Informatisation du Dictionnaire Hydrographique International

Laurent Romary, Patrice Bonhomme

▶ **To cite this version:**

Laurent Romary, Patrice Bonhomme. Informatisation du Dictionnaire Hydrographique International. [Rapport de recherche] 1997. inria-00460457

HAL Id: inria-00460457
https://inria.hal.science/inria-00460457

Submitted on 1 Mar 2010

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Informatisation du

Dictionnaire Hydrographique International

Etude de faisabilité réalisée pour le Service Hydrographique et Océanographique de la Marine

Laurent Romary et Patrice Bonhomme
CRIN-CNRS & INRIA Lorraine
B.P. 239, F-54506 Vandœuvre lès Nancy
romary@loria.fr

0. CONTEXTE DE L'ETUDE 1

1. DESCRIPTION GENERALE DU DHI 1

2. CONTENU DES ENTREES 2

2.1 ENTREES DU VOLUME ANGLAIS 2
2.2 ENTREES DU VOLUME FRANÇAIS 3
2.3 MECANISMES DE POINTAGE 4

3. CODAGE DES ENTREES 4

3.1 VERS UNE PERSPECTIVE « EDITORIALE » 4
3.2 UTILISER LA TEI : POURQUOI ? 5
3.3 STRUCTURE GENERALE DU DICTIONNAIRE 8
3.3.1 L'ENTETE TEI POUR LE DHI 8
3.3.2 REPRESENTATION EN SGML 9
3.4 STRUCTURE GENERALE DES ENTREES 10
3.5 REPRESENTATION DES REFERENCES CROISEES 12
3.6 CHOIX EDITORIAUX A ADOPTER 15
3.6.1 OPERATIONS D'EDITION 15
3.6.2 QUELQUES SOLUTIONS A ENVISAGER 16
3.7 EXTENSIONS POSSIBLES 19

4. ENVIRONNEMENTS D'EDITION ET DE MANIPULATION DE DONNEES SGML 21

4.1 LES DIFFERENTES OPERATIONS ENTRANT DANS LE CYCLE DE VIE D'UN DOCUMENT ELECTRONIQUE 21
4.2 ENVIRONNEMENTS D'EDITION HORS LIGNE 22
4.2.1 LES DTD TEI ET TEI LITE 22
4.2.2 UN EDITEUR SIMPLE, EMACS 22
4.2.3 UN EDITEUR PROFESSIONNEL, AUTHOR/EDITOR 24
4.2.4 GENERATION DE FORMATS RTF ET HTML, JADE 24
4.3 EXPERIMENTATION D'UN OUTIL EN LIGNE 25

4.4 L'AVENIR DE SGML (ALIAS XML) **26**

5. BILAN DE L'ETUDE **27**

ANNEXE 1 - UN EXTRAIT DU DHI CODE EN SGML **29**

ANNEXE 2 - EXTRAIT DU DHI MIS EN PAGE AUTOMATIQUEMENT PAR UNE FEUILLE DE STYLE DSSSL ERREUR ! SIGNET NON DEFINI.

ANNEXE 3 - EXTRAIT D'UNE FEUILLE DE STYLE DSSSL PERMETTANT DE PRODUIRE UN DOCUMENT RTF A PARTIR D'UN FICHIER TEI **37**

0. Contexte de l'étude

L'Organisation Hydrographique Internationale (OHI) publie régulièrement un dictionnaire hydrographique dans ses deux langues officielles (anglais et français) et en espagnol. Ce document, édité jusqu'à présent sous forme « papier », devrait faire l'objet d'un projet de version « électronique ». Le Bureau Hydrographique International (BHI) a demandé aux membres du groupe de travail de l'OHI chargé de l'entretien du dictionnaire d'étudier indépendamment les différentes offres pouvant satisfaire ce projet.

C'est dans ce cadre que le service hydrographique et océanographique de la marine (SHOM) a demandé au CRIN une étude de faisabilité permettant d'identifier les éléments à mettre en œuvre pour ce projet.

Remarque 1 : Les exemples de ce rapport sont extraits de la 5ème édition anglaise et de la 4ème ou de la 5ème édition à venir pour le volume français.

Remarque 2 : du point de vue de la terminologie adoptée, nous avons gardé la notion de volume (pour les différentes langues au sein du DHI) pour désigner les versions électroniques correspondantes.

1. Description générale du DHI

Le dictionnaire hydrographique est publié sous la forme d'un volume indépendant par langue. Chaque volume contient une Préface, une Introduction, ainsi qu'une suite de chapitres correspondant à un découpage alphabétique du dictionnaire. Chacune de ces parties est formée d'une suite d'entrées, sans regroupement particulier par homographe (pas de suite d'entrées correspondant au même terme). Si nécessaire, les entrées sont redécomposées en plusieurs acceptions.

On remarquera que le DHI est d'une complexité similaire à différents projets de dictionnaires informatisés spécialisés que l'on peut trouver actuellement sur le web, par exemple :

- NEVADA DIVISION OF WATER PLANNING, WATER WORDS DICTIONARY, A Compilation of Technical Water, Water Quality, Environmental, and Water-Related Terms
  http://www.state.nv.us/cnr/ndwp/dict-1/waterwds.htm

- DICTIONARY OF ABBREVIATIONS AND ACRONYMS IN GEOGRAPHIC INFORMATION SYSTEMS, CARTOGRAPHY, AND REMOTE SENSING by Philip Hoehn and Mary Larsgaard October 1997 Version 3.0
  http://www.lib.berkeley.edu/EART/abbrev.html
- DICTIONARY OF TECHNICAL TERMS FOR AEROSPACE USE, Web edition edited by Daniel R. Glover, Jr., NASA Lewis Research Center, Cleveland, Ohio
  http://sulu.lerc.nasa.gov/dictionary/intro.html

Il existe par ailleurs un certain nombre de projets beaucoup plus ambitieux (mais dont la mise en œuvre se déroule sur plusieurs années), par exemple :

- l'informatisation du *Trésor de la Langue Française* à l'INaLF-CNRS (Institut National de la Langue Française)
  http://www.ciril.fr/~mastina/TLF
- le *Middle English Dictionary* à l'Université du Michigan
  http://www.hti.umich.edu/dict/med/

2. Contenu des entrées

2.1 Entrées du volume anglais

Le volume anglais du DHI (DHI-EN) s'articule autour de trois champs principaux observés dans leur forme la plus simple dans certaines entrées telles que :

3	**abrasion**. The wearing away or rounding of surfaces by friction.

Le *mot-vedette*, marqué en gras, représente le point d'entrée dans le dictionnaire et sert de base à l'organisation alphabétique du volume. Le mot-vedette, peut être soit simple ('abrasion'), soit composé ('absolute error'). Pour les entrées composées (souvent des structures adjectif+nom), l'ordre des mots peut être conservé ('absolute error') ou inversé pour mettre en tête le deuxième terme comme point d'accès au dictionnaire. Une notation particulière est alors adoptée ('acceleration: angular').

La *définition* suit immédiatement le mot-vedette et se compose d'une ou de plusieurs phrases non structurées autour de rubriques particulières. Les éventuelles indications de domaine ('In astronomy') ou de renvoi ('Also called achromat', 'See abberation of light') sont intégrées dans

la continuité du texte sans marque typographique particulière (mis à part la mise en évidence des références croisées sur lesquelles nous reviendrons).

Un *numéro d'ordre*, en tête de colonne, identifie l'entrée - ou l'acception quand il y a plusieurs sens associés à une entrée - de façon à ce qu'elle puisse être référencée dans les autres volumes du DHI. Ce numéro est susceptible de varier d'une édition du DHI à une autre, en fonction des opérations d'insertion ou de suppression effectuées sur des entrées.

Tant pour le volume anglais que pour le volume français la partie définition de l'entrée peut devenir plus complexe, soit parce que sont décrits plusieurs sens pour le même mot-vedette :

> **Aberration,***f*
> 2**a)**En ASTRONOMIE, […].
> **b)**En optique, […].

soit pour détailler le sens d'une entrée par des indications d'usage :

> 331**Azimut** *m* **géodésique.**
> Angle dièdre entre les demi-plans limités par la normale à l'ELLIPSOÏDE DE RÉFÉRENCE au point d'observation, et contenant respectivement l'axe de l'ellipsoïde, et le point d'observé ou la direction considérée.
> Se compte de 000° à 360° dans le sens des aiguilles d'une montre.

2.2 Entrées du volume français

Les entrées du volume français ont globalement la même structure que celles du volume anglais, avec les différences suivantes :

- le mot-vedette, ou éventuellement la première partie de celui-ci dans le cas de mots composés est systématiquement suivi d'une indication grammaticale[1] (cf. ci-dessus **Azimut** *m* **géodésique**) ;
- le numéro en tête de ligne indique l'équivalent de l'entrée dans le volume anglais (ou d'une sous-acception de l'entrée).

[1] Dans la quatrième édition, on pouvait rencontrer une telle indication (entre parenthèses), notamment quand il existe un risque d'ambiguïté quand à la catégorie morphologique du mot-vedette. Le choix semble avoir été fait de généraliser ces indications pour la cinquième édition.

2.3 Mécanismes de pointage

Par mécanisme de pointage, on signifie ici toute manière de faire référence, à partir d'un mot ou une entrée d'un dictionnaire à une autre entrée décrite ou non dans le même dictionnaire. Le dictionnaire hydrographique réalise différents types de pointages, à savoir :

- des références à partir de mots apparaissant dans certaines entrées (définitions ou indications d'usage) et définis par ailleurs dans le même volume. Ces références sont marquées en lettres capitales dans les volumes anglais et français ;
- des références à des termes dont le sens est lié au mot-vedette de l'entrée courante, toujours dans le même volume. Ces références sont indiquées en capitales, précédées d'une mention explicite (En anglais *See, See also*; En français *Voir, Voir aussi*) ;
- des références à des synonymes, indiquées en italiques et précédées d'une mention (On dit aussi…) ;
- des références du volume français vers le volume anglais, sur la base des numéros d'ordre décrits plus haut.

3. Codage des entrées

3.1 Vers une perspective « éditoriale »

Dans la mise en oeuvre d'une version électronique d'un dictionnaire existant au préalable sous forme papier, il est classique de distinguer trois points de vue pouvant guider ce processus :

- le point de vue *typographique* vise à préserver dans les moindres détails la forme (bidimensionnelle) du document initial (ruptures de page, colonage, marques typographiques etc.) ;
- le point de vue *éditorial* s'intéresse plutôt à la structure linéaire du dictionnaire en identifiant les entrées et les champs qui les composent dans leur ordre d'apparition, ainsi que l'essentiel des marques de ponctuation qui structurent les champs ;
- le point de vue *lexical* se démarque complètement de l'organisation de l'information sur le support papier d'origine pour ne voir dans le dictionnaire qu'une base de données parfaitement structurée ;

Dans le présent rapport, nous envisageons une représentation qui puisse se déduire aisément de la structure actuelle du DHI telle que disponible à partir du traitement de texte (MS Word) qui a

servi à l'éditer, tout en passant au niveau d'abstraction nécessaire pour identifier les champs de façon logique (par exemple distinguer que telle information correspond à une indication grammaticale) par opposition à un simple marquage typographique qui pourrait s'avérer ambigu (mot en italique). C'est pourquoi nous envisageons ici une perspective résolument éditoriale, avec pour conséquences :

- la préservation de l'ordre des entrées et des champs qui les composent dans la représentation informatique ;
- la transformation de toute indication typographique (italique, gras, majuscule) en marquage explicite de la signification de ces indications ;
- le maintien de toutes les marques de ponctuation qui ne peuvent se déduire directement de la structure du dictionnaire informatisé.

3.2 Utiliser la TEI : pourquoi ?

L'informatisation du DHI s'inscrit dans un mouvement général de conversion de documents existant initialement sous forme papier et convertis en un format exploitable électroniquement, afin d'en améliorer l'utilisation et en faciliter éventuellement l'évolution. Il est ainsi clair que d'autres entreprises du même type voient le jour de par le monde et qu'il faut tenir compte des choix qui ont été faits et de l'expérience acquise par d'autres. Du point de vue de l'édition électronique de documents, la norme SGML[2] s'est imposée comme l'une des meilleures solutions pour représenter de l'information structurée, tant au sein des entreprises qu'au niveau académique. C'est ainsi cette même norme qui a servi de base à une réflexion internationale initiée en 1987 dans le cadre de la TEI, *Text Encoding Initiative*. Cette initiative regroupe, sous l'égide des principales sociétés savantes du domaine[3], la plupart des chercheurs de sciences

[2] SGML (Standard Generalized Markup Language) est un standard reconnu par ISO (norme ISO 8879). Dans un document SGML, une balise telle que <entry> représente le début de l'élément, et </entry> en représente la fin. Des couples attribut=valeur, insérés dans la balise ouvrante d'un élément permettent d'apporter des informations supplémentaires au niveau de la représentation considérée (par exemple pour fournir un numéro d'ordre à une suite d'éléments de type <entry>). L'organisation des balises entrent elles est soumise à une syntaxe définie par l'utilisateur (SGML n'est en fait qu'un méta-langage de description de documents) et appelée DTD (Document Type Definition).

[3] ACH (Association for Computers and Humanities), ALLC (Association for Literary and Linguistic Computing), ACL (Association for Computational Linguistics).

humaines (linguistes, philologues, historiens, etc.) ayant à manipuler des informations textuelles sous forme électronique. Ce travail véritablement exemplaire de collaboration internationale a mené à la définition d'un ensemble de directives ("TEI guidelines"), sous la forme d'une part d'une DTD modulaire permettant de traiter différents types de documents (prose, poésie, théâtre, transcription d'échanges oraux, dictionnaires…) et d'autre part une documentation précise publiée en 1992 et disponible directement sur le web.

Dans le cas des dictionnaires, les directives sont relativement complètes et couvrent un large ensemble de phénomènes susceptibles d'être rencontrés dans une activité d'informatisation. Parmi ceux-ci, on peut mentionner :

- la représentation de la structure générale d'un dictionnaire en grandes divisions et entrées, avec de possibles regroupements d'entrées correspondant à des homonymes etc. ;
- la structure interne des entrées en différentes acceptions (éventuellement hiérarchiques) ;
- les différentes informations reliées à une entrée ou une acception tels que les informations morphologiques (forme orthographique, phonétique etc.), grammaticales (catégorie syntaxique, genre, nombre etc.), les définitions, les exemples, l'étymologie, les traductions éventuelles dans d'autres langues, les indications d'usage, les références croisées à d'autres entrées, des notes etc.

Ce cadre général présente par ailleurs une grande souplesse grâce à l'utilisation de différents attributs permettant de mieux cibler la représentation sur les caractéristiques propres d'un dictionnaire donné. A titre d'illustration, nous mentionnons le codage possible de l'entrée "dab" du *Collin's Student Dictionary*. Voici tout d'abord l'entrée telle qu'elle apparaît dans le dictionnaire :

> **dab** /d*!ab/, **dabs, dabbing, dabbed.1.**VB WITH OBJ AND ADJUNCT If you **dab** a substance onto a surface, you put it there with quick, light, strokes. If you **dab** a surface with something, you touch it quickly and lightly with that thing. *She dabbed some powder on her nose. He dabbed the cuts with disinfectant.* **2.** COUNT N A **dab** of something is a small amount of it that is put onto a surface. *She returned wearing a dab of rouge on each cheekbone.* **3.**PHRASE If you are a **dab hand** at something, you are good at doing it; an informal British use.

Et son codage conforme à la DTD de la TEI. On notera en particulier l'indication des variations flexionnelles du verbe et l'indication d'usage du nom. Par ailleurs, on observe que certains champs peuvent être répétés.

```
<entry>
      <form>
            <orth>dab</orth>
            <pro>/d*!ab/</pro>
      </form>
      <form type=infl>
            <orth>dabs</orth>
            <orth>dabbing</orth>
            <orth>dabbed</orth>
      </form>
      <gramGrp>
            <pos>subst. fém.</pos>
      </gramGrp>
      <sense n='1'>
            <usg type=gram>VB with OBJ and ADJUNCT</usg>
            <def>If you dab a substance onto a surface, you put it there with
            quick, light, strokes. If you dab a surface with something, you
            touch it quickly and lightly with that thing.</def>
            <eg>She dabbed some powder on her nose.</eg>
            <eg> He dabbed the cuts with disinfectant.</eg>
      </sense>
      <sense n='2'>
            <usg type=gram>COUNT N</usg>
            <def>A dab of something is a small amount of it that is put onto a
            surface.</def>
            <eg> She returned wearing a dab of rouge on each cheekbone.</eg>
      </sense>
      <sense n='3'>
            <usg type=gram>PHRASE</usg>
            <def> If you are a dab hand at something, you are good at doing
            it; an informal British use.</def>
      </sense>
</entry>
```

3.3 Structure générale du dictionnaire

La TEI structure un document électronique en deux parties principales :

- d'une part un entête contenant l'ensemble des informations permettant de documenter le texte électronique et éventuellement sa source. Cette partie est identifiée à l'aide de l'élément <teiHeader> ;
- d'autre part le contenu informationnel proprement dit, inclus dans l'élément <text>.

3.3.1 L'entête TEI pour le DHI

Tout document TEI doit obligatoirement être précédé d'un entête (élément <teiHeader>) qui contient toutes les informations permettant d'identifier et de décrire le contenu informationnel associé. La DTD de la TEI subdivise cet entête en quatre grandes parties :

- la description du fichier électronique (<fileDesc>), avec son titre, son auteur, les personnes ou institutions responsables de la distribution du document, ainsi que toute information relative à la source bibliographique du document quand celui-ci n'est pas « primaire » (c'est le cas de tout document qui résulte de l'électronisation d'une version papier). C'est aussi dans cette partie (élément <availability>) que peuvent être déclarées les contraintes d'accès et de diffusion du document électronique ;
- une description du contenu informationnel du document (<profileDesc>) ;
- un section éditoriale permettant de préciser les choix de codage spécifiques utilisés pour l'informatisation du document (<encoding Desc>) ;
- l'ensemble des opérations de révision « majeures » subies par le document (<revisionDesc>).

Dans un premier temps, l'entête de la version électronique du DHI peut ne contenir que l'élément <fileDesc> (au demeurant le seul obligatoire), sous une forme pouvant correspondre à l'exemple qui suit et qui pourra être affinée en interaction avec le BHI :

```
<teiheader>
      <filedesc>
            <titlestm>
                  <title>Dictionnaire Hydrographique International</title>
                  <title type="GMD">une version électronique</title>
            </titlestmt>
            <sourcedesc>
                  <biblstruct>
                        <monogr>
```

```
                        <editor>Organisation Hydrographique
                        Internationale</editor>
                        <title>Dictionnaire Hydrographique</title>
                        <edition>5ème édition</edition>
                        <imprint>
                              <publisher>Bureau Hydrographique
                              International</publisher>
                              <pubplace>Monaco</pubplace>
                              <date teiform="date">1997</date>
                        </imprint>
                  </monogr>
            </biblstruct>
      </sourcedesc>
</filedesc>
</teiheader>
```

3.3.2 Représentation en SGML

La version électronique du DHI peut être structurée sur trois niveaux :

1. une organisation générale de l'élément <text> en <front>, contenant la préface et l'introduction, <body>, contenant le corps du dictionnaire et <back>, contenant d'éventuels annexe et index ;
2. un deuxième niveau de découpage de <body> notamment en une suite de divisions (élément) correspondant aux différentes lettres de l'alphabet ;
3. un découpage des divisions sous la forme de la suite des entrées de dictionnaire correspondantes (suite d'éléments <entry>).

```
<text>
      <front>
            <div type="preface">…</div>
            <div type="introduction">…</div>
      </front>
      <body>
            <div>
                  <head>A</head>
                  <entry>…</entry>
                  <entry>…</entry>
                  <!-- … -->
                  <entry>…</entry>
```

```
            </div>
            <div>
            <!-- … -->
            </div>
      </body>
      <back>…</back>
</text>
```

3.4 Structure générale des entrées

Comme nous l'avons vu, une entrée élémentaire (pour le volume français) s'articule autour de la forme orthographique, d'indication grammaticale et d'une définition. La TEI identifie ces différentes informations de façon générale en autorisant l'utilisation d'éléments plus spécifiques à un niveau inférieur. On dispose ainsi des éléments suivants :

- pour décrire la forme de l'entrée, <form>, élément qui peut contenir la représentation orthographique qui nous intéresse ici (élément <orth>), mais qui peut aussi contenir des informations concernant la prononciation par exemple ;
- pour apporter des indications grammaticales, l'élément <gramGrp> va contenir un sous-élément <pos> (pour *part of speech*) indiquant la catégorie grammaticale du mot-vedette ;
- enfin, l'ensemble des informations sémantiques (indications d'usage, définition, exemples etc.) réunies dans un ou plusieurs éléments <sense>, qui pour nous contiendra une définition (<def>) et éventuellement des indications d'usage (<usg>).

A titre d'exemple, voici la forme que peut prendre la représentation de l'entrée « Amarres » du DHI :

```
<entry>
      <xr><xptr doc="SHOM-TEI-EN" from="ID (3322)"></xr>
      <form><orth>Amarres</orth>.</form>
      <gramgrp><pos>f</pos></gramgrp>
      <sense>
            <def>Matériel (câbles, cordages, chaînes, etc.) servant à tenir un
            navire le long d'un quai.</def>
      </sense>
</entry>
```

Dans certains cas, l'entrée peut être subdivisée en différentes acceptions ou sens, éventuellement identifiés à l'aide de l'attribut 'n'. Ainsi, l'entrée « Alidade » aura la représentation suivante :

```
<entry>
```

```
<form><orth>Alidade</orth></form>
<gramgrp><pos>f</pos></gramgrp>
<sense n="a">
<xr><xptr doc="SHOM-TEI-EN" from="ID (84-452)"></xr>
<def>Règle munie d'un dispositif de visée pouvant tourner autour du centre d'un cercle gradué.</def></sense>
<sense n="b"><def>Partie mobile d'un THÉODOLITE.</def></sense>
<sense n="c"><def>Dans les LEVÉS à la PLANCHETTE TOPOGRAPHIQUE, règle munie d'un dispositif de visée et permettant de porter les DIRECTIONS sur la minute.</def></sense>
<sense n="d"><def>Dispositif de visée s'adaptant aux COMPAS et aux RÉPÉTITEURS, muni de PINNULES ou d'une LUNETTE (<ref>alidade à lunette</ref>) pour faciliter la prise des RELÈVEMENTS.</def></sense>
<sense n="e">
<xr><xptr doc="SHOM-TEI-EN" from="ID (2415-2416)"></xr>
<def>Dans un instrument, bras mobile pourvu d'un index et servant à faire des mesures angulaires, comme par exemple l'alidade d'un SEXTANT DE MARINE qui pivote autour du centre du LIMBE et porte le VERNIER ou le MICROMÈTRE, ou encore le bras mobile d'un RAPPORTEUR A ALIDADE.</def></sense>
</entry>
```

Enfin, on peut envisager un découpage plus fin des entrées du DHI par rapport à la version papier actuelle, en identifiant précisément des variantes sémantiques. Ainsi, pour « Accomodation », la définition primaire est suivi d'un usage particulier. Nous suggérons deux entrées de type <sense>, dont l'une intègre, en plus de la définition, une indication d'usage.

> 19 **Accommodation**. *f* Faculté de l'oeuil humain permettant de maintenir une vision nette des objets quelle que soit leur distance. En STEREOSCOPIE, faculté des yeux humains d'obtenir la vision stéréoscopique par superposition de deux images.

```
<entry>
      <form>
            <orth>accommodation</orth>
      </form>
      <gramGrp>
            <pos>f</pos>
      </gramGrp>
      <sense>
            <def>Faculté de l'oeil humain permettant de maintenir une vision nette des objets quelle que soit leur distance.</def>
      </sense>
```

```
      <sense>
            <usg type="domaine">En STEREOSCOPIE</usg>,
            <def>faculté des yeux humains d'obtenir la vision stéréoscopique
            par superposition de deux images.</def>
      </sense>
</entry>
```

3.5 Représentation des références croisées

L'utilisation de SGML et plus particulièrement de la TEI permet d'envisager la représentation de différents types de liens. Le premier mécanisme, qui est une instanciation d'un mécanisme général à SGML, permet d'identifier un pointeur d'un élément vers un autre à l'intérieur du même document sur la base de l'attribut 'id' caractérisant de façon unique un élément dans un document SGML. Dans ce cadre, la TEI utilise les éléments suivants :

<ptr>, un élément vide dont l'attribut 'target' va contenir le pointeur, ou

<ref>, un élément pouvant contenir une description explicite, pointant lui aussi à l'aide de l'attribut target.

Ces éléments sont classiquement utilisés pour des renvois à l'intérieur d'un texte, ainsi :

```
      See especially <ref target="sec12">section 12 on page 34</ref>.
```

ou

```
See especially <ptr target="sec12">.
```

permet de pointer vers une division du même texte déclaré de la façon suivante :

```
<div id="sec12"><head>Les identificateurs…</head>
```

Le deuxième mécanisme introduit dans la TEI permet de pointer sur des parties d'un autre document que celui qui contient le pointeur. Ces références inter-documents s'appuient sur les éléments suivants dans la TEI :

<xptr>, un élément vide dont les attributs 'doc' et 'from' vont respectivement contenir une référence au document dans lequel le pointeur est à interpréter[4], et une formule permettant d'atteindre un élément particulier de ce document.

<xref>, un élément pouvant contenir du texte et reposant sur les même attributs 'doc' et 'from'.

L'attribut 'from' décrit ce que l'on appelle une échelle de positionnement (*location ladder*) reposant sur un ensemble de mots-clef permettant soit des accès directs aux parties d'un

[4] Remarque technique : la valeur de ce document est une entité SGML qui doit être déclarée en début de document.

document (racine du document, élément possédant un 'id' particulier), soit des accès en relatif par la description d'un chemin dans la structure SGML (cf. annexe 2). Sans entrer dans les détails de ces échelles de positionnement, nous pouvons illustrer leur fonctionnement de celles-ci à l'aide d'un exemple.

Supposons qu'un premier document (« doc1 ») contienne une division identifiée comme suit :

```
<div id="sec12"><head>Les identificateurs…</head>
```

Un deuxième document pourra pointer sur celui-ci à l'aide d'un élément <xptr> de la façon suivante :

```
Voir en particulier <xptr doc="doc1" from="ID (SEC12)">…
```

Une autre possibilité peut être d'utiliser une formule plus complexe, ainsi :

```
Voir en particulier <xptr doc="doc1" from="DESCENDANT (2 DIV) (4 P) CHILD (1 QUOTE LANG LAT)">
```

indique un pointeur accédant successivement à la deuxième division, au quatrième paragraphe, puis au premier fils direct de type 'quote' et dont l'attribut 'lang' a pour valeur 'lat' (i.e. la première citation latine).

Les deux mécanismes présentés ci-dessus sont bien adaptés aux différents types de référence que l'on peut rencontrer dans le DHI et plus généralement dans tout dictionnaire. Le premier mécanisme peut ainsi être utilisé pour tout renvoi interne correspondant à des synonymes ou des redirections sur des entrées plus complètes. Ainsi, « Aberration diurne » n'est pas définie en tant que tel, mais pointe sur l'entrée général « Aberration ». On aura donc pour cette dernière la représentation suivante :

```
<entry id="aberration">
      <form><orth>Aberration</orth></form>
      <gramgrp><pos>f</pos></gramgrp>
      <sense n="a"><def>En ASTRONOMIE, […].</def></sense>
      <sense n="b"><def>En optique, […].</def></sense>
</entry>
```

et pour « Aberration diurne » un pointeur sur celle-ci :

```
<entry>
      <form type=part><orth>Aberration</orth></form>
      <gramgrp><pos>f</pos></gramgrp>
      <form type=part><orth>diurne</orth>.</form>
      <sense>
```

```
      <xr><lbl>Voir </lbl><ref target="aberration ">ABERRATION</ref>.</xr>
      </sense>
</entry>
```

On remarquera, que nous avons adopté une représentation plus complète intégrant l'élément <ref> à l'intérieur d'un élément <xr> encadrant plus généralement tout type de renvoi dans un dictionnaire. Ceci permet en particulier de marquer de façon explicite les segments de texte qualifiant le renvoi (e.g. *Voir, Voir aussi…*) à l'aide de l'élément <lbl> (pour *label*, étiquette en anglais). Une telle représentation est susceptible de simplifier les choix de mise en page suivant le format de sortie (e.g. RTF, HTML ou autre) envisagé.

De la même façon, il est possible de coder les renvois synonymiques. Par exemple pour « Abaque » :

```
<entry>
      <form><orth>Abaque</orth>.</form>
      <gramgrp><pos>m</pos></gramgrp>
      <sense>
      <def>Diagramme indiquant […] .</def>
      <xr type="syn"><lbl>On dit aussi </lbl><ref>monogramme</ref>.</xr>
      </sense>
</entry>
```

Dans le cas des renvois sur les équivalents de traduction dans le volume anglais, on utilisera le mécanisme de pointage externe, en supposant que les entrées dans le document anglais ont été correctement identifiées (cf. supra pour les problèmes éditoriaux que cela pose). Ainsi, si nous reprenons l'entrée « Abaque », celle-ci est associée à un équivalent dans le document anglais de la façon suivante (le document est ici référencé par l'indication " SHOM-TEI-EN") :

```
<entry>
      <xr type="trans"><xptr doc="SHOM-TEI-EN" from="ID (3458)"></xr>
      <form><orth>Abaque</orth>.</form>
      <gramgrp><pos>m</pos></gramgrp>
      <sense>
      <def>Diagramme indiquant […] .</def>
      <xr type="syn"><lbl>On dit aussi </lbl><ref>monogramme</ref>.</xr>
      </sense>
</entry>
```

L'attribut 'type' associé à l'élément <xr> permet de différentier les différents pointeurs, notamment dans la perspective de les visualiser de façon différente (impression papier ou visualisation électronique) ou de leur donner un comportement différent (accès électronique).

3.6 Choix éditoriaux à adopter

La mise en place d'un mécanisme de pointage, que celui-ci soit manuel ou automatisé, pose le problème du maintien de la cohérence entre le pointeur et l'objet pointé. Ainsi, le DHI dans sa version papier actuelle ne peut être consulté du français vers l'anglais qu'à la condition de disposer de deux éditions compatibles pour ces deux volumes, faute de quoi les références numériques n'ont plus aucun sens. En effet, toute opération d'édition du volume anglais engendre un décalage des entrées qui conduit *in fine* à une renumérotation complète de l'ouvrage, opération qui doit être accompagnée d'une remise à jour des volumes français et espagnol. De fait, maintenir la cohérence d'un tel système revient à continuellement viser une cible mouvante sans véritable certitude qu'à un moment ou à un autre un lien ne devienne erronée par mégarde au cours de telle ou telle opération d'édition.

Le passage à une version informatisée doit s'accompagner d'une réflexion en profondeur sur les mécanismes qui peuvent résoudre ces problèmes, sans introduire pour autant une charge de travail trop importante ni pour les rédacteurs du dictionnaire, ni pour les utilisateurs qui vont concevoir une version papier à partir de la version électronique ou tout simplement consulter en ligne le dictionnaire. Dans cette section, nous allons dans un premier temps analyser les conséquences de différentes opérations d'édition sur la gestion de la cohérence des pointeurs à l'intérieur du DHI, puis faire différentes propositions techniques et éditoriales pour préserver au mieux l'intégrité du document multilingue.

3.6.1 Opérations d'édition

La structure d'un dictionnaire, et particulièrement du DHI, comme une suite d'entrées organisées « à plat » à l'intérieur de sections alphabétiques immuables fait que les modifications à prendre en compte dans le cadre de la gestion des pointeurs se limitent à celles portant sur les entrées (<entry>) et éventuellement les sous-découpages en acceptions (<sense>). Trois opérations peuvent alors être identifiées, accompagnées des conséquences sur la structure du document :

- **ajout d'une entrée ou d'une acception dans le volume cible (anglais)** : dans le système de numérotation actuel, ceci introduit un décalage dans les numéros d'ordre. Par ailleurs, le terme introduit n'a pas a priori d'équivalent dans les autres langues du DHI ;
- **ajout d'une entrée ou d'une acception dans un volume source (e.g. français ou espagnol)** : cette opération ne modifie bien sûr en rien le système de numérotation, mais introduit un terme sans équivalent dans le volume cible. Si cette opération suit l'ajout d'un terme dans la version anglaise, le pointeur entre les deux doit alors être reconstitué ;
- **modification en profondeur d'une entrée** : il ne s'agit pas là de corrections cosmétiques, mais par exemple de la réécriture d'une définition qui changerait le sens de l'acception. Dans ce cas, bien que le pointeur ne soit pas perdu, il est possible d'insidieusement altérer le lien entre la source (par exemple française) et la cible, car les sens peuvent très bien ne plus être équivalents ;
- **suppression d'une entrée dans le volume cible** : une telle opération a deux conséquences. D'une part elle décale la numérotation des mots-vedette dans le dictionnaire cible, et d'autre part, elle peut entraîner la perte d'un équivalent pour les entrées source qui pointaient éventuellement sur elle. C'est l'une des opérations les plus délicates à réaliser car elle peut laisser des pointeurs en suspens.
- **suppression d'une entrée dans un volume source** : cette opération entraîne la perte éventuelle d'un équivalent dans la langue considérée, à moins qu'il ne s'agisse d'un choix global sur l'ensemble des volumes, choix difficile à gérer au coup par coup (cf. nos propositions ci-dessous).

3.6.2 Quelques solutions à envisager

D'un point de vue technique, nous suggérons d'adopter les choix suivants, afin de faciliter la gestion des pointeurs dans le cadre de représentation que nous avons suggéré à la section 3.5 :

- adoption d'un système de numérotation indépendant du numéro d'ordre des entrées du dictionnaire anglais. Le système que nous proposons se déduit de la forme du mot-vedette en tenant compte du fait que celui-ci soit éventuellement composé (en remplaçant les blancs et les apostrophes de séparation par des signes _) et en supprimant accents et majuscules. Ce système présente l'avantage d'être effectivement unique pour chaque entrée et d'être calculable (ou vérifiable) automatiquement .

Exemples (cf. aussi l'usage dans un pointeur pour l'entrée « Aberration » dans la section 3.5) :

L'entrée « Globe », aura pour clef d'entrée globe ;

« Mouillage de quarantaine », mouillage_de_quarantaine ;

« Télédétection », teledetection.

- préservation de toutes les entrées au cours d'une opération d'édition en faisant correspondre une suppression au marquage de l'entrée correspondante à l'aide de l'attribut "status"[5]. L'attribut "status" d'une entrée donnée aurait ainsi par défaut la valeur "active" et prendrait la valeur "deleted" lors d'une opération de suppression. De la sorte, on préserve la validité de tout pointeur sur l'entrée correspondant, tout en s'autorisant la possibilité de contrôler l'intégrité des différentes entrées dans les différentes langues. Nous suggérons aussi d'utiliser cet attribut pour préserver les anciennes versions (status="old") des entrées du dictionnaire quand des opérations de refonte profonde sont effectuées ;
- utilisation de deux attributs associés à l'élément <entry> pour indiquer la date de création (date-created) et la date de dernière modification (date-modified) afin d'assurer un meilleur suivi éditorial de l'ensemble ;
- mise en place d'un module de vérification automatique de la cohérence des liens au sein du DHI. Ce module serait chargé à la fois de repérer les liens rendus caduques par les opérations de suppression virtuelle, et le repérage des entrées ou acceptions ne possédant pas d'équivalents dans les autres langues du DHI;

D'un point de vue éditorial, les propositions techniques précédentes doivent s'accompagner d'un ensemble de modes opératoires particuliers :

- on s'interdira de modifier telle quelle une entrée du dictionnaire pour en redéfinir le sens en profondeur. Une opération de ce type passera par la création d'une nouvelle entrée et le passage de l'ancienne au statut "deleted" ;
- des sessions de synthèse éditoriale régulière devront avoir lieu entre les différents comités de rédaction associés à chacune des langues du DHI. Ce ne sera qu'à l'occasion de ces sessions

[5] Cet attribut fait déjà partie de la TEI est est utilisé notamment pour indiquer le statut global d'un texte dans l'entête TEI (pour la balise <availability>). Il semble relativement naturel d'en étendre l'usage, à l'aide d'une extension de la DTD TEI, aux besoins que nous exprimons ici.

qu'il sera décidé de faire passer les entrées du statut de "deleted" au statut "old", pouvant lui même conduire à une suppression effective de l'entrée concernée si telle est le choix du comité éditorial (par exemple pour ne pas inutilement alourdir les différentes bases de données ;

- au cours des sessions de synthèse éditoriale, on identifiera les termes dans chacune des langues ne possédant pas d'équivalents (grâce à l'outil mentionné ci-dessus) et l'on choisira ou non d'en générer dans les autres langues. On remarquera ici que techniquement (au sens de SGML) il n'est bien sûr pas indispensable que toutes les entrées possèdent des équivalents de traduction.

Remarque : bien que dans un premier temps nous nous soyons appuyé sur l'hypothèse d'une rétroconversion *a minima* du DHI en gardant le principe d'un pointage de tous les volumes source vers le seul volume cible représenté par l'anglais, nous suggérons d'envisager à terme d'aller vers une plus grande indépendance éditoriale des différents volumes du dictionnaire, tout en augmentant les possibilités de parcours en introduisant un mécanisme de double pointage par paires de langues. Il serait ainsi tout aussi facile de passer d'une entrée en anglais à un équivalent français que l'inverse.

Concrètement, on peut garder la structure des entrées actuelles en ajoutant simplement à chaque entrée du dictionnaire anglais les pointeurs vers les équivalents dans les autres langues. Ainsi, si l'entrée française à la forme suivante (on adopte ici le système d'identification suggéré plus haut) :

```
<entry id="abaque">
      <xr type="trans"><xptr doc="SHOM-TEI-EN" from="ID (nomogram)"></xr>
      <form><orth>Abaque</orth></form>
      <gramgrp><pos>m</pos></gramgrp>
      <sense>
      <def>Diagramme indiquant […] .</def>
      <xr type="syn"><lbl>On dit aussi </lbl><ref>monogramme</ref>.</xr>
      </sense>
</entry>
```

L'entrée anglaise aurait une structure similaire pointant vers l'entrée française :

```
<entry id="nomogram">
      <xr type="trans"><xptr doc="SHOM-TEI-FR" from="ID (abaque)"></xr>
```

```
      <form><orth>nomogram</orth></form>
      <sense>
      <def>A DIAGRAM showing, […] .</def>
      </sense>
</entry>
```

On remarquera que le pointage de l'anglais vers les autres langues pourra être calculé automatiquement à partir des pointeurs directs.

3.7 Extensions possibles

Bien que la proposition de codage faite dans les sections précédentes recouvre au plus près les informations contenues dans les éditions actuelles du DHI (l'objectif premier étant d'assurer une rétroconversion intégrale), l'utilisation de SGML dans le cadre des directives de la TEI peut permettre à terme d'étendre les fonctionnalités de l'ouvrage dans sa version électronique par l'ajout de différents champs (ou éléments au sens SGML du terme) à l'intérieur ou en complément des champs existants.

- indications des sources bibliographiques en cas d'emprunts à d'autres ouvrages. Cette situation est fréquente dans le cas du DHI qui résulte souvent d'un travail de compilation à partir de différentes sources spécialisées. On peut envisager un codage assez fin de ces sources en utilisant les éléments proposés par la TEI. L'exemple qui suit (extrait de l'entrée quarantaine du *Trésor de la Langue Française*) illustre ainsi la mention d'une source (élément <bibl>) à l'intérieur d'un exemple (élément <eg>) associé à l'entrée *quarantaine* :

```
<eg>
      <cit>
            <q>Innocent IV (...) accorda un an et quarante jours
            d'indulgence (...). Sixte IV (...) cinquante années et
            autant de quarantaines d'indulgence à tous les fidèles (...)
            qui visiteraient les églises de l'ordre de saint
            François</q>
            <bibl> (<author>Montalembert</author>, <title>Ste
            élisabeth</title>, <date>1836</date>, <biblScope>p.
            321</biblScope>)</bibl>.
      </cit>
</eg>
```

- insertion de notes éditoriales. Suite au travail des comités éditoriaux, il est parfois nécessaire de mémoriser certaines informations spécifiques (ambiguïtés, incertitudes, propositions

d'évolution ou de nouvelles entrées). L'élément <note> peut être utilisé à cet effet et filtré avant toute opération de présentation ;

- indication systématique du domaine d'usage pour chacune des entrées ou des acceptions. Les différentes discussions que nous avons eues avec le SHOM montrent que de nombreux termes relèvent de domaines ou sous-domaines particuliers tels que la météorologie, l'optique ou la mécanique des fluides qui sont souvent associés à l'hydrographie par nature. L'utilisation systématique de l'élément <usg type='dom'> en tête d'entrée permettrait de filtrer ces « emprunts » pour par exemple mieux gérer la cohérence éditoriale du dictionnaire en le confrontant à des bases terminologiques spécialisées (cf. les ouvrages du même type trouvés sur le web et mentionnés dans la section 1.) ;
- indications de prononciation. En plus de la forme orthographique, de telles indications peuvent être importantes lorsque le dictionnaire multilingue est utilisé dans le cadre d'échanges internationaux. Il suffit d'ajouter un élément <pron> à l'intérieur des indications morphologiques (<form>) ;
- pointage sur diverses sources textuelles. Le dictionnaire actuel ne contient quasiment aucun exemple d'emploi des mots. Une manière d'ajouter une telle information serait de pointer sur des textes informatisés qui seraient associés au DHI informatisé (on pense en particulier aux publications du BHI qui à terme pourraient être elles aussi normalisées sous forme électronique).

Comme la forme électronique est indépendante du format de représentation visé (cf. section 4), ce n'est pas parce qu'on étend la quantité d'information contenue à l'intérieur du dictionnaire informatisé que l'on va alourdir pour autant l'aspect de la version papier qui pourra en être tirée ou la navigation sur une éventuelle version mise sur le web. Ainsi, les notes éditoriales insérées dans le dictionnaire pourront rester à usage strictement interne aux comités de rédaction du DHI et ne jamais apparaître dans aucune version publique. De la même façon, un outil automatique de navigation du DHI pourra proposer différents niveaux de précisions suivant ce que cherche un utilisateur, par exemple un niveau simple d'une part ne présentant que le mot et ses équivalents dans les autres langues, et d'autre part un niveau plus élaboré faisant effectivement apparaître les définitions.

Enfin, au delà des quelques propositions que nous avons faites ci-dessus, les éditeurs futurs pourront toujours choisir d'ajouter d'autres éléments au format que nous avons suggéré, soit en

reprenant des possibilités déjà offertes dans le cadre de la *Text Encoding Initiative*, soit en ajoutant (la TEI prévoit ce mécanisme) d'autres éléments tout à fait spécifiques au contexte d'utilisation du DHI.

4. Environnements d'édition et de manipulation de données SGML

4.1 Les différentes opérations entrant dans le cycle de vie d'un document électronique

Dans le cadre de représentation que nous avons suggéré, différentes étapes sont à envisager :

1. rétroconversion du dictionnaire existant, à partir des documents en format Microsoft Word qui sont disponibles pour les volumes français et anglais. Cette opération n'est à réaliser qu'une seule fois à partir du moment où il est décidé que toutes les opérations d'édition ultérieures se font sur la base de la représentation en SGML. Elle peut ainsi être coordonnée par le CRIN-CNRS & INRIA Lorraine en liaison étroite avec les centres responsables des différents volumes (le SHOM pour le français), et sur la base d'outils d'usage courant faciles à installer (par exemple l'éditeur EMACS en mode SGML) ;
2. mise en place d'un environnement d'édition « professionnel » au niveau des différents comités éditoriaux. Il s'agit de définir un environnement convivial qui permette aux comités éditoriaux de faire évoluer le DHI de façon autonome et sans nécessairement posséder de connaissances précises sur SGML et le format sous-jacent adopté. Nous suggérons l'utilisation d'un environnement construit autour d'un produit tel que Author/Editor de la société Softquad ;
3. réalisation de différents filtres permettant de générer entre autre des versions papier à partir du format SGML. De telles filtres peuvent être réalisés à l'aide de feuilles de style au format DSSSL reconnu par l'outil Jade présenté plus loin. Ils permettraient en particulier de générer automatiquement une version qui puisse être relue sur un traitement de texte courant (format RTF reconnu par Microsoft Word), ou encore une version HTML simple pouvant servir de base à l'édition d'un CDROM. C'est à l'aide d'une feuille de style expérimentale de ce type qu'a été généré l'extrait mis en page en annexe 2 ;
4. intégration des versions électroniques sur un serveur web accessible, soit de façon simplifiée sous la forme d'un ou plusieurs documents html, soit via un outil plus complet implanté à cet effet.

Bien que toutes les expérimentations que nous avons menées pour réaliser cette étude aient été faites en utilisant la DTD complète de la TEI (ce qui nous a donné en particulier toute la

souplesse requise pour des tests de rétroconversion), il sera nécessaire de définir une DTD simplifiée pour la mise en place de l'environnement d'édition finale. En effet, une telle DTD permettra de bien contraindre et cadrer le travail d'édition n'autorisant qu'un minimum d'options à un moment donné. En effet, les utilisateurs non spécialistes en informatique ne doivent pas être bloqués par un environnement qui deviendrait trop complexe car trop ouvert. Une telle DTD simplifiée pourrait d'une part être testée au cours de l'opération de rétroconversion pour s'assurer qu'elle couvre bien les données existantes, et d'autre part intégrer certaines propositions d'ajouts d'informations supplémentaires que suggérerait le BHI suite à la présente étude.

Dans les sections qui suivent, nous présentons succinctement les outils pouvant servir de base à l'opérationalisation de notre étude de faisabilité. Tous ces outils, pour lesquels nous indiquons des pointeurs html officiels, sont disponibles au minimum sous Unix et Windows 95/NT, et souvent sur Macintosh.

4.2 Environnements d'édition hors ligne

4.2.1 Les DTD TEI et TEI Lite

Toute la documentation, les DTD, ainsi que de nombreux exemples sont disponibles gratuitement sur le web. Il existe par ailleurs une liste de discussion (TEI-L) très active où participent l'ensemble des acteurs de la communauté TEI.

La TEI est utilisable sous deux formes. La TEI P3 est la DTD complète pour coder tout type de ressource linguistique (ou presque) notamment les dictionnaires mono et multilingues. Il existe une version simplifiée de cette DTD : la TEI Lite, qui ne contient pas de module pour représenter les dictionnaires, mais qu'il peut être intéressant de consulter pour quelqu'un souhaitant normaliser la représentation de textes électroniques (par exemple pour les publications du BHI).

TEI P3
    http://etext.virginia.edu/TEI.html
    ftp://ftp-tei.uic.edu/pub/tei/dtd/
TEI Lite
    http://www.uic.edu/orgs/tei/lite/

4.2.2 Un éditeur simple, EMACS

Emacs est un éditeur disponible sur quasiment toutes les plate-forme informatiques actuellement et qui est l'outil de base de nombreux spécialistes, notamment pour programmer. C'est un outil très riche intégrant par exemple des fonctions élaborées de recherche et de remplacement utilisant

des expressions régulières. Par ailleurs, sa popularité a conduit au développement d'un environnement complet pour éditer des documents SGML en respectant les contraintes de la DTD[6]. Cet environnement repose sur le mode "PSGML" et utilise le parseur[7] "nsgmls" (cf. ci-dessous). C'est typiquement un environnement de ce type qui permet de contrôler de façon convenable une opération de rétroconversion d'un dictionnaire tel que le DHI.

Xemacs (19.15 et >) pour Unix :

http://www.xemacs.org/

NTEmacs (19.34 et >) pour Windows 95/NT :

http://www.cs.washington.edu/homes/voelker/ntemacs.html

Psgml (1.0.1)

Psgml est un mode majeur SGML pour éditeur Emacs. Psgml contient un parseur simple de DTD et peut fonctionner avec n'importe quelle DTD. Les fonctions fournies, incluent :

- les menus et les commandes pour l'insertion des balises à partir d'une liste des éléments SGML valides dans le contexte,
- l'identification des erreurs liées à la structure,
- l'édition des valeurs des attributs dans une boite de dialogue séparée contenant les informations sur les types et les valeurs,
- une édition basée sur la structure SGML en non WYSIWYG.

Il est facilement configurable est permet, par exemple, l'ajout de fonctionnalités supplémentaires pour le menu principal ou les menus contextuels.

Remarque :

Le package psgml (1.0) est intégré à la distribution de xemacs-19.16.

[6] Par exemple, l'insertion d'une balise est contrainte par le contexte des éléments déjà introduits. Cependant, l'environnement ne présente qu'une seule vue à l'ustilisateur, à savoir l'ensemble du document avec ses balises.

[7] Un *parseur* (ou "parser" en anglais) est un logiciel qui lit un document SGML et en valide la structure, c'est à dire s'assure de la conformité de ce document vis à vis de la DTD, interprète les entités, vérifie la cohérence des pointages de type ID/IDREF et produit une sortie normalisée qui peut être utilisée par d'autres outils ayant moins de fonctionnalités de vérification.

SP (1.2)

http://www.jclark.com/sp/

Psgml possède un mini parseur SGML mais pour une ultime vérification d'un document SGML, il est nécessaire d'utiliser un véritable parseur 100% SGML. Le parseur le plus proche de la norme SGML et le plus utilisé dans le monde de l'édition est **nsgmls** livré avec le package de *James Clark*, **sp**. Il est possible d'utiliser ce parseur directement depuis le mode **Psgml**.
nsgmls parse et valide un document SGML en fonction de la DTD et fournit une liste d'erreur précise en indiquant le type, la ligne et la colonne de l'erreur.

4.2.3 Un éditeur professionnel, Author/Editor

Author/Editor est un véritable éditeur professionnel intégrant toutes les fonctionnalités d'un traitement de texte, mais travaillant exclusivement sur des documents au format SGML en contrôlant l'édition à partir de la DTD qui lui est fourni, et présentant des vues simplifiées à l'utilisateur à l'aide de feuilles de style qui doivent être redéfinies pour chaque DTD. Une fois un environnement d'édition défini, il est possible par exemple de ne mettre à disposition d'un utilisateur final qu'une représentation du dictionnaire où celui-ci va éditer une à une les différentes entrées.
http://www.softquad.co.uk/
Pour Windows (version 3.5) ou Mac (version 3.1)

$995 US

$1195 CDN

Pour Unix (version 3.1)

$1495 US

$1795 CDN

4.2.4 Génération de formats RTF et HTML, Jade

La norme SGML est supposée être logique et non *visuelle*. **Jade** permet d'effectuer des transformations et d'appliquer une feuille de style spécifique à la DTD du document traité afin d'obtenir des publications sous des formats différents et imprimables. **Jade** utilise un document SGML en entrée et une spécification DSSSL pour fournir les formats de sortie suivant :

• RTF (Rich Text Format) qui peut être lu par les principaux traitements de texte (Word, Framemaker etc.);

• TeX, format de mise en page utilisé dans le monde Unix et pouvant fournir une version PostScript (directement imprimable) du document;

• un autre document SGML avec une DTD différente (HTML par exemple).

Jade (1.0)

http://www.jclark.com/jade/

4.3 Expérimentation d'un outil en ligne

La génération d'une simple page html à partir du dictionnaire SGML, même si elle présente un avantage de simplicité, ne permet pas une réelle navigation au sein de l'ouvrage en fonction des souhaits de l'utilisateur à un instant spécifique de sa consultation. C'est pour cette raison que nous avons expérimenté au cours de l'été 1997 le développement d'un outil de consultation de dictionnaire qui permettent de faire des requêtes, via le web, sur un dictionnaire tel que le DHII (DHI Informatisé) qui serait mis au format SGML/TEI et placé sur un serveur.

L'outil préliminaire que nous avons implanté possède les caractéristiques suivantes :

- l'outil a été programmé entièrement en Java dans une configuration faisant communiquer ce que l'on nomme une Applet (partie exécutée sur le poste de celui qui consulte le dictionnaire) et une Servlet (partie exécutée sur le serveur) ;
- la partie Servlet repose sur un analyseur de structures XML (proche de SGML cf. infra) disponible gratuitement sur le web et implanté par Microsoft ;
- l'utilisateur sélectionne un mot et demande ses caractéristiques. L'affichage de celle-ci est paramétrisable à l'aide d'une feuille de style.

Le bilan de cette expérience montre que le développement d'un tel outil est tout à fait réalisable dans le cadre des technologies accessibles actuellement. Le développement d'une version plus complète (incluant en particulier le parcours de liens dans le dictionnaire) permettrait en particulier d'envisager une interface qui puisse être la même entre une version accessible via le web et une version que l'on placerait comme outil de consultation sur un CDROM. Par rapport à une page html, un tel outil apporterait un meilleur contrôle des accès et une bonne évolutivité si on souhaite ajouter des fonctionnalités. Ainsi, cet outil pourrait faciliter le travail éditorial par consultation directe, par les personnes habilitées, de l'état des différents volumes, incluant les

entrées marquées comme étant à détruire et les éventuelles notes attachées. Pour les autres personnes qui consulteraient le dictionnaire, seules les informations valides seraient présentées.

4.4 L'avenir de SGML (alias XML)

A l'heure où de nombreuses évolutions viennent constamment perturber le paysage informatique dans le domaine des normes de représentation et des langages informatiques, il peut être important de faire le point sur la durée de validité des propositions faites dans se rapport. En particulier, le consortium W3[8] a récemment (mars 1997) émis différentes propositions autour d'un « nouveau » langage de représentation et d'échange de documents structurés sur l'Internet: XML. Ce format devrait, aux dires de Jan Bozac lors de la conférence TEI'10 qui s'est tenue à Providence en novembre 1997, s'imposer comme le support principal des informations accessibles sur Internet et pallier en particulier les limitations de HTML.

Il se trouve que XML est directement dérivé de SGML, puisqu'il en reprend l'essentiel de la syntaxe, en simplifiant celle-ci en différents points rendant son implantation plus aisée. Par ailleurs, la définition du langage XML s'accompagne d'un travail en profondeur d'une part sur la définition d'un mécanisme de liens entre documents inspiré des pointeurs de la TEI et d'autre part sur la mise en place d'une notion de feuille de style issue de la norme DSSSL. Au bilan, il s'avère que ces évolutions devraient avoir un effet très bénéfique pour tous les projets qui jusqu'ici se sont appuyés et s'appuient encore sur le standard SGML. Bien plus, la présence au sein des groupes de travail sur XML de personnes fortement impliquées dans la TEI telles que Michael Sperberg-McQueen (co-éditeur de la TEI) ou Steve de Rose ont permis que les mécanismes de pointages sur lesquels nous nous sommes appuyés dans ce rapport se généralisent sur le web, et soient donc à terme interprétables par tout navigateur du type Netscape ou Microsoft Internet Explorer.

Conrètement, l'usage qui est fait de SGML par la plupart des projets rend les documents parfaitement compatibles avec la norme XML telle qu'elle se présente actuellement. Le travail de rétroconversion et l'environnement de travail que nous proposons en bilan de cette étude (cf. section 5.) se situe donc dans un cadre qui semble garantir une parfaite pérennité aux données du

[8] Ce consortium réunit la plupart des acteurs industriels (Sun, Netscape, Microsoft etc.) et universitaires (MIT, INRIA) dans le domaine de l'Internet. Il vise à définir des standards concernant les documents et les programmes transitant sur les réseaux internationaux (cf. http://www.w3.org/)

DHI, ainsi qu'à terme une large flexibilité d'utilisation et de distribution. A terme tous les documents techniques ont vocation à rester des documents SGML, sachant que tous les outils d'édition sauront passer de l'un à l'autre sans difficulté (ainsi les versions à venir d'Author/Editor intégrerons les deux fonctionnalités, de même, la dernière version de nsgmls intègre - de façon transparente - la lecture des deux types de fichier). Pour reprendre les termes de Jan Bozac lors de la conférence TEI'10, nous pouvons et nous devons continuer à travailler dans le cadre que nous nous sommes fixés.

5. Bilan de l'étude

Au terme de cet étude, il nous semble tout à fait possible de passer en un temps raisonnable de la version papier du DHI telle qu'elle existe actuellement en anglais et en français à une version électronique qui couvre l'ensemble des informations existantes actuellement. Bien plus, nous avons vu que le format proposé permet d'aboutir à une vision différente, plus souple et évolutive du dictionnaire, car son électronisation rendra possible une mise à jour continuelle et raisonnée de ses entrées. A tout moment, il sera possible malgré tout de figer des « éditions » spécifiques lorsque l'on souhaitera publier une version papier ou réaliser un CDROM à partir de la version informatisée. Enfin, l'utilisation d'un format de représentation qui soit compatible avec les standards internationaux adoptés à la fois par les éditeurs professionnels et la communauté académique garantit la pérennité des informations ainsi codées. A terme, nous suggérons que l'expérience d'informatisation du DHI se généralise à l'ensemble des publications du BHI afin d'en assurer une meilleure gestion et distribution.

Pour l'heure, nous récapitulons ici l'ensemble des étapes et/ou opérations à prendre en compte si le BHI décidait de se lancer effectivement dans la réalisation d'un DHII. Pour chacune de ces étapes, nous avons estimé le coût en hommes-mois et au final une estimation du coût total, dans l'hypothèse que seuls les volumes français et anglais seraient concernés dans un premier temps. Il est évident que tous les autres volumes existants (e.g. espagnol) ou à venir bénéficieraient de l'expérience acquise sur les deux langues officielles de l'OHI et pourraient en particulier s'appuyer immédiatement sur les mêmes environnements d'édition.

Les étapes qui suivent ont été approximativement évaluées en hommes-mois, sachant que suivant la complexité du travail, la tâche peut être confiée à des personnes plus ou moins qualifiées. L'ensemble du travail correspondrait à première vue à un budget de 250KF sur 2 ans maximum.

Rétroconversion des volumes français et anglais du DHI (total **8 hm**)

Rétroconversion semi-automatique (1 hm par volume)

Relecture complète des entrées, pour en corriger le formatage SGML (3 hm par volume, fait par des stagiaires de licence/maîtrise Industries de Langue connaissant SGML)

Définition d'une DTD simplifiée pour le DHI, validation sur les données rétroconverties (DTD "DHI") (**1hm**)

Mise en place d'un environnement d'édition du DHI électronique (total **4 hm**)

Choix et test d'un éditeur de documents SGML (Author/Editor de la société Softquad), intégration de la DTD DHI (1 hm)

Définition d'une feuille de style pour la DTD DHI (1 hm)

Production d'un manuel d'utilisation simplifié (1 hm)

Installation et test de l'environnement sur site (1 hm)

Implantation d'un module de validation des liens au sein du DHI (**3 hm**)

Détection des liens absents ou pointant vers des entrées ayant le statut "deleted"

Repérage des entrées ne possédant pas d'équivalents dans les autres langues

Définition d'une feuille de style DSSSL permettant la génération de documents formatés pour l'impression d'une version papier et pour la génération d'une version HTML (**1 hm**)

Définition d'un outil d'accès au DHI via Internet (**6 hm**)

Contrôle des utilisateurs

Interface d'exploration incluant recherche par mot et parcours des liens

Annexe 1 - Un extrait du DHI codé en SGML

Ci dessous, on trouvera les 10 premières entrées codées à partir du fichier qui doit servir à la 5ème édition française. La première partie comporte les déclarations nécessaires pour utiliser la partie “dictionnaire” de la DTD TEI, suivie du document lui-même. On retrouvera d’une part l’entête TEI et d’autre part, les entrées codées à l’intérieure d’une seule division.

L’ensemble constitue une première ébauche qui a servi aux auteurs de ce rapport à évaluer la faisabilité de l’opération de rétroconversion. En particulier, les renvois internes (notés à l’aide de majuscules dans le document d’origine et laissés tels quel ici) n’ont pas été transformés en pointeurs sur d’autres entrées.

```
<!DOCTYPE TEI.2 PUBLIC "-//TEI P3//DTD Main Document Type//EN" [
<!-- Mettre INCLUDE ou IGNORE en fonction des besoins -->
<!-- Base tag sets first, EXACTLY ONE -->
<!ENTITY % TEI.prose 'IGNORE'>
<!ENTITY % TEI.verse 'IGNORE'>
<!ENTITY % TEI.drama 'IGNORE'>
<!ENTITY % TEI.spoken 'IGNORE'>
<!ENTITY % TEI.dictionaries 'INCLUDE'>
<!ENTITY % TEI.terminology 'IGNORE'>
<!-- Mixed bases -->
<!ENTITY % TEI.general 'IGNORE'>
<!ENTITY % TEI.mixed 'IGNORE'>
<!-- Additional tag sets -->
<!ENTITY % TEI.linking 'INCLUDE'>
<!ENTITY % TEI.analysis 'IGNORE'>
<!ENTITY % TEI.fs 'IGNORE'>
<!ENTITY % TEI.certainty 'IGNORE'>
<!ENTITY % TEI.transcr 'IGNORE'>
<!ENTITY % TEI.textcrit 'IGNORE'>
<!ENTITY % TEI.names.dates 'INCLUDE'>
<!ENTITY % TEI.nets 'IGNORE'>
<!ENTITY % TEI.figures 'IGNORE'>
<!ENTITY % TEI.corpus 'IGNORE'>
<!ENTITY % ISOlat1 PUBLIC "ISO 8879:1986//ENTITIES Added Latin 1//EN">
%ISOlat1;

<!ENTITY SHOM-TEI-EN SYSTEM "SHOM-TEI-EN.sgml" SUBDOC>
]>
<tei.2>
<teiheader type="text" status="new">
<filedesc>
<titlestmt>
       <title>Dictionnaire Hydrographique International</title>
       <title type="GMD">une version électronique</title>
</titlestmt>
<publicationstmt>
       <p></p>
       <!-- one of (authority distributor publisher p) -->
</publicationstmt>
<sourcedesc default="NO">
       <biblstruct default="NO">
       <monogr>
       <editor role="editor">Organisation Hydrographique Internationale</editor>
```

```
      <title>Dictionnaire Hydrographique</title>
      <edition>4ème édition</edition>
      <imprint>
      <publisher>Bureau Hydrographique International</publisher>
      <pubplace>Monaco</pubplace>
      <date>1992</date>
      </imprint>
      </monogr>
      </biblstruct>
</sourcedesc>
</filedesc>
</teiheader>
<text>
<body>
<div type="lettre" n="A">
      <head>A</head>
<entry>
<xr><xptr doc="SHOM-TEI-EN" from="ID (3458)"></xr>
<form>
<orth>Abaque. </orth>
</form>
<gramgrp>
<pos>m</pos>
</gramgrp>
<sense>
<def>
Diagramme indiquant les relations entre plusieurs variables à l'aide de réseaux de
courbes graduées appropriées. Il permet de résoudre graphiquement des équations liant
ces diverses variables.</def>
<xr type="syn"><lbl>On dit aussi </lbl><ref>monogramme.</ref></xr>
</sense>
</entry>
<entry><form type=part>
<orth>Abaque </orth>
</form>
<gramgrp>
<pos>m</pos>
</gramgrp>
<form type=part>
<orth>(d'échelle).</orth>
</form>
<sense>
<def>Abaque permettant de déterminer l'échelle d'une carte en
un point donné lorsque celle-ci est fortement variable.</def>
</sense>
</entry>
<entry>
<xr><xptr doc="SHOM-TEI-EN" from="ID (648)"></xr>
<form type=part>
<orth>Abattre </orth>
</form>
<gramgrp>
<pos>vi</pos>
</gramgrp>
<form type=part>
<orth>en carène.</orth>
</form>
<sense>
<def>Coucher volontairement un navire sur un bord.</def>
</sense>
</entry>
<entry>
<xr><xptr doc="SHOM-TEI-EN" from="ID (167)"></xr>
```

<form type=part>
<orth>Aberration </orth>
</form>
<gramgrp>
<pos>f </pos>
</gramgrp>
<form type=part>
<orth>annuelle.</orth>
</form>
<sense>
<xr>Voir <ref>ABERRATION.</ref></xr>
</sense>

</entry>
<entry>
<xr><xptr doc="SHOM-TEI-EN" from="ID (773)"></xr>
<form type=part>
<orth>Aberration </orth>
</form>
<gramgrp>
<pos>f</pos>
</gramgrp>
<form type=part>
<orth>chromatique.</orth>
</form>
<sense>
<xr>Voir <ref>ABERRATION.</ref></xr>
</sense>

</entry>
<entry>
<xr><xptr doc="SHOM-TEI-EN" from="ID (4905)"></xr>
<form type=part>
<orth>Aberration </orth>
</form>
<gramgrp>
<pos>f</pos>
</gramgrp>
<form type=part>
<orth>de sphéricité.</orth>
</form>
<sense>
<xr>Voir <ref>ABERRATION.</ref></xr>
</sense>

<form>
<orth> Aberration.</orth>
</form>
<gramgrp>
<pos>f</pos>
</gramgrp>
<sense n="a">
<def>
<xr><xptr doc="SHOM-TEI-EN" from="ID (2)"></xr>
<def>En ASTRONOMIE, l'aberration de la lumière est le déplacement apparent de la position d'un CORPS CÉLESTE, due à la combinaison de la VITESSE de la lumière et de celle d'un observateur à la surface de la TERRE. L'aberration de la lumière due à la ROTATION de la TERRE sur son AXE est appelée ABERRATION DIURNE. Celle due à la RÉVOLUTION de la TERRE autour du SOLEIL est nommée ABERRATION ANNUELLE.</def>
</sense>
<sense n="b">
<def>En optique, défaut affectant un SYSTÈME OPTIQUE lorsque tous les RAYONS LUMINEUX issus d'un point objet ne convergent pas exactement en un point image de

```
position bien définie. <ref>L'aberration sphérique</ref> provient du fait que les
RAYONS ayant utilisé des zones différentes d'une LENTILLE ou d'un MIROIR convergent à
des distances différentes de cette LENTILLE ou de ce MIROIR. <ref>L'aberration
chromatique</ref> provient des différences d'INDICE DE RÉFRACTION des verres DU
SYST&Egrave;ME OPTIQUE en fonction de la couleur de la lumière, imparfaitement
corrigées de sorte qu'à chaque couleur correspond un FOYER différent.</def>
</sense>
</entry>
<entry>
<xr><xptr doc="SHOM-TEI-EN" from="ID (1414)"></xr>
<form type=part>
<orth>Aberration </orth>
</form>
<gramgrp>
<pos>f</pos>
</gramgrp>
<form type=part>
<orth>diurne.</orth>
</form>
<sense>
<xr>Voir <ref>ABERRATION.</ref></xr>
</sense>

</entry>
<entry>
<xr><xptr doc="SHOM-TEI-EN" from="ID (1411-4155)"></xr>
<form type=part>
<orth>Aberration </orth>
</form>
<gramgrp>
<pos>f</pos>
</gramgrp>
<form type=part>
<orth>radiale.</orth>
</form>
<sense>
<def>ABERRATION d'une lentille qu'on corrige sur la PHOTOGRAPHIE en dépla&ccedil;ant
l'IMAGE le long d'un rayon partant du POINT PRINCIPAL.</def>
</sense>
</entry>
<entry>
<xr><xptr doc="SHOM-TEI-EN" from="ID (5660)"></xr>
<form>
<orth>Abioseston.</orth>
</form>
<gramgrp>
<pos> m</pos>
</gramgrp>
<sense>
<def> Ensemble des particules détritiques en suspension dans l'eau.</def>
</sense>
</entry>
<entry>
<xr><xptr doc="SHOM-TEI-EN" from="ID (3)"></xr><form>
<orth>Abrasion.</orth>
</form>
<gramgrp>
<pos>f</pos>
</gramgrp>
<sense>
<def>Action d'user par frottements.</def>
</sense>
</entry>
```

```
</div>
</body>
</text>
</tei.2>
```

Annexe 2 - Extrait du DHI mis en page automatiquement par une feuille de style DSSSL

Le fichier généré automatiquement, comprenant à la fois les pages de titre (à partir des informations contenues dans l'entête) et les 10 premières entrées du dictionnaire), à l'aide de l'outil "jade" est du RTF (le format d'échange « standard » de Microsoft). La version correspondante en html est directement consultable sur le web à l'adresse suivante :

http://www.loria.fr/~romary/SHOM/SHOM.html

Dictionnaire Hydrographique International

une version électronique

Organisation Hydrographique Internationale

Dictionnaire Hydrographique

4ème édition

Bureau Hydrographique International
Monaco
1992

A

Abaque. - *m* - Diagramme indiquant les relations entre plusieurs variables à l'aide de réseaux de courbes graduées appropriées. Il permet de résoudre graphiquement des équations liant ces diverses variables. On dit aussi *monogramme.*

Abaque - *m* - **(d'échelle).** Abaque permettant de déterminer l'échelle d'une carte en un point donné lorsque celle-ci est fortement variable.

Abattre - *vi* - **en carène.** Coucher volontairement un navire sur un bord.

Aberration - *f* - **annuelle.** Voir *ABERRATION.*

Aberration - *f* - **chromatique.** Voir *ABERRATION.*

Aberration - *f* - **de sphéricité.** Voir *ABERRATION.*

Aberration. - *f* - a) En ASTRONOMIE, l'aberration de la lumière est le déplacement apparent de la position d'un CORPS CÉLESTE, due à la combinaison de la VITESSE de la lumière et de celle d'un observateur à la surface de la TERRE. L'aberration de la lumière due à la ROTATION de la TERRE sur son AXE est appelée ABERRATION DIURNE. Celle due à la RÉVOLUTION de la TERRE autour du SOLEIL est nommée ABERRATION ANNUELLE.

b) En optique, défaut affectant un SYSTÈME OPTIQUE lorsque tous les RAYONS LUMINEUX issus d'un point objet ne convergent pas exactement en un point image de position bien définie. *L'aberration sphérique* provient du fait que les RAYONS ayant utilisé des zones différentes d'une LENTILLE ou d'un MIROIR convergent à des distances différentes de cette LENTILLE ou de ce MIROIR. *L'aberration chromatique* provient des différences d'INDICE DE RÉFRACTION des verres DU SYSTÈME OPTIQUE en fonction de la couleur de la lumière, imparfaitement corrigées de sorte qu'à chaque couleur correspond un FOYER différent.

Aberration - *f* - **diurne.** Voir *ABERRATION.*

Aberration - *f* - **radiale.** ABERRATION d'une lentille qu'on corrige sur la PHOTOGRAPHIE en déplaçant l'IMAGE le long d'un rayon partant du POINT PRINCIPAL.

Abioseston. - *m* - Ensemble des particules détritiques en suspension dans l'eau.

Abrasion. - *f* - Action d'user par frottements.

Annexe 3 - Extrait d'une feuille de style DSSSL permettant de produire un document RTF à partir d'un fichier TEI

On voit dans cet exemple qu'à chaque élément peut être associé une fonction qui définit des caractéristiques de présentation associées au contenu correspondant.

```
;; ==== BASE TAG SET FOR PRINTED DICTIONARIES ========================

(element ENTRY
         (ENTRY-PARAGRAPH))

(element FORM
         (BOLD-SEQUENCE))

(element GRAMGRP
         (OUTPUT-GRAMGRP))

(element SENSE
         (if (string? (attribute-string "n"))
             (OUTPUT-SENSE-N (attribute-string "n"))
           (OUTPUT-SENSE)))

(element DEF
         (OUTPUT-DEF))

(element USG
         (USG-PARAGRAPH))

(element LBL
         (score-seq 'after))


;; ==== FUNCTIONS ============================================================

(define (ENTRY-PARAGRAPH)
 (make paragraph
       use: p-style
       space-before: *para-sep*
       first-line-start-indent: (- (* 2 (ULSTEP)))
       start-indent: *entry-start-indent*
       quadding: (ALIGN "JUSTIFY")
       (process-children-trim)))

(define (USG-PARAGRAPH)
 (make paragraph
       use: p-style
       space-before: 0em
       space-after: 0em
       first-line-start-indent: (/ *entry-start-indent* 2)
       start-indent: *entry-start-indent*
       quadding: (ALIGN "JUSTIFY")
       (process-children-trim)))

(define (OUTPUT-GRAMGRP)
```

```
  (make sequence
        font-posture: 'italic
        (literal " - ")
        (process-children-trim)
        (literal " - ")))

(define (OUTPUT-SENSE-N attribn)
  (make sequence
        (literal " ")
        (literal attribn)
        (literal ") ")
        (process-children-trim)))

(define (OUTPUT-SENSE)
  (make sequence
        (literal " ")
        (process-children-trim)))

(define (OUTPUT-DEF)
  (make sequence
        (literal " ")
       (process-children-trim)))

(define (OUTPUT-LBL)
 (make sequence
       under-mark: single
       (process-children-trim)))

;; ====================================================================
```