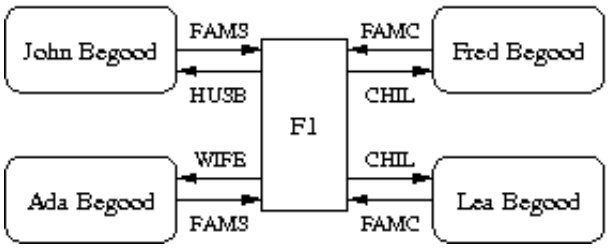
Projet XML

Présentation

Le but de ce projet est d'écrire une DTD, un schéma XML et une feuille de style XSLT pour la traduction en XHTML de documents généalogiques. Les données généalogiques sont fournies par des fichiers Gedcom.

Format Gedcom

Le standard Gedcom (version 5.5.1) est une description d'un format de transmission de fichiers. Les fichiers sont en ASCII ou Unicode. Tous les fichiers d'exemple fournis utilisent le codage Latin-1 (iso-8859-1). Voici quelques exemples de fichiers au format Gedcom.



Liens de la famille Begood

- Un exemple minimaliste constitué d'une seule famille avec les parents et deux enfants. Les liens de parenté sont représentés sur le dessin ci-dessus.
- Un autre petit exemple avec quelques tags supplémentaires pour les dates et lieux de naissance et de décès.
- Un autre exemple un peu plus complet.
- La généalogie de la famille royale d'Angleterre.
- La généalogie de la famille royale de France.

De manière purement syntaxique, le fichier est organisé ligne par ligne mais il décrit des données qui sont structurées de manière hiérarchique. Cette structure hiérarchique est décrite à travers les entiers qui apparaissent au début de chaque ligne du fichier. Les caractères blancs comme les espaces ou les tabulations servent uniquement de séparateurs. Les espaces en début de chaque ligne sont en particulier ignorés. Chaque ligne est constituée des trois éléments suivants.

- d'un entier positif appelé *niveau*
- d'un mot clé appelé *tag*
- d'une valeur qui s'étend jusqu'à la fin de la ligne

dans cet ordre. La valeur peut éventuellement être absente (vide). Une ligne typique a donc la forme suivante.

```
1 NAME John /Begood/
```

où le niveau est `1`, le tag est `NAME` et la valeur est `John /Begood/`. Certaines lignes particulières possèdent un *identificateur* qui est inséré entre le niveau et le tag. Un identificateur est une chaîne de caractères alphabétiques qui commence et se termine avec le caractère arrobas `@`. Une ligne avec un identificateur a donc la forme suivante.

```
0 @I23@ INDI
```

où le niveau est `0`, l'identificateur est `@I23@`, le tag est `INDI` et la valeur est absente.

Les données du fichier sont organisées en *entrées*. Chaque entrée possède un *type*, une *valeur* et des *attributs* (dans la terminologie Gedcom sans lien avec la terminologie XML) qui sont eux-mêmes d'autres entrées. Une entrée est décrite sur plusieurs lignes. Elle commence sur une ligne dont le tag et la valeur donnent respectivement le type et la valeur de l'entrée. Les attributs sont décrits sur les lignes suivantes. Cette description

comprend toutes les lignes suivantes qui ont un niveau strictement supérieur à celui de la première ligne. La description de l'entrée s'arrête donc à la première ligne qui a un niveau inférieur ou égal à celui de l'entrée. Par exemple le fragment de fichier

```
0 @I23@ INDI
  1 NAME John /Begood/
  1 BIRT
    2 DATE 12 nov 1966
    2 PLAC Paris
0 ...
```

décrit une entrée de type INDI. Cette entrée n'a pas de valeur mais elle a deux attributs qui sont des entrées de type NAME et BIRT. L'entrée de type NAME a une valeur qui est John /Begood/ mais n'a pas d'attribut. L'entrée de type BIRT n'a pas de valeur mais elle a deux attributs qui sont des entrées de type DATE et PLAC. Ces deux entrées ont chacune une valeur mais pas d'attribut.

Le format GEDCOM impose que le niveau d'un attribut soit égal au niveau de son entrée augmenté d'une unité. Les entrées de niveau 0 sont appelées *enregistrements*. Seuls, les enregistrements ne sont pas attribut d'une autre entrée.

Un fichier est une séquence d'enregistrements. Chaque enregistrement représente un sommet, c'est-à-dire un individu ou une famille. Il a un identificateur qui est donné sur sa première ligne entre le niveau et le tag. Cet identificateur doit être unique dans toute la base. Le tag est INDI pour les individus et FAM pour les familles. Aucun ordre sur les enregistrements dans le fichier n'est imposé. Les individus et les familles peuvent apparaître dans n'importe quel ordre.

Les informations relatives à un sommet sont représentées par des attributs de niveau 1. Pour les individus, il y a

- Un attribut de type NAME dont la valeur est la suite des prénoms et noms. Le nom est mis entre deux barres obliques comme dans John /Begood/.
- Un attribut de type FAMC dont la valeur est l'identificateur de l'unique famille où l'individu est enfant.
- Des attributs de type FAMS dont leur valeur est l'identificateur d'une famille où l'individu est parent
- Un attribut de type SEX de valeur M ou F.
- Un attribut de type OBJE qui permet d'associer un document multimédia externe.

Pour les familles, il y a

- Un attribut de type HUSB dont la valeur est l'identificateur de l'individu qui est le père.
- Un attribut de type WIFE dont la valeur est l'identificateur de l'individu qui est la mère.
- Des attributs CHIL dont leur valeur est l'identificateur d'un individu qui est issu de cette famille.
- Un attribut de type OBJE qui permet d'associer un document multimédia externe.

Une entrée peut avoir plusieurs attributs de même type. Une famille a par exemple plusieurs attributs de type CHIL si plusieurs individus sont enfants de cette famille. Par contre certains attributs comme ceux de type FAMC ne peuvent être présents qu'une seule fois dans une entrée. Un individu n'est bien entendu issu que d'une seule famille.

La norme GEDCOM est prévue pour des données qui sont souvent partielles. Aucun attribut n'est obligatoire dans une entrée. Par exemple, un enregistrement de type INDI peut très bien ne pas avoir d'attribut de type SEX ou même de type NAME si le sexe ou le nom de l'individu ne sont pas connus. Le format n'impose aucun ordre sur les attributs ou les enregistrements. Même si dans les fichiers d'exemple, les enregistrements de type INDI sont avant les enregistrements de type FAM, ceci n'est pas obligatoire.

L'information est bien entendue redondante, dans la mesure où un individu qui est enfant dans une famille a un attribut de type FAMC qui pointe en retour sur cette famille. De même pour les parents.

Le standard GEDCOM permet d'associer à un individu ou à une famille un document multimédia dans un fichier externe. Pour cela, l'enregistrement de l'individu ou de la famille contient un attribut de type OBJE. Cette entrée contient alors elle-même trois attributs de type FORM, TITL et FILE dont les valeurs donnent respectivement le format du document multimédia, son titre et l'adresse du fichier (URL par exemple).

En plus des individus et des familles, un document Gedcom contient généralement un enregistrement `HEAD` au début et un enregistrement `TRLR` à la fin. Le premier contient des information générales sur le document et le second est vide. Ces deux enregistrements pourront être ignorés dans un premier temps.

Travail demandé

- Écrire une DTD pour des documents XML de données généalogiques provenant de données initialement au format GEDCOM qui est décrit ci-dessus. Il faut ensuite écrire un outil de traduction du format GEDCOM vers le nouveau format XML. Ce programme doit traduire chaque document Gedcom en un document XML valide pour la DTD. Ce programme de traduction doit être le plus simple possible et il peut supposer que le fichier Gedcom fourni est valide. Le langage de programmation est libre. La DTD doit la plus précise possible tout en validant les documents.
- Écrire une schéma XML pour les documents XML de données généalogiques. L'objectif est bien sûr de fournir un schéma plus précis que la DTD. Tous les documents XML produits par votre traducteur doivent être valides pour le schéma. Le schéma doit le plus précis possible tout en validant les documents.
- Écrire des programmes XSLT pour transformer en page XHTML les fichiers XML obtenus à partir des fichiers Gedcom. Les liens entre familles et individus doivent être rendus par des liens hypertextes. Il est aussi demandé d'ajouter un index alphabétique des individus, les listes des enfants et conjoints pour chaque individu.

Les tags suivants doivent absolument être pris en compte dans votre travail : `INDI`, `FAM`, `NAME`, `TITL`, `SEX`, `PLAC`, `DATE`, `DIV`, `BIRT`, `DEAT`, `BURI`, `MARR`, `CHR`, `FAMC`, `FAMC`, `FAMS`, `FAMS`, `HUSB`, `HUSB`, `WIFE` et `CHIL`. Les autres tags peuvent être ignorés dans un premier temps.

Modalités pratiques

Le projet s'effectue en binôme. Il donnera lieu à une soutenance individuelle à la fin du semestre. La note tiendra compte de la lisibilité et de la simplicité des solutions choisies ainsi que de la faculté à répondre aux questions. Bien que les soutenances s'effectuent en binôme, les notes sont individuelles.

Le rendu du projet doit comprendre les éléments suivants

- les sources du programme de traduction pour convertir les fichiers Gedcom en document XML.
- les fichiers `complet.xml` … `royal.xml` contenant la traduction en XML des fichiers Gedcom fournis.
- un fichier `gedcom.dtd` contenant une DTD qui valide les documents `complet.xml` … `royal.xml`.
- un fichier `gedcom.xsd` contenant un schéma qui valide les documents `complet.xml` … `royal.xml`.
- un fichier `xml2html.xsl` contenant une feuille de style XSLT permettant de traduire en XHTML les documents `complet.xml` … `royal.xml`.
- les fichiers `complet.html` … `royal.html` obtenus par application de la feuille de style XSLT `xml2html.xsl`.
- un document PDF contenant un rapport du projet.