L'informatisation du Dictionnaire hydrographique international : normalisation et utilisation

Laurent Romary, Patrice Bonhomme, Bessero Gilles

L'informatisation du *Dictionnaire hydrographique international* : normalisation et utilisation

1 Du support papier à l'information électronique

L'informatisation d'un ouvrage du type dictionnaire doit pouvoir se situer dans la perspective plus large de la transformation de l'information textuelle – quelle que soit son origine : littéraire, manuscrit, etc. – d'un support papier, par essence relativement pérenne dans sa forme et son contenu, à un représentation électronique pour laquelle ces simples notions de forme et de contenu sont à la limite même difficiles à définir. Sans entrer dans une analyse fine de différents concepts qui accompagnent la notion de livre électronique ou simplement de dictionnaire informatisé[1], nous pouvons tenter de dégager quelques éléments de réflexion qui risquent de subordonner le succès de l'entreprise de numérisation d'un dictionnaire.

Considérons tout d'abord le support papier qui sert de référence. Celui-ci contient de nombreuses informations de natures différentes qui résultent de choix liés à la fois au contenu, à la réalisation pratique de l'ouvrage, ou éventuellement à des considérations plus esthétiques. Les indications typographiques d'un dictionnaire permettent en particulier de séparer les différents champs contenus dans une entrée (indications grammaticales, définition, exemples, etc.). Leur signification est donc intimement liée à une *codification* préalable que des indications introductives, ou encore une pratique intertextuelle, permettent de comprendre. Ces indications typographiques sont d'ailleurs souvent associées à une certaine codification de la matière linguistique (abréviations pour les marques de catégorie grammaticales par exemple) qui en facilite encore plus la lecture.

De son côté, le support informatique impose ses propres contraintes qui nécessairement introduiront un écart par rapport à la version papier de référence. Il est ainsi clair qu'il serait vain d'attendre une lisibilité du contenu qui soit *équivalente* lorsque l'on passe du papier à l'ordinateur. De façon plus large, on peut immédiatement percevoir qu'en aucun cas on ne peut obtenir une fidélité parfaite

1. On trouvera notamment une réflexion très intéressante sur les usages possibles d'un dictionnaire informatique dans le rapport de D. Piotrowski (1996).

par rapport au texte d'origine, quand bien même cette notion aurait un sens. Sans entrer dans ce débat, on peut s'interroger sur la façon dont il faut utiliser les informations contenues dans un dictionnaire papier. Prenons, par exemple, le cas d'un mot marqué en caractère gras dans une entrée de dictionnaire. Différentes solutions nous sont offertes pour coder ceci de façon électronique. On peut tout simplement occulter cette information et ne garder que la chaîne de caractère. On peut aussi – sous un format à déterminer – mémoriser l'indication typographique telle quelle (/le mot est en gras/). Enfin, on peut attacher au mot une indication *sémantique* correspondant à la signification du marqueur typographique initial (ex. /le mot est une entrée du dictionnaire/). On constate que ces différents choix dépendent intimement de ce que l'on souhaite privilégier lorsque l'on informatise un texte, de sa structure physique (son aspect extérieur) ou de sa structure logique (l'organisation de son contenu). Il s'agit là de choix éditoriaux relatifs à la matière électronique sur lesquels nous reviendrons dans le cas particulier du *Dictionnaire hydrographique international (DHI).*

Une autre raison d'admettre un certain écart entre le texte tel qu'il apparaît sur support papier et une représentation électronique a trait aux nouveaux usages que l'on souhaite pouvoir associer au texte électronique. En effet, la souplesse du support électronique, par exemple par la possibilité d'ajouter des informations qui ne seront pas immédiatement visibles à l'utilisateur (à des fins d'indexation par exemple), ou encore d'annoter dynamiquement les données proposées, permet d'imaginer bien d'autres accès qu'une recherche par mots-vedettes dans un dictionnaire ou d'autres visualisations que de présenter les entrées une à une dans leur intégralité. Dès lors, il faut envisager une représentation informatique d'un texte et *a fortiori* d'un dictionnaire comme une structure ouverte[2] propre à devenir la matière première de nombreux usages différents, et non plus comme un produit fini, dans l'esprit de ce que peut être un ouvrage édité.

Une fois posée la question du contenu informationnel que l'on désire mettre sur support électronique, l'étape suivante consiste à définir un format de représentation de ce contenu.

C'est là l'objectif de l'étude qui a été menée conjointement par le Service hydrographique et océanographique de la marine (Shom) et le Laboratoire lorrain de recherche en informatique et ses applications (Loria), sous l'égide de l'Organisation

2. On n'est pas loin ici de la notion d'*œuvre ouverte* telle que définie par U. Eco (1965)

hydrographique internationale (OHI). Nous allons voir dans cet article comment les considérations précédentes ont guidé nos choix pour aboutir à un cadre qui permette d'envisager des opérations d'édition et de parcours qui puissent être les plus souples possibles.

2 Contexte du travail

L'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 Shom a demandé au laboratoire Loria une étude de faisabilité permettant d'identifier les éléments à mettre en œuvre pour ce projet[3].

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*[4] ;
- *Dictionary of Abbreviations and Acronyms in Geographic Information Systems, Cartography, and Remote Sensing* by Philip Hoehn and Mary Larsgaard, October 1997, Version 3.0[5] ;

3. Les exemples de ce rapport sont extraits de la 5e édition anglaise et de la 4e ou de la 5e édition à venir pour le volume français.

4. www.state.nv.us/cnr/ndwp/dict-1/waterwds.htm.

5. 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[6].

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)[7] ;
- le *Middle English Dictionary* à l'Université du Michigan[8].

3 Contenu des entrées

3.1 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).

6. sulu.lerc.nasa.gov/dictionary/intro.html.
7. www.ciril.fr/~mastina/TLF.
8. www.hti.umich.edu/dict/.

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. »

3.2 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[9] (*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).

9. Dans la 4e édition, on pouvait rencontrer une telle indication (entre parenthèses), notamment quand il existe un risque d'ambiguïté quant à la catégorie morphologique du mot-vedette. Le choix semble avoir été fait de généraliser ces indications pour la 5e édition.

3.3 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.

4 Codage des entrées

4.1 Vers une perspective « éditoriale »

Dans la mise en œuvre 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 (Ide & Véronis, 1995) :

- Le point de vue *typographique* vise à préserver dans les moindres détails la forme (bidimensionnelle) du document initial (ruptures de page, colonnes, 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.

4.2 Utiliser la TEI : pourquoi ?

L'informatisation du *DHI* s'inscrit bien 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[10] 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[11], la plupart des chercheurs de sciences humaines (linguistes, philologues, historiens,

10. SGML (*Standard Generalized Markup Language*) est un standard reconnu par l'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 entre 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*).

11. ACH (*Association for Computers and Humanities*), ALLC (*Association for Literary and Linguistic Computing*), ACL (*Association for Computational Linguistics*).

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 telles 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, les 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. À 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>
```

4.3 Structure générale du dictionnaire

La TEI structure tout document électronique en deux parties principales :

- d'une part, un en-tête 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>.

Nous ne détaillerons pas ici la structure de l'en-tête TEI, mais nous insistons malgré tout sur l'importance d'une documentation associée à tout texte électronique afin de gérer le plus précisément possible son histoire et ses caractéristiques principales. C'est en particulier un moyen sûr de traiter convenablement les versions correspondant aux opérations d'édition que le document subit.

Du point de vue du contenu, 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 annexes 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>).

On obtient ainsi un canevas de représentation tel que suit :

```
<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>
```

4.4 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, dans le cas du *DHI*, contiendront une définition (<def>) et éventuellement des indications d'usage (<usg>).

À titre d'exemple, voici de façon simplifiée 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 ACCOMMODATION, 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'œil humain permettant de maintenir une vision nette des objets quelle que soit leur distance. En STÉRÉOSCOPIE, 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'œil humain permettant de maintenir une vision nette des objets quelle
        que soit leur distance.</def>
    </sense>
    <sense>
        <usg type="domaine">En STÉRÉOSCOPIE</usg>,
        <def>faculté des yeux humains d'obtenir la vision stéréoscopique par superposition de
        deux images.</def>
    </sense>
</entry>
```

4.5 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 propre 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 :

<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[12], 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-clés permettant soit des accès directs aux parties d’un document (racine du document, élément possédant un ‘id’ particulier), soit des accès en relatif par la description d’un chemin dans la structure. Sans entrer dans les détails de ces échelles de positionnement, nous pouvons illustrer leur fonctionnement à l’aide d’un exemple.

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

<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

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

correspondant à des synonymes ou des redirections sur des entrées plus complètes. Ainsi, ABERRATION DIURNE n'est pas défini 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="berration"> 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 (p.ex. « 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 (p.ex. 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).

4.6 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 erroné 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.

4.6.1 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 sur les sous-découpages en acceptions (<sense>). Trois opérations (ajout, modification, suppression) 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)[13] – 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 (p.ex. 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-vedettes 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).

13. On utilise ici la terminologie de *volume source* et de *volume cible* pour désigner respectivement l'origine et la destination des références externes de volume à volume au sein du *DHI*. Dans la structure actuelle de celui-ci, le volume cible sera ainsi nécessairement l'anglais.

4.6.2 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 4.5 :

- 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 est 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[14] :

  L'entrée GLOBE, aura pour clef d'entrée *globe* ;

  L'entrée MOUILLAGE DE QUARANTAINE aura pour clef d'entrée *mouillage_de_quarantaine* ;

  L'entrée TÉLÉDÉTECTION aura pour clef d'entrée *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'[15]. 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.

---

14. *Cf.* aussi l'usage dans un pointeur pour l'entrée ABERRATION dans la section 4.5.

15. Cet attribut fait déjà partie de la TEI et est utilisé notamment pour indiquer le statut global d'un texte dans l'en-tête 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.

- 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 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 tel 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.

Bien que dans un premier temps, nous nous soyons appuyés sur l'hypothèse d'une rétroconversion *a minima* du *DHI* en gardant le principe d'un pointage de tous les volumes sources 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.

4.7 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.
- 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, 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 précision 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 d'autres éléments tout à fait spécifiques au contexte d'utilisation du *DHI* (la TEI prévoit ce mécanisme).

5 Conclusions et perspectives

Le travail présenté dans cet article était initialement guidé par les contraintes propres à un ouvrage particulier et à une communauté d'éditeurs et d'utilisateurs relativement spécifique. Il ressort cependant que la méthodologie adoptée nous semble suffisamment générale pour que nous en dégagions ici les grandes lignes :

- distinguer la représentation interne des données lexicographiques de leur utilisation future, de façon à ce que les contraintes logiques priment sur les contraintes de présentation ;
- adopter un cadre normatif qui permette non seulement de s'appuyer sur l'expérience acquise au sein d'autres projets similaires, mais aussi d'assurer une compatibilité de l'ouvrage que l'on met sous une forme électronique avec d'autres ouvrages similaires ;
- définir, en lien avec les utilisateurs de l'ouvrage, une politique éditoriale qui accompagne les choix techniques envisagés.

Au total, il semble que les avancées technologiques récentes permettent d'envisager de façon cohérente et conviviale la mise en œuvre de dictionnaires informatisés multilingues. Dans le domaine maritime, où la communication est un facteur primordial, notre travail peut servir de base à la définition de systèmes d'information plus globaux, intégrant ouvrages de référence, dictionnaires, transmission de messages au sein d'un même poste de travail embarqué.

Laurent Romary,
Patrice Bonhomme,
Laboratoire lorrain de recherche en informatique et ses applications (Loria),
Nancy.

Gilles Bessero,
Service hydrographique et océanographique de la Marine (Shom),
Paris,
France.

Bibliographie

Eco (U.), 1965 : *L'œuvre ouverte*, Paris, Seuil (réédition : Points, Seuil, 1979).

Ide (N.) et Véronis (J.), réd., 1995 : *Text Encoding Initiative, Background and Context*, Dordrecht, Boston et London, Kluwer Academic Publishers.

Piotrowski (D.), réd., 1996 : *Lexicographie et informatique: autour de l'informatisation du Trésor de la langue française. Actes du colloque international de Nancy (29, 30, 31 mai 1995)*, Paris, Didier-Érudition.

Sperberg-McQueen (M.) et Burnard (L.), 1994 : *Guidelines for Electronic Text Encoding and Interchange*, Chicago et Oxford, Text Encoding Initiative.