

L 'Analyse Conceptuelle.

**Identification, extraction et description
des connaissances dans les textes**

Peter Stockinger

Institut National des Langues et Civilisations Orientales

Maison des Sciences de l'Homme

Paris 1992

Sommaire

1) PROBLÉMATIQUE GÉNÉRALE	3
2) LE SYSTÈME DE REPRÉSENTATION DE DOCUMENTS	7
2.1) TYPES DE DOCUMENTS.....	7
2.2) DESCRIPTION DES TROIS TYPES DE DOC.....	15
2.2.1) LE DOC.TEXTE	15
2.2.2) LE DOC.THEMA.....	19
2.2.3) LE DOC.CONFIG-THEMA.....	23

1) **Problématique générale**

La description et l'exploitation d'une expertise textuelle recouvre les phases de l'extraction, de la description et de la modélisation des "connaissances" contenues dans un document/un ensemble de documents ou encore dans un discours d'expert(s) en vue de leur stockage sous forme de "bibliothèques" (électroniques ou traditionnelles).

Ces "bibliothèques" peuvent satisfaire à une multitude de besoins concrets telles que la mise à jour (modification, adjonction, suppression, ... d'informations) ciblée et locale, la réutilisation (partielle ou complète) d'une méthodologie déjà établie dans d'autres procédures du transfert de l'expertise du même type (cf. la problématique de la portabilité d'une méthodologie), la recherche d'informations pertinentes dans des bases documentaires, la production de documents assistée par ordinateur, la constitution de bases de données et de connaissances, la navigation "intelligente" dans une base de données multimédias, etc.

La problématique à laquelle doit satisfaire le transfert de l'expertise est, par définition, décrite pendant la phase des préalables de l'analyse conceptuelle.

Nous ne traitons ici que le composant de la *thématique*, c'est-à-dire du *contenu* (les informations, les connaissances contenues dans un document ou un discours d'expert) à extraire, décrire et modéliser.

Cela signifie concrètement que l'analyste doit avoir à sa disposition un document établi à la fin des investigations préalables à l'analyse conceptuelle proprement dite lui précisant surtout:

- les objets à décrire,
- le ou les standards selon lesquels les objets doivent être décrits,
- éventuellement le ou les destinataires et les objectifs visés,
- éventuellement des indications ou "files conducteurs" spécifiant davantage le processus même du transfert de l'expertise.

Voici un premier exemple d'un tel document: *Identification et détermination de la problématique du transfert de l'expertise des maladies tropicales*

1) OBJET A DECRIRE

- les maladies tropicales: maladies parasitaires
- les aspects: symptomatologie, processus du diagnostic, types du traitement

2) STANDARD CHOISI:

- standard biomédical dont portent garants:
 - * les experts (l'expert): nom
 - * les documents (le document): titre

3) DESTINATAIRE ET OBJECTIF:

- apprentissage de la nosographie pour aide-soignants travaillant dans des centres médicaux en Afrique Noire,

- aide au diagnostic pour le personnel medical non-medecin travaillant dans des centres médicaux en Afrique Noire,
- aide pour le choix du traitement médical pour le personnel medical non-medecin dans des centres médicaux en Afrique Noire.

4) INDICATIONS PARTICULIERES

- i) ne prendre en considération que les parties descriptives et explicatives du corpus (oral, écrit), à exclure: ...
- ii) faire attention aux tournures linguistiques semantiquement équivalentes et choisir une terminologie standardisable,
- iii) faire attention aux tournures linguistiques ne rentrant pas dans le cadre du standard biomédical et expliciter leur(s) sens.

Deuxième exemple : *Identification et détermination de la problématique du transfert de l'expertise concernant la caractérisation des îles grecques*

1) OBJET A DECRIRE:

- les îles grecques: Arki, Corfu, Cephalonie, ...
- les aspects: localisation des îles, climat, paysage, habitants, activités artisanales, ...

2) STANDARD CHOISI

- standard "touristique généralement accessible" dont portent garants:
 - * l'expert (les experts): nom
 - * le document (les documents): titre

3) DESTINATAIRE ET OBJECTIF:

- apprentissage non-spécialisé
- renseignements pratiques
- guides de circuits

4) INDICATIONS PARTICULIERES:

- à prendre en considération que les aspects descriptifs et évaluatifs du corpus
- à prendre en considération seulement les données permettant de caractériser les aspects de l'objet à décrire,
- faire attention aux tournures linguistiques semantiquement équivalentes et essayer de mettre en place une terminologie stabilisée.

Ces documents peuvent évidemment être plus détaillés. Il ne s'agit ici que de documents-types dont la pertinence reside dans leurs structure et organisation internes.

La méthodologie à l'aide de laquelle le transfert de l'expertise se fait est:

- soit d'ordre analytique ("general"),
- soit d'ordre spécifique adapté à l'objet de l'expertise.

Le deuxième cas s'impose si on dispose déjà d'un métalangage approprié à l'objet de l'expertise et si ce métalangage est considéré comme pertinent pour l'expertise à mener.

Lors des investigations antérieures ou parce qu'il existe comme "théorie" acceptée", on peut avoir un métalangage de description de symptômes "mal à la tête" suivant:

- [types de symptômes]:
- [intensité]:
- [fréquence]:
- [localisation]
- [durée]:
- [accompagné par]:
- [indique maladie_x]:
- [indique maladie_x] si [type_x]
- [se soigne avec objet_x]:
- [se soigne avec objet_x] si [type_x]
- etc.
-

Lorsque il faut extraire et modéliser ce type de connaissances, on aura recours directement à ce métalangage descriptif à condition qu'il soit accepté par l'expert ou le commanditaire. On n'aura pas recours au métalangage analytique qui est - d'un point de vue empirique - moins pertinent que le métalangage spécifique.

Le métalangage spécifique, cependant, se construit selon les mêmes principes que le métalangage analytique; il comportera notamment toujours:

- la distinction entre le composant pragmatique et le composant sémantique (thématique),
- une structuration du composant pragmatique selon les indices de validité et les actes de langage,
- une structuration du composant sémantique selon des critères taxinomique, paronymique, attributif, fonctionnel, modal, spatio-temporel, narratif ou encore rhétorique,
- les principes généraux déterminant la construction de tous les métalangages (principes référentiel, pragmatique, générique, de la modularité, de l'équilibre metastable et le principe métalinguistique).

Construire les différents métalangages spécifiques selon les mêmes principes et structures est en effet indispensable pour garantir la cohérence et la compatibilité interne du système de métalangages et donc de leur comparaison, évaluation, spécification, modification, etc ainsi que de leur *portabilité*, c'est-à-dire de leur réutilisabilité dans d'autres applications.

Si le transfert de connaissances de l'expertise ne peut avoir recours à un métalangage spécifique, il évoluera en suivant le métalangage analytique décrit dans Stockinger (1992) ou encore dans Greimas (1966, 1979).

Le déroulement du transfert de l'expertise peut être caractérisé comme suit:

1) extraction proprement dite de données textuelles pertinentes sous forme d'une liste ordonnée d'énoncés en un langage relativement standardisé et repérage-regroupement des indicateurs pragmatiques;

2) regroupement des énoncés en suivant un principe d'homogénéité sémantique en créant des catégories thématiques nouvelles ou en utilisant des catégories thématiques déjà existantes;

3) regroupement des catégories thématiques sous forme d'une configuration (ou de plusieurs configurations) qui constituera la théorie;

4) validation de la théorie sur le même corpus, un corpus autre ou encore par rapport aux attentes d'un destinataire qui peut aussi être le commanditaire.

Ces quatre tâches énumérées constituent le **processus de la description** dans le module du transfert de l'expertise.

Ce processus peut être visualisé par un système de représentation sous forme de **documents** (électroniques ou "traditionnels", réels ou virtuels) qui constituent des langages de représentation.

Dans les deux chapitres suivants, nous détaillerons:

- 1) Le système de représentation de documents,
- 2) Le processus de la description.

2) Le système de représentation de documents

Le transfert de l'expertise se concrétise donc sous forme de plusieurs types de documents (à support papier et/ou électronique). Les documents constituent une bibliothèque devant permettre le stockage et l'archivage, l'accès, la mise à jour et la diffusion ("personnalisée") d'informations ou de connaissances pertinentes ainsi que la création de bases de données et de connaissances.

2.1) Types de Documents

Nous distinguons les trois types suivants de documents:

1) *document de données textuelles pertinentes* (= "doc.texte" correspondant à la première tâche du processus de la description);

2) *document thématique de l'objet de l'expertise* (= "doc.thema" correspondant à la deuxième tâche du processus de la description);

3) *document de la configuration thématique* (= "doc.config" correspondant à la troisième tâche du processus de la description).

Voici des exemples simples des trois types de documents:

Doc.texte:

DOC:

Ile Arki -

1. est connue, dans l'Antiquité, sous le nom d'Akrité.
2. est une île isolée, sauvage et peu remarquable.
4. se trouve à 1 km au nord-est de Patmos.
5. a peu d'*habitants* -
 - 5.1. se consacrent à la pêche et à l'élevage.
 - 5.2. mènent une vie modeste et à l'écart des commodités d'aujourd'hui.
6. est cernée par de nombreux *récifs et îlots*
 - 6.1. ont une petite taille
 - 6.2. servent au pâturage des chèvres.

Doc.texte:

DOC:

Ascaridiose -

1. est une parasitose intestinale.
2. est localisé dans le monde entier.
3. est très répandue chez l'enfant.
4. provoque parfois des troubles digestifs.

Doc.index:

DOC:

Ile Arki -

LOCALISATION

1.
2.

CLIMAT

1. ...

HABITANT

ACTIVITE:

1. ...
2. ...
3. ...

HISTOIRE:

1. ...
2. ...

PAYSAGE

....

Doc.thema:

CONTENU:

Ascaridiose -

TYPE_MALADIE

1. est une parasitose intestinale

EPIDEMIOLOGIE

LOCALISATION_GEOGRAPHIQUE

1. est localisé dans le monde entier

POPULATION_CIBLE

1. est très répandue chez l'enfant

SYMPTOME

1. provoque parfois des troubles digestifs

Doc.config.-thema.

DOC:

"AGRICULTURE" - Définition canonique

AGRICULTURE -

1. est un type d'activité professionnelle,
2. s'exerce dans un certain endroit,
3. fournit un certain type de produits,
4. s'appuie sur un ensemble de méthodes et d'instruments,
5. est réglementée par un ensemble de règles déontologiques,
6. se caractérise par une certaine *evolution* -
 - 6.1. est de type social,
 - 6.2. est de type technique,
 - 6.3. est de type économique

.....

Systeme de Représentation

[AGRICULTURE]---(est-un)---> [METIER]

(agt)---> [AGRICULTEUR]

(obj)--->[PRODUIT_AGRICOLE]

.....

Doc.config.-thema.

CONTENU:

"ASCARIDIOSE"

Définition canonique:

ASCARIDIOSE -

1. est une maladie parasitaire.
2. est localisée géographiquement

Ajoutons tout de suite une précision de taille: on distinguera, par définition, pour chaque type de documents deux parties:

- une partie **meta-doc**,
- une partie **doc**.

i) *Méta-Doc et Doc*

La distinction entre la partie "méta-document" et la partie "document" correspond à la distinction entre le contenu d'une proposition et la validité ou la prise en charge de la validité du contenu dans notre méthodologie, i.e.:

- c'est le Baedeker qui dit que l'île Arki ne possède que peu d'habitants,
- c'est le guide bleu qui dit que l'île Corfu ...
- c'est le manuel M qui dit que ...
- c'est pour le destinataire D que l'auteur A du manuel M dit que ...
- c'est l'auteur A qui affirme que l'agriculture est ...
- c'est l'auteur A qui affirme que les habitants de l'île Arki ...

La distinction entre les deux parties ne doit donc pas se limiter à une simple distinction graphique et visuelle (sur l'écran) mais elle doit permettre à un utilisateur (cogniticien, rédacteur, apprenant, ...) d'une part de *distinguer* explicitement entre le contenu et la validité du contenu et d'autre part de *comparer* et d'*évaluer* entre différents standards, i.e.:

- selon le Baedeker et le Guide Bleu de Hachette, les habitants de l'île Arki ... mais selon le Guide de Routards ...
- le manuel A affirme que l'agriculture est; le manuel B affirme que l'agriculture est..., le manuel C affirme que l'agriculture est ...: les différences entre ces trois manuels portent sur, les équivalences portent sur:;
- etc.

Donc:

a) pour le transfert de l'expertise, l'analyste doit disposer d'une *interface graphique distinguant entre deux parties* - le méta-doc et le doc lui permettant d'introduire dans le méta-doc toutes les informations relatives à la validité du doc et dans le doc toutes les informations concernant soit un référent soit une théorie (une description) générique ou spécifique du référent;

b) pour accéder, évaluer (et mettre à jour) une expertise archivée dans la bibliothèque, l'utilisateur doit avoir la *possibilité de "poser de questions"* au document concernant l'auteur d'une expertise, sa validité temporelle, sa validité par rapport à d'autres auteurs, sa validité intersubjective (c'est-à-dire pour tel ou tel type de destinataire), etc.

ii) la récursivité entre les parties méta-doc et doc

Souvent on est confronté à des expertise du type suivant:

"Dans le Guide du Musée des Arts Africains et Océaniens, on affirme que selon les Dogons les statuettes de formes étaient inventées par les Tellem. Il s'agit d'un peuple"

On a ici:

méta-doc: Guide du Musée
 doc: statuette chez les Dogons
 méta-doc: Dogons
 doc: inventeur des statuettes = Tellem.

(ce n'est pas, à strictement parler, le Guide du Musée qui affirme que les Tellem sont les inventeurs de statuettes chez les Dogons mais ce sont les Dogons qui l'affirment, donc: Le Guide du Musée affirme que les Dogons possède des statuettes et que les Dogons affirment que les Tellem les ont inventées...).

Il faut donc prévoir la possibilité qu'à l'intérieur de la partie doc puisse s'ouvrir de nouveau une partie graphique méta-doc et une autre partie graphique doc (et ainsi de suite, s'il en faut).

Ici, de nouveau, il ne s'agit pas d'avoir uniquement des distinctions graphiques; il faut, en plus, la possibilité à un analyste-utilisateur d'évaluer, de comparer, etc (cf. point 1).

Exemples:

A) presence d'une unité textuelle "métadoc" et d'une unité textuelle "doc", i.e.:

METADOC:
DOC:

B) dans l'unité textuelle "métadoc" doivent figurer, entre autre, le nom de l'auteur (de l'analyste), l'identité du corpus/du texte, le domaine de référence du corpus/du texte, la date de la conception-rédaction du doc et d'autres informations jugées pertinentes pour la bonne utilisation du doc, i.e.:

doc.texte.:

<p>METADOC: nom-analyste: document: Baedeker 1991 texte: IleArki (1 page) segment: paragraphe "situation et généralités" (5 lignes) domaine: tourisme date analyse:.... date rédaction: </p>
<p>DOC: <i>Ile Arki</i> - 1. est connue, dans l'Antiquité, sous le nom d'Akrité. 5. a peu d'<i>habitants</i> - 5.1. se consacrent à la pêche et à l'élevage. 5.2. mènent une vie modeste et à l'écart des commodités d'aujourd'hui.</p>

doc.thema

METADOC:

nom-analyste:

document: Baedeker 1991

texte: IleArki (1 page)

segment: paragraphe "situation et généralités" (5 lignes)

domaine: tourisme

date analyse:....

date rédaction:

.....

DOC:

Ile Arki -

LOCALISATION

1.

2.

CLIMAT

1. ...

HABITANT

ACTIVITE:

1. ...

2. ...

3. ...

HISTOIRE:

1. ...

2. ...

PAYSAGE

....

doc.config.-thema.:

METADOC:

nom-analyste:

document: Baedeker 1991

texte: IleArki (1 page)

segment: paragraphe "situation et généralités" (5 lignes)

domaine: tourisme

date analyse:....

date rédaction:

.....

DOC:

Configuration Thématique: "AGRICULTURE"

Définition canonique

AGRICULTURE -

1. est un type d'activité professionnelle,
2. s'exerce dans un certain endroit,
3. fournit un certain type de produits,
4. s'appuie sur un ensemble de méthodes et d'instruments,
5. est réglementée par un ensemble de règles déontologiques,
6. se caractérise par une certaine *evolution* -
 - 6.1. est de type social,
 - 6.2. est de type technique,
 - 6.3. est de type économique

.....

Système de Représentation

[AGRICULTURE]---(est-un)---> [METIER]

(agt)---> [AGRICULTEUR]

(obj)--->[PRODUIT_AGRICOLE]

.....

2.2) Description des trois types de doc.

2.2.1) Le doc.texte

A) Caractérisation

Le doc.texte réunit, dans un langage sémi-normalisé, les données brutes mais pertinentes d'un corpus référant à une thématique commune (choisie par l'expert, l'analyste ou encore indiqué dans le corpus).

Si la thématique est *maladies parasitaires*, on réunira les données qui réfèrent à cette thématique mais, étant donné la généralité intrinsèque de la thématique, on prendra garde de la surspécification de cette thématique. On évitera, par exemple, de prendre en compte un certain type des caractéristiques spécifiant les caractéristiques "primaires" de la thématique à décrire.

On pourra, par exemple, avoir une expertise comme celle-ci:

"Les maladies parasitaires sont dues à la présence chez un individu (homme ou animal), d'un être vivant, animal ou végétal, qui vit au dépense de son hôte. ... Le plasmodium, par exemple, est le parasite responsable du paludisme, maladie parasitaire qui sévit ..."

Or, la thématique choisie est *maladies parasitaires* en general. On ne prendra en compte l'expertise portant sur le plasmodium ni celle introduisant le paludisme - expertise trop spécifique par rapport à la thématique choisie.

Il faut, en effet, respecter le *niveau de pertinence* choisi ou imposé. Un indicateur fort (mais ni nécessaire ni seul) du changement d'un niveau de pertinence est le *critère taxinomique*, i.e.

niveau _i	>	niveau _{i+n}
maladies parasitaires (description...)	>	paludisme (description incluant le parasite plasmodium...)

Un autre indicateur fort est le critère pragmatique du *standard* et de ses *indices de validité*, i.e.:

standard ₁	vs	standard ₂	vs	standard ₃	vs etc
standard bio-médical actuel		standard biomédical passé		standard de la médecine populaire	

B) Le langage sémi-normalisé

La standardisation du langage utilisé dans le doc.texte portent essentiellement sur sa forme et sa "grammaire":

- chaque unité "doc" doit être initialisée par le nom de l'objet à analyser (i.e. "Ile Arki") suivi d'un petit trait (i.e. "-" donc: "Ile Arki -") qui indique le fait que tout ce qui lui suivra à droite dépend (sémantiquement parlant) de l'objet qui se trouve à gauche de lui);

- les unités qui suivent le petit trait sont appelés "quasi-phrases" et cela d'une part pour pouvoir les différencier des phrases- (ou paragraphes-) occurrences dont elles sont issues et d'autre part pour insister sur les faits qu'elles possèdent une forme linguistique normalisée (avec, notamment, à la tête, le prédicat suivi des arguments sur lesquels il porte);

- chaque quasi-phrase est numérotée à gauche et clôturée par un point (".") à droite;

- une quasi-phrase peut comporter un ou plusieurs arguments (sous forme de SN, SP, etc) ;

- un argument d'une quasi-phrase peut être caractérisé par des quasi-phrases regroupés. Si ce cas se présente, alors on met en relief l'argument qui est typifié par un ou plusieurs quasi-phrases et on fait suivre d'un trait la quasi-phrase à laquelle l'argument appartient. Les quasi-phrases dépendant d'un argument particulier d'une autre quasi-phrase seront également numérotées mais le premier chiffre sera évidemment toujours celle que comporte la quasi-phrase dont fait partie l'argument.

Exemple:

DOC:

Ile Arki -

1. est connue, dans l'Antiquité, sous le nom d'Akrité.

.....

.....

5. a peu d'*habitants* -

5.1. se consacrent à la pêche et à l'élevage.

5.2. mènent une vie modeste et à l'écart des commodités d'aujourd'hui.

Remarque: on peut envisager, à court terme, une standardisation des expressions prédicatives à l'aide de la méthodologie de l'analyse conceptuelle portant sur la composante thématique en y distinguant quelques grands **types** d'expressions prédicatives tels que:

- type taxinomique,
- type type paronymique,
- type attributif,
- type fonctionnel,
- type modale,
- etc.

Par définition, chaque expression prédicative figurant à la tête d'une quasi-phrase dans un doc.texte sera interprété comme un **token** (une "occurrence") d'un de ces types canoniques - stratégie qui permettra non seulement, d'ores et déjà, de "mettre de l'ordre" dans la liste (ouverte) des quasi-phrases d'un doc.texte mais aussi, à moyen terme, de passer directement - à l'aide du doc.config.-thema. - des données brutes dans un corpus au doc.thema tout en rendant superflue l'existence physique du doc.texte (le doc.texte deviendra ainsi un document virtuel).

En respectant les quelques consignes données à propos de la forme et "grammaire" du doc.texte, on

- extrait ("manuellement") des phrases-occurrences pertinentes du corpus et
- les convertit dans le format-standard du doc.texte.

Exemple: La Peste (Maladie Parasitaire)

<p>Corpus: La peste est une maladie des rongeurs transmise par les puces, et atteignant accidentellement l'homme. Ayant provoqué de graves épidémies autrefois, la peste guérit parfaitement avec les antibiotiques modernes.</p>	<p>doc.texte: <i>Peste -</i> 1. est une maladie des rongeurs. 2. est transmise par les puces. 3. atteint l'homme accidentellement. 4. a provoqué de graves épidémies autrefois. 5. guérit parfaitement avec les antibiotiques modernes.</p>
---	---

La comparaison du texte à gauche (le corpus) et du texte à droite (le doc.texte) fait apparaître différentes opérations simples à effectuer lors du processus de la conversion:

opération de suppression

La peste est une maladies des rongeurs --->	... est une maladie des rongeurs
--	-------------------------------------

opération de copie

(La peste) est une maladies des rongeurs --->	... est une maladie des rongeurs
--	-------------------------------------

opération de complétion prédicative

(...) transmise par les puces --->	est transmise par les puces
---------------------------------------	-----------------------------

opération de restitution prédicative

(...) et atteignant accidentellement l'homme ---->	atteint l'homme accidentellement.
(...) ayant provoqué autrefois... ---->	a provoqué autrefois

Si on se donne l'objectif de la constitution *assistée par ordinateur* (donc, au moins, sémi-automatique) du doc.texte dans son format standard décrit ci-dessus, il faudra:

- déterminer la classe de ces opérations et définir les règles correspondantes et puis
- écrire un module de conversion qui sera lié à un analyseur lexico-syntaxique auquel on donnera, comme seule information le nom de la thématique qui sert comme référent.

Il faut néanmoins noter que, dans une telle perspective d'automatisation de la conversion de données pertinentes en quasi-phrases, l'opération de la condensation-réduction sémantique reste entièrement ouverte: l'activité de la condensation-réduction compactifie des données jugées redondantes ou synonymes en une seule quasi-phrase. Cette activité est évidemment très importante dès qu'on doit s'attaquer à un corpus quantitativement important.

C) Indérêts et Objectifs du doc.texte

Voici quelques possibilités d'utilisation du doc.texte:

- a) stockage et archivage de données pertinentes à propos d'une thématique;
- b) comparaison entre les données pertinentes d'une même thématique traitée par différents experts/corpus;
- c) mise-à-jours de données pertinentes: soit suppression, soit adjonction, soit modification de données/du doc.texte;
- d) utilisation des quasi-phrases sans leur tête prédicative comme un système de mots-clés co-occurents dans la recherche d'information;
- e) utilisation des quasi-phrases avec leur tête prédicative et le nom de la thématique comme des "inputs" pour la production (mais pas pour la génération !) de documents assistée par ordinateur à partir d'un ou de plusieurs doc.texte traitant de la même thématique.

D) Le doc.texte comme document virtuel

Le doc.texte est un document physique stocké, par exemple, dans une base de données dont les données sont fournies par de l'information déjà traitée.

Cependant, déjà pour des raisons techniques, il faut envisager le doc.texte comme un document virtuel, c'est-à-dire dont l'existence physique se réduit à la durée ou à la période pendant laquelle quelqu'un en besoin.

La réalisation du doc.texte comme document virtuel présuppose néanmoins un système très performant et sophistiqué incluant en particulier:

- un système gérant l'organisation et la structure internes du doc.texte;
- un analyseur lexico-syntaxique de haut niveau permettant la recherche automatisée de données pertinentes dans un corpus;
- un système de conversion de données pertinentes en un ensemble de quasi-phrases telles qu'elles sont utilisées dans le doc.texte.

Comme perspective intermédiaire on peut envisager de laisser à l'analyste le choix - à l'aide d'une simple commande - s'il veut ou non stocker un doc.texte.

2.2.2) Le doc.thema

A) Caractérisation

Le doc.thema permet de regrouper les quasi-phrases du doc.texte en un sous-ensemble sémantiquement homogène réunit sous un thème particulier de la thématique traitée.

Au lieu d'avoir donc, comme dans le doc.texte, d'une part le nom de la thématique et d'autre part un ensemble plus ou moins important mais sémantiquement pas structuré de quasi-phrases, on a dans le doc.thema:

- le nom de la thématique,
- un ou plusieurs thèmes constituant la thématique,
- éventuellement des conditions (des indices de validité) particulières déterminant l'existence de tel ou tel thème,
- une documentation sous forme de quasi-phrases provenant du doc.texte/de doc.textes ou encore, à partir du moment où on dispose du doc.texte virtuel, directement du corpus.

La structure - en soi simple - du doc.thema est la suivante:

- nom de la thématique suivi d'un tiré,
- noms de thèmes apparaissant sous forme séquentielle et (si nécessaire) hiérarchique,
- quasi-phrases dont le format est défini dans le doc.texte.

Exemple:

Doc.thema:

DOC:

Ascaridiose -

TYPE_MALADIE

1. est une parasitose intestinale

EPIDEMIOLOGIE

LOCALISATION_GEOGRAPHIQUE

1. est localisé dans le monde entier

POPULATION_CIBLE

1. est très répandue chez l'enfant

SYMPTOME

1. provoque parfois des troubles digestifs

La constitution du doc.thema peut se faire selon les trois façons suivantes:

- a) l'analyste extrait les thèmes dont il a besoin;
- b) l'analyste utilise *sans modification* les thèmes tels qu'ils sont définis par le graphe conceptuel dans le doc.config.-thema;
- c) l'analyste utilise *avec modification* les thèmes tels qu'ils sont définis par le graphe conceptuel dans le doc.config.-thema.

La première possibilité correspond à une situation où l'analyste n'a pas à sa disposition une définition conceptuelle (une "théorie") ou n'as pas à sa disposition une définition conceptuelle qu'il trouve satisfaisante. Il forge donc sa propre définition (sa propre "théorie").

Cette éventualité se présente, d'un point de vue de l'utilisateur, comme suit:

- * l'analyste veut convertir un doc.texte en doc.thema,
- * on lui affiche, s'ils existent, les doc.config.thema contenant les définitions et les graphes conceptuels pouvant l'intéresser,
- * l'analyste les consulte et les ferme, un à un,
- * on lui propose de créer un doc.config.thema sous les modes "provisoire" ou "acceptés",
- * on lui affiche un doc.config.thema vide et un espace de travail,
- * l'analyste élabore dans l'espace de travail sa définition et son graphe conceptuel et les introduit dans le doc.config.thema,
- * on utilise automatiquement les concepts apparaissant dans le graphe conceptuel du doc.config.thema "accepté", les convertit en thème et édite un doc.thema où figurent les concepts sous forme de thèmes séquentiellement et hiérarchiquement arrangés (note: les concepts se trouvant sous la dominance d'un concept; doivent apparaître dans le doc.thema

sous forme hiérarchique; le concept central du graphe marqué par un symbol spécial est toujours le nom de la thématique),

- * on lui propose ensuite de consulter les doc.thema référant à la même thématique dont on dispose déjà,

- * (l'analyste consulte, un à un les doc.thema soit sur écran soit dans forme imprimée),

- * l'analyste remplit ensuite les thèmes dans le doc.thema nouvellement créé avec les quasi-phrases provenant de son doc.texte (ou encore, s'il le souhaite de doc.thema déjà existants),

- * on clôt la session avec la demande de sauvegarde et de stockage du doc.thema et du doc.config.thema créés par l'analyste.

La deuxième et la troisième possibilité correspond, t à une situation où l'analyste à recours à une définition/à un graphe conceptuel déjà existant dans un doc.config.thema.

La deuxième possibilité, la plus simple et la plus rapide, est une application directe d'une théorie déjà existante sur une description à faire.

Une session de travail peut se concevoir comme suit:

- * l'analyste veut convertir un doc.texte en un doc.thema,

- * on lui affiche la ou les doc.config.thema contenant les définitions et graphes conceptuels qui l'intéressent,

- * l'analyste en choisit une et lance une commande signifiant qu'il l'accepte telle qu'elle,

- * si des doc.thema constitués à l'aide du doc.config.thema coisi par l'analyste existent déjà, on le lui signifie,

- * à sa demande, on les affiche un par un et, s'il le souhaite, on les imprime pour qu'il puisse les évaluer et comparer par rapport à son doc.texte,

- * puis, on le demande s'il souhaite de modifier un doc.thema déjà existant en y apportant les précisions provenant de son doc.texte ou s'il souhaite de créer un nouveau doc.thema,

- * s'il choisit l'option de la modification d'un doc.thema, on lui donne la possibilité de le modifier directement en lui permettant soit de supprimer de quasi-phrases, soit de les modifier soit encore d'y ajouter celles provenant de son propre doc.texte; enfin on lui demande à la fin de la session s'il veut sauvegarder et stocker uniquement le doc.thema modifié ou à la fois le doc.thema modifié et le doc.thema originaire,

- * si l'analyste choisit l'option de la création d'un nouveau doc.thema, on utilise automatiquement les concepts apparaissant dans le graphe conceptuel du doc.config.-thema, les convertit en thèmes et édite un doc.thema vide où figure les concepts sous forme de thèmes séquentiellement et hiérarchiquement arrangés (note: les concepts; se trouvant sous la dominance d'un concept; doivent apparaître dans le doc.thema sous forme hiérarchique; le concept central du graphe marqué par un symbol spécial est toujours le nom de la thématique),

- * l'analyste remplit ensuite les thèmes dans le doc.thema vide avec les quasi-phrases provenant de son doc.texte (ou encore, s'il le souhaite de doc.thema déjà existants),

- * on clôt la session avec une demande de sauvegarde et du stockage du doc.thema créée par l'analyste.

La troisième possibilité est une *application contrôlée* par l'analyste d'une théorie déjà existante sur une description à faire.

Une session de travail se présente comme suit:

- * l'analyste veut convertir un doc.texte en un doc.thema,
- * on lui affiche la ou les doc.config.thema contenant les définitions et graphes conceptuels pouvant l'intéresser,
- * l'analyste en choisit un ou plusieurs graphes conceptuels et indique qu'il veut le (les) retravailler,
- * on lui crée un espace de travail et on lui donne la possibilité d'y exporter le ou les graphes conceptuels choisi(s) par lui,
- * après le travail de modification entreprise par l'analyste sur le ou les graphes conceptuels, on le demande s'il souhaite de modifier le(s) doc.config.-thema concerné(s) ou s'il souhaite de créer un nouveau (note: au cas où l'analyste choisit l'option de modification, on le demande à la fin de la session s'il veut sauvegarder et stocker uniquement le doc.config.-thema modifié ou aussi les doc.config.-thema originaires),
- * on crée pour l'analyste un nouveau doc.thema avec les concepts du graphe conceptuel convertis en thèmes organisés séquentiellement et hiérarchiquement,
- * on lui propose de consulter des doc.thema déjà existants et on lui permet, le cas échéant, d'importer des quasi-phrases de doc.thema existants dans le sien nouvellement créé,
- * l'analyste introduit ensuite dans son doc.thema les quasi-phrases provenant de son doc.texte,
- * à la clôture de la session, on propose à l'analyste de sauvegarder et de stocker son doc.thema.

B) Intérêts et Objectifs du doc.thema

Voici quelques possibilités d'utilisation du doc.thema:

- a) stockage et archivage pas seulement de quasi-phrases sémantiquement non-structurées d'une thématique mais aussi de thèmes spécifiant la thématique,
- b) possibilité de comparaison et d'évaluation de chaque thème d'une thématique avec les quasi-phrases correspondantes,
- c) possibilité d'évaluation de la modularité de chaque thème, c'est-à-dire de sa relative indépendance par rapport à une thématique donnée et donc de sa ré-apparition dans d'autres thématiques (cf. le thème "localisation spatiale" dans la thématique "maladie parasitaire" pouvant réapparaître dans des thématiques fort différentes),
- d) possibilité d'une mise à jour sémantiquement sélective et ciblée pouvant porter, localement, sur tel ou tel thème particulier et ses quasi-phrases tout en laissant inchangé le reste de la thématique,
- e) possibilité d'utilisation de thèmes comme critères conceptuels pour la conception de bases de données ou encore comme noms de listes, d'objets ou de règles pour la conceptions de bases de connaissances,

f) possibilité d'utilisation de la structure du doc.thema comme moyen de structuration du contenu dans la production de documents assistée par ordinateur (exemple: tel thème - tel paragraphe),

g) possibilité d'utilisation selective de thèmes d'une thematique/de plusieurs thématiques dans la production de documents assistée par ordinateur.

C) Le doc.thema comme document virtuel

On peut envisager de traiter l'unité doc.thema comme une unité virtuelle dont l'existence physique se limite aux moments dont quelqu'un en a besoin. Mais il s'agit ici d'une perspective de recherche de plusieurs années.

La réalisation du doc.thema comme unité virtuelle présuppose:

- un système très performant permettant de supprimer le doc.texte comme unité physique (cf. le chapitre consacré au doc.texte),

- une bibliothèque importante de définitions et de graphes conceptuels (cf. le doc.config.-thema) intégrant forcément des opérations telles que l'abstraction, la spécification, la jonction de graphes, etc. ainsi que des opérations telles qu'elles apparaissent dans les différents systèmes de la logique non-monotone (logique de la révision épistémique, logique temporelle, logique modale, logique dialogique, ...),

- un système permettant de faire "communiquer" les graphes conceptuels avec les données pertinentes dans un corpus.

2.2.3) Le doc.config-thema

A) Caractérisation

Le doc.config.-thema contient la "théorie" ou encore le "standard" à l'aide de laquelle (duquel) l'analyste (le commanditaire, le spécialiste-expert, l'utilisateur, ...) apprehende une expertise.

Une théorie ou un standard est représenté dans le doc.config-thema:

- d'une part sous forme d'une définition-description à l'aide de laquelle un référent est appréhendée,

- d'autre part sous forme d'un graphe conceptuel (cf. Sowa 1984) décrivant la définition dans un système de représentation uniformisé et homogène pour lequel il existe de procédures de conversion en langages formels (logique de prédicats, logiques non-monotones, ...) (cf. Sowa 1984, Fargues et al 1986, etc).

a) la définition-description

La fonction de la définition-description dans le doc.config.-thema est plutôt de nature didactique ayant comme objectif principal de faire comprendre la théorie ou le standard à

quelqu'un qui n'a pas l'habitude de travailler avec les graphes conceptuels. C'est dans ce sens que les "phrases" de la définition doivent être construites selon les schémas formels qui leurs sont sous-jacents (cf. Sowa 1984, Stockinger 1992). Cette exigence permettra d'envisager d'ailleurs la conversion (au moins sémi-automatique) d'une définition-description en graphes conceptuels.

Une session de travail visant l'élaboration d'une définition-description peut se concevoir comme suit:

- * l'analyste demande qu'on lui ouvre un doc.config.-thema vide pour la création d'une définition-description portant,
- * on lui indique d'écrire le nom de la thématique pour laquelle la définition-description devra être valable,
- * si le nom de la thématique existe déjà dans la bibliothèque des doc.config.-thema ou s'il y a des noms sémantiquement voisins, on lui propose de consulter les doc.config.-thema. qui peuvent l'intéresser (note: dès la constitution d'une bibliothèque des doc.config.-thema, il faut penser à l'élaboration d'une structure -d'un schéma directeur - de cette bibliothèque qui sera en fait une métathéorie - une "métaclasse" - des doc.config.-thema à laquelle devra être associé un petit analyseur sémantique),
- * si l'analyste le souhaite, on lui édite (et, au cas échéant, imprime) les différentes définitions disponibles,
- * après avoir consulté les définitions disponibles (ou s'il n'y en pas ou bien encore s'il n'en veut pas), l'analyste doit indiquer qu'il est prêt de commencer sa modélisation conceptuelle,
- * on lui ouvre un espace de travail et on lui demande s'il veut travailler ou non simultanément à partir d'un ou de plusieurs doc.config.-thema,
- * si l'analyste veut travailler à partir d'un ou de plusieurs doc.config.thema, on les affiche ou (s'il y en plusieurs) on les lui tient à sa disposition (ils sont activés),
- * l'analyste élabore sa définition-description et demande, à la fin, de l'importer dans le doc.config.-thema vide,
- * on importe la définition-description dans le doc.config.-thema, le demande s'il veut le sauvegarder et stocker dans la bibliothèque des doc.config.-thema,
- * on le demande également s'il veut supprimer ou conserver les doc.config.-thema modifiés lors de son travail de modélisation conceptuelle.

b) le graphe conceptuel

Il est essentiel de bien comprendre que le graphe conceptuel figurant dans un doc.config.-thema ne doit pas seulement être un objet graphique dont l'existence se limiterait à la "page" où il apparaît dans le doc.config.-thema. *Le graphe conceptuel est, bien entendu, une visualisation graphique (et typographique) d'une base de connaissances dont il fait partie et à laquelle on a accès grâce à lui dans le système de doc.texte, doc.thema et doc.config.-thema.* Entendu dans ce sens, le graphe conceptuel constitue une interface graphique entre l'utilisateur-l'analyste et la base de connaissances.

Ceci dit, il faut néanmoins laisser ouverte la possibilité de considérer le graphe conceptuel comme pur objet graphique qui sera stocké dans la bibliothèque des doc.config.-

thema où il pourra être consulté avec d'autres graphes pertinents pour une application donnée en vue de la constitution d'une base de connaissances.

Outre sa fonction d'interface graphique, le graphe conceptuel joue un rôle actif dans la constitution du doc.thema dans la mesure où le doc.thema est constitué à partir de sa structure, c'est-à-dire ses concepts, relations et règles (cf. le chapitre consacré au doc.thema).

Une session de travail peut se concevoir comme suit:

- * après avoir créé une nouvelle définition-description, l'analyste demande qu'on lui ouvre le doc.config.-thema contenant celle-ci pour la création d'un graphe conceptuel,
- * on lui indique d'écrire le nom de la thématique pour laquelle la définition-description devra être valable,
- * s'il existe déjà un ou plusieurs graphes conceptuels plus ou moins correspondants dans la bibliothèque des doc.config.-thema on lui propose de consulter les graphes conceptuels qui peuvent l'intéresser (note: dès la constitution d'une bibliothèque des doc.config.-thema, il faut penser à l'élaboration d'une structure -d'un schema directeur - de cette bibliothèque qui sera en fait une métathéorie - une "métaclasse" - des doc.config.-thema à laquelle devra être associé un système d'inférence conceptuelle permettant de comparer et dévaluer les graphes conceptuels); s'il existe un graphe conceptuel semblable dans une base de connaissances, on l'indique à l'analyste,
- * si l'analyste le souhaite, on lui édite (et, au cas échéant, imprime) les différents graphes conceptuels disponibles dans la bibliothèque ainsi que celui (ceux) qui se trouve(nt) dans la (les) base(s) de connaissances déjà constituée(s),
- * après avoir consulté les définitions disponibles (ou s'il n'y en pas ou bien encore s'il n'en veut pas), l'analyste doit indiquer qu'il est prêt de commencer sa modélisation conceptuelle,
- * on lui ouvre un espace de travail et on lui demande s'il veut travailler ou non simultanément à partir d'un ou de plusieurs doc.config.-thema,
- * si l'analyste veut travailler à partir d'un ou de plusieurs doc.config.thema, on les affiche ou (s'il y en plusieurs) on les lui tient à sa disposition (ils sont activés); si l'analyste souhaite de modifier un graphe conceptuel qui est activé dans une base de connaissances, on en fait une copie et on lui propose la copie de ce graphe pour la modélisation conceptuelle,
- * on propose à l'analyste un menu graphique contenant les composants constituant un graphe conceptuel (boîte, flèche, cercle, encadré, symboles logiques spéciaux, jeux de police, ...),
- * l'analyste élabore son graphe conceptuel et demande, à la fin, de l'importer dans son doc.config.-thema,
- * on importe le graphe conceptuel dans le doc.config.-thema, le demande s'il veut le sauvegarder et stocker dans la bibliothèque des doc.config.-thema,
- * on le demande également s'il veut supprimer ou conserver les doc.config.-thema modifiés lors de son travail de modélisation conceptuelle.

Notons encore que le document de la configuration thématique doit pouvoir se différencier, si besoin en est, en un document de configuration générique et en des documents de configurations spécifiques, i.e.: la configuration "agriculture" telle qu'elle est

utilisée par un auteur, un document, etc. et les configurations "agriculture dans un tel ou tel endroit", "agriculture de tel ou tel produit particulier", etc. En ce qui concerne les configurations spécifiques, il ne s'agit pas seulement de spécifier tel ou tel concept appartenant au niveau de la définition canonique mais aussi de la possibilité de voir apparaître de nouveaux concepts se comportant d'une certaine façon comme des paramètres locaux propres à la configuration spécifique en question.

On aura donc:

- des *documents de configuration thématique générique* (ex.: doc. "agriculture selon le Baedeker")
- des *documents de configurations spécifiques* (ex.: doc. "agriculture insulaire" selon le Baedeker", "agriculture monoculture selon le Baedeker", ...)

B) Intérêts et Objectifs du doc.config.thema.

Voici quelques possibilités d'utilisation du doc.config.-thema:

- a) comparaison et évaluation de différents "standards" ou "théories" à propos d'un référent,
- b) portabilité d'une méthodologie conceptuelle d'une application vers une autre du même type,
- c) jonction de deux ou plusieurs méthodologies conceptuelles pour des applications peu génériques,
- d) possibilité d'utilisation pour la conception de bases de données et de connaissances,
- e) possibilités d'utilisation pour la conception de bases de données multimédias (iconographiques, filmiques, ...)
- f) possibilité d'utilisation pour des systèmes intégrant plusieurs bases de connaissances,
- etc.