     </body>
30   </text>
31 </TEI>
```

Le même code que précédemment en TEI

Valide vs bien formé

Valide (*valid XML document*) vs bien formé (*well-formed XML document*)

- ▶ "Bien formé" renvoie au **langage** et signifie que le document respecte les règles précédemment mentionnées (l'élément est entre chevron, une balise ouverte doit être fermée ...)
- ▶ "Valide" renvoie au **vocabulaire** et signifie que le document répond aux exigences de la TEI (d'où l'expression "valide contre TEI ALL")
- ▶ Un **schéma**, qui est une sorte de dictionnaire qui permet de contrôler que le vocabulaire est bien utilisé, et donc que le document est valide

Valide vs bien formé

Cour 02
Les
standards
des
humanités
numériques

Simon Gabay

Langage

Vocabulaire

Valide vs bien formé

- ▶ L'emploi d'un vocabulaire précis est une restriction du champ des possibles
- ▶ Un document bien formé n'est pas nécessairement valide
- ▶ Un document valide est nécessairement bien formé

Manipulation

Cour 02
Les
standards
des
humanités
numériques

Simon Gabay

Langage

Vocabulaire

```
1 <?xml version="1.0" encoding="UTF-8"?>
2 <?xml-model href="http://www.tei-c.org/release/xml/tei/custom/schema/relaxng/tei_all.rng"
3   type="application/xml" schematypens="http://relaxng.org/ns/structure/1.0"?>
4 <TEI xmlns="http://www.tei-c.org/ns/1.0">
5   <teiHeader>
6     <fileDesc>
7       <titleStmt>
8         <title>Exercice pour le cours d'humanités numériques</title>
9       </titleStmt>
10      <publicationStmt>
11        <p>Université de Neuchâtel</p>
12      </publicationStmt>
13      <sourceDesc>
14        <p>Cours original</p>
15      </sourceDesc>
16    </fileDesc>
17  </teiHeader>
18  <text>
19    <body>
20      <head>Filmographie</head>
21      <div>
22        <head>Films</head>
23        <p>STAR WARS, INDIANA JONES, TITANIC</p>
24      </div>
25      <div>
26        <head>Réalisateurs</head>
27        <p>GEORGE LUCAS, STEVEN SPIELBERG, JAMES CAMERON</p>
28      </div>
29    </body>
30  </text>
31 </TEI>
```

Un schéma permet de contrôler que le code est valide contre la TEI

Défauts de la TEI

Cour 02
Les
standards
des
humanités
numériques

Simon Gabay

Langage

Vocabulaire

La TEI pose des problèmes

- ▶ Elle force à utiliser un standard, par définition générique, et qui ne convient pas exactement à nos données
- ▶ Elle nécessite un apprentissage, notamment pour respecter le sémantisme du vocabulaire

Pourquoi la TEI?

Cour 02
Les
standards
des
humanités
numériques

Simon Gabay

Langage

Vocabulaire

Alors pourquoi la TEI?

- ▶ Elle simplifie l'échange d'information (interopérabilité des données)
- ▶ Elle force à adopter de bonnes pratiques, notamment en ce qui concerne les métadonnées
- ▶ Elle donne accès à une communauté qui donne de l'aide ...
- ▶ ... et qui développe des outils disponibles pour tous!

Manipulation

Cour 02
Les
standards
des
humanités
numériques

Simon Gabay

Langage

Vocabulaire

```
1 <?xml version="1.0" encoding="UTF-8"?>
2 <?xml-model href="http://www.tei-c.org/release/xml/tei/custom/schema/relaxng/tei_all.rng"
3   type="application/xml" schematypens="http://relaxng.org/ns/structure/1.0"?>
4 <TEI xmlns="http://www.tei-c.org/ns/1.0">
5   <teiHeader>
6     <fileDesc>
7       <titleStmt>
8         <title>Exercice pour le cours d'humanités numériques</title>
9       </titleStmt>
10      <publicationStmt>
11        <p>Université de Neuchâtel</p>
12      </publicationStmt>
13      <sourceDesc>
14        <p>Cours original</p>
15      </sourceDesc>
16    </fileDesc>
17  </teiHeader>
18  <text>
19    <body>
20      <head>Filmographie</head>
21      <div>
22        <head>Films</head>
23        <p>STAR WARS, INDIANA JONES, TITANIC</p>
24      </div>
25      <div>
26        <head>Réalisateurs</head>
27        <p>GEORGE LUCAS, STEVEN SPIELBERG, JAMES CAMERON</p>
28      </div>
29    </body>
30  </text>
31 </TEI>
```

Un **<TEIheader>** permet de fournir des métadonnées.