Développement d'un outil terminologique multilingue pour l'e-Learning

Cahier des charges

Mokhtar Ben Henda & Henri Hudrisier

Version 1.0 **Mai 2010**

Table

1. Cadre général

L'AUF et Cartago se sont investis depuis 2000 dans la participation active au soutien de la diversité culturelle et linguistique dans les normes des TICE au sein du WG1 de l'ISO/CEI JTC1 SC36. La prise en charge de l'animation du WG1 offre à la délégation AUF une occasion importante pour accomplir une série d'actions dans ce sens :

- rendre efficace, réellement, largement multilingue et multi-écriture les travaux de terminologie du WG1,
- réaliser une gestion intégrée des corpus de textes normatifs sur les TICE,
- aboutir à une gestion intégrée des textes réglementaires et pédagogiques dans de multiples langues et nations référents à ces collections de concepts, liés à des termes dans leurs langues et contextes,
- proposer aux partenaires du WG1 et de l'AUF un outil de travail terminologique multilingue conforme aux normes internationales.

2. Cadre technique

La délégation AUF travaille dans un cadre technique de développement de normes e-Learning qui nécessite un ensemble d'outils terminologiques lui permettant de faire face aux difficultés à fonctionner avec les autres WGs au sein du SC36 et avec les différentes délégations nationales dans des projets connexes. Les choix actuels de la délégation AUF convergent vers les solutions suivantes :

- Une application informatique multilingue libre et ouverte pour la gestion intégrée des corpus de textes normatifs sur les TICE et de textes réglementaires et pédagogiques dans plusieurs langues,
- Un environnement collaboratif d'accès et de partage de ressources terminologiques avec les partenaires du WG1 et des délégations nationales partenaires dans les projets AUF et Cartago,
- Un modèle de DTD pour une base de données XML selon des normes terminologiques en vigueur, particulièrement TMF (norme ISO 10241:1992).

3. Cadre normatif

Toutes les actions de la délégation AUF à propos de la terminologie multilingue normalisée s'inscrivent dans le cadre référentiel des normes suivantes :

	Norme	Titre	Auteur
1.	ISO 704:2009	Travail terminologique -- Principes et méthodes	TC37/SC1
2.	ISO 1087-1:2000	Travaux terminologiques -- Vocabulaire -- Partie 1: Théorie et application	TC37/SC1
3.	ISO 10241 :1992	Normes terminologiques internationales -- Élaboration et présentation	TC 37/SC 2
4.	Geneter 10241-Annexe C	Normes terminologiques internationales -- Élaboration et présentation-modèle générique. http://www.geneter.org/docs/tutorialGeneter10241_V01.htm	TC 37/SC 2
5.	ISO 15188:2001	Lignes directrices pour la gestion de projets de normalisation terminologique	TC 37/SC 2
6.	ISO/IEC 12620:2009	Terminologie et autres ressources langagières et ressources de contenu -- Spécification de catégories de données et gestion d'un registre de catégories de données pour les ressources langagières	TC 37/SC 3
7.	ISO/IEC 16642:2003	Applications informatiques en terminologie -- Plate-forme pour le balisage de terminologies informatisées	TC 37/SC 3
8.	ISO 11197	Registres de métadonnées (RM)	JTC 1/SC 32
9.	ISO/IEC2382-36:2008	Vocabulaire -- Partie 36: Apprentissage, education et formation	JTC 1/SC 36

4. Les besoins exprimés

La situation actuelle autour de l'activité de la délégation AUF et de l'Alliance Cartago montre d'évidence que seule une gestion intégrée du modèle terminologique (ISO 10241:TMF) avec celle des textes normatifs et réglementaires dans une bibliothèque numérique largement multilingue permettrait de résoudre réellement les questions de liaisons et de synergies qui se posent concrètement à l'équipe des experts du WG1. L'environnement du modèle Greenstone (http://www.greenstone.org/) nous parait propice pour les raisons suivantes :

- Greenstone est un ensemble de logiciels dont le but est de donner accès à des collections d'informations constituant une bibliothèque numérique,
- C'est un logiciel libre (Open Source) sous licence publique générale de GNU (GPL),
- Il tourne sur toutes les versions des systèmes d'exploitation Windows, Unix/Linux et Mac OS-X. Il est compatible avec les serveurs Apache,
- Il est hautement interopérable grâce à sa conformité aux normes actuelles, particulièrement celles des archives ouvertes (OAI-PMH),
- Il a des capacités d'imports/exports de collections dans le format METS,
- Il est conforme aux schémas de métadonnées de Dublin Core et de RFC 1807 (format de notices bibliographiques),
- Son éditeur interne « *Greenstone's Metadata Set Editor* » permet aussi de développer de nouveaux schémas de métadonnées ersonnalisés
- Des "Plug-ins" sont également utilisés pour ingérer les métadonnées externe préparé sous différents formes, particulièrement en XML, MARC, CDS/ISIS, ProCite, BibTex, reporter, OAI, DSpace, METS

- *Greenstone* est multimédia. Il intègre les documents en formats texte (PDF, PostScript, Word, RTF, HTML, texte, latex, des archives ZIP, Excel, PPT, Courrier électronique, du code source) ; tous les formats de documents graphiques (y compris les GIF, JIF, JPEG, TIFF), les documents audio et vidéo (MP3, MPEG, MIDI, AVI etc.),
- Greenstone est multilingue. Son interface est disponible en arabe, arménien, bengali, catalan, croate, tchèque, chinois (simplifié et traditionnel), néerlandais, anglais, persan, finnois, français, galicien, géorgien, allemand, grec, hébreu, hindi, indonésien, italien, japonais, le kannada, le kazakh, kirghiz, le letton, le Maori, le mongol, le portugais (BR et versions PT), russe, serbe, espagnol, thaï, turc, ukrainien, vietnamien.

Greenstone offre ainsi un environnement de travail collaboratif multilingue, libre et ouvert, qui saurait aider la délégation AUF à atteindre les objectifs annoncés ci-haut.

De ce fait, nous considérons nécessaire d'étudier la faisabilité des alternatives stratégiques et techniques suivantes et de les appliquer rapidement pour réaliser ces objectifs :

1. Vérifier la capacité d'intégration de l'application *Greenstone* pour toute extension XML d'un module terminologique conforme au schéma TMF. Cela nécessite une analyse plus poussée des possibilités techniques de *Greenstone* pour l'intégration de modules XML externes, pour la création de flux de données de et vers ses bases de données et pour l'exportation des données dans des formats supplémentaires à ses formats natifs,
2. Vérifier la capacité de *Greenstone* à accepter la fouille de texte dans ses collections de données numériques multilingues et multi-écritures,
3. Développer en langage XML et tester une (ou plusieurs) application(s) terminologiques du type TMF compatible avec l'environnement *Greenstone*,
4. Faire en sorte que l'outil à développer et le projet à construire puissent être déclinés en autant de langues et de systèmes d'écriture au-delà du cadre de la Francophonie (Sud-Est asiatique, Europe, rassemblement des NBs et des Liaisons du SC36…),
5. Développer des routines de collecte et de versement systématiques de données terminologiques dans une base de données multilingue centralisée,

5. Gestion des priorités

Ces besoins s'échelonnent sur plusieurs phases de réalisation. Ils seront accomplis en fonction d'un ordre de priorité qui commence par la conception d'une application informatique évolutive. Cette application sera enrichie à chaque étape par des fonctionnalités plus avancées et des ressources terminologiques plus variées.

La priorité en cette phase initiale se résume en ce qui suit :

1. Développer une base de données avec une interface intuitive en XML, capable de gérer les terminologies multilingues disponibles de la norme ISO 2382-36,
2. Veiller à ce que la base de données soit conforme aux normes citées précédemment et que la réalisation des terminologies se fasse dans un cadre méthodologique largement défini et en partie normalisé,
3. L'application définie en point (1) devra avoir la capacité :
    a. De produire des éditions de vocabulaires multilingues parallèles (dans un premier temps en anglais-français) sur le modèle de la norme ISO/IEC 2382-

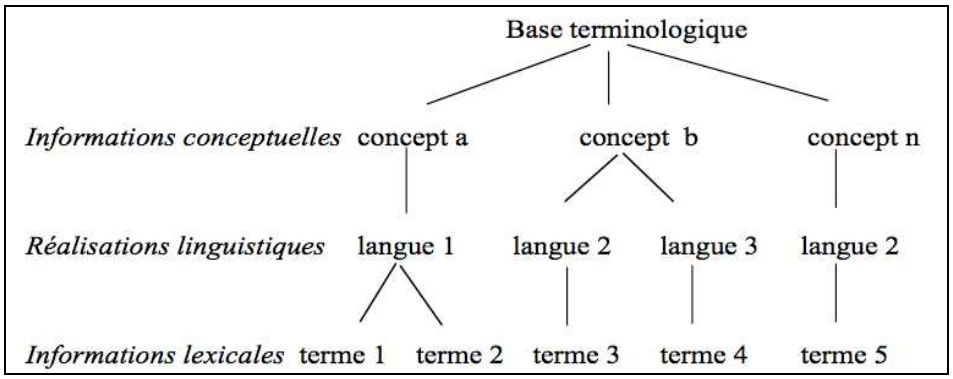
36,

b. De générer des versements cohérents et compatibles en direction de la terminologie centralisée du JTC1 mais aussi en direction de la terminologie de l'ISO fondée sur une approche conceptuelle (onomasiologique et conforme aux normes de l'ISO TC37 cités comme références),
c. De gérer l'aspect éditorial, administratif, normatif et référentiel ainsi que la non complétude ou complétude des états de réalisation des différentes entrées terminologiques de la base (mais aussi les différentes versions globales de la base)
d. De s'insérer dans un cadre de travail en réseau collaboratif convivial et fiable qui devra communiquer avec un outil déjà existant (l'environnement *Greenstone*).

6. Spécifications techniques

L'application à définir devrait respecter les spécifications générales suivantes :

- être rédigée en langage XML,
- être munie de fonctions d'import/export en format normé ISO 2382 et en format texte formaté,
- pouvoir être intégrée à un environnement collaboratif comme celui de *Greenstone*,
- permettre la saisie des termes et des définitions en plusieurs langues selon le modèle TMF suivant :



- respecter les directives des normes en vigueur pour les opérations d'entrée, de sortie et d'administration des données terminologiques. Ces directives sont décrites ci-après :

6.1. L'entrée des données

L'entrée de données est tributaire de deux conditions clés :
- la structure de la base de données terminologique,
- les mécanismes de saisie.

A. Structure de la base terminologique

La structure d'une base de données terminologique doit être conforme à la norme ISO 10241 :1992 (TMF). Elle doit reproduire le schéma suivant :

❖ **Le niveau racine « TDC » (*Terminological data Collection*)**

Il s'agit de la collection de données contenant des informations sur les concepts dans un domaine spécifique. Ce niveau hiérarchique supérieur est celui de la base terminologique elle-même. Ce niveau TDC est subdivisé en 3 sous niveaux :

1. Information générale (GI),
2. Information complémentaires (CI),
3. Entrées terminologiques (TE).

❖ **Les niveaux Informations Générales (GI) et Informations Complémentaires (CI)**

Ces niveaux (ou nœuds) constituent deux registres dans lesquels sont stockées les données référentielles importantes pour administrer ou faire fonctionner la base terminologique. Ces données n'appartiennent pas directement à la collection terminologique.

❖ **Le niveau terminologique (TE)**

Le niveau TE (Terminological Entry) constitue le cœur de l'application dans lequel s'accomplissent les opérations de saisie des données terminologiques par langue. C'est l'entrée qui contient les informations sur les unités terminologiques. Malgré son nom, c'est exclusivement le niveau du concept commun à toutes les langues. C'est à ce niveau que l'on peut décrire les caractéristiques du système (ou graphe) de concept : générique ou partitif.

Le niveau (TE) est subdivisé en 3 sections :

1. Une section de langue (LS),
2. Une section de terme (TS),
3. Une section de terme composé (CTS).

❖ **La section de langue (LS)**

La section de langue (LS) identifie la langue des ressources traitées. C'est une section de l'entrée terminologique contenant des informations relatives à la langue. Comme son nom l'indique, c'est plus une section qu'un niveau. Le LS intégre tout ce qui est dépendant des langages. Elle s'oppose clairement au niveau des concepts (TE). C'est à ce niveau hiérarchique que l'on ouvre des langues. L'ouverture d'un LS est un préalable indispensable à l'ouverture d'un TL (*Term Level*) qui est un niveau encore hiérarchiquement inférieur dans chaque langue du schéma TMF.

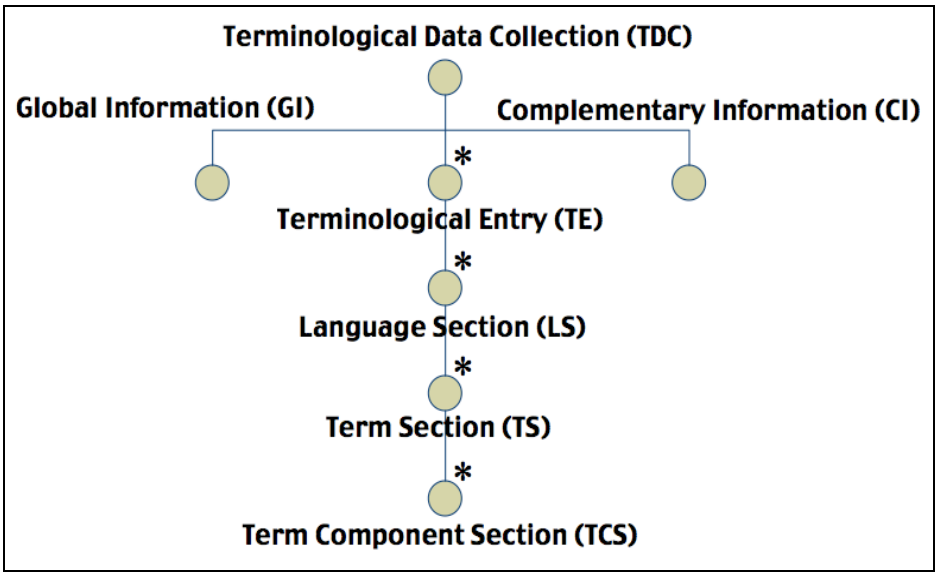
❖ **La section de terme (TS)**

Une section de termes (TS) permet de saisir les termes qui correspondent à la définition du concept saisie dans le (TE). C'est à ce niveau que s'ouvrent un ou plusieurs termes supposés être tous plus ou moins synonymes dans autant de LS qu'il y a de langues dans la base. C'est à ce niveau que peut se faire une description morphosyntaxique des termes : genre, nombre, catégorie du discours (nom, adjectif, verbe, syntagme, mots valise..), abréviation, acronyme.

❖ **La section de terme composé (TCS)**

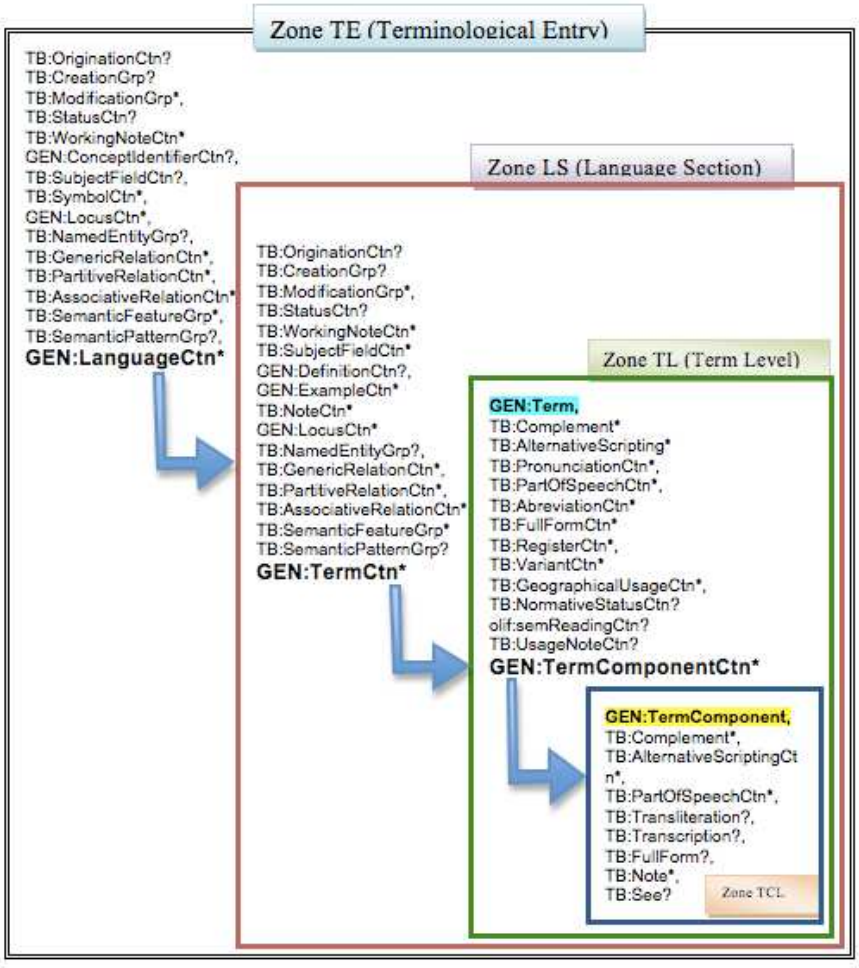
La section (TCS) permet de décrire les composants d'un terme. Dans le cas concret du SC36 et dans beaucoup de terminologies d'instances normatives, nous devons gérer une proportion importante de mots valises ou de syntagmes. Beaucoup de terminoticiens le considèrent comme un niveau de détail inutile. Formellement chaque élément d'un terme composé, dans chaque langue permet d'ouvrir un TCL (*Term Component Level*). Il est possible donc d'ouvrir autant d'items pour un terme qu'il y a de composants dans le terme. L'utilité réside en ce que ce TCL facilite une vision comparée (largement multilingue et potentiellement

sémantiquement assistée) des modes de génération des nouveaux concepts d'un champs terminologique donné (néologisme ou néonymes).



La structure du modèle TMF

L'illustration suivante schématise l'hiérarchisation entre les sections d'une entrée terminologique dans le modèle TMF



B. Les mécanismes de saisie

L'application à développer en première phase devrait permettre une saisie manuelle des données via une interface graphique conviviale. L'entrée des données devrait ensuite évoluer vers de nouvelles formes plus sophistiquées de saisie comme l'importation de corpus terminologiques externes et la fouille de texte dans des collections de documents numériques.

L'interface de saisie manuelle devrait correspondre à la structuration TMF décrite ci-avant. Elle devrait permettre l'entrée des données sur 4 niveaux hiérarchiques imbriqués :

- une entrée terminologique de niveau concept,
- une entrée linguistique de niveau variante linguistique,
- une entrée terminologique de niveau terme dans une langue,
- une entrée terminologique de niveau terme composé dans une langue.

La composition de chaque niveau d'hiérarchie est définie dans la norme ISO 10241:1992.

Des exemples d'écrans de saisie dans une application similaire sont donnés en ressources complémentaires décrites sous le point 7.

L'interface doit respecter les critères d'accessibilité permettant l'identification facile des différentes zones de saisie (répartition écran, couleurs, etc.). Elle devrait également permettre d'utiliser des valeurs de données prédéfinies dans des listes fermées déroulantes.

6.2. Les éléments de sortie (les produits)

L'édition des listes devrait correspondre dans sa structure à la norme ISO 2382-36. Elle doit se conformer, pour cela, aux spécifications d'élaboration et de présentation définies par la normes ISO 10241:1992. Les deux textes de normes sont fournis en complément à ce cahier des charges.

Les listes terminologiques produites doivent être paramétrables en fonction du nombre des langues saisies et/ou choisies en sortie.

Ci-après sont présentés des modèles de listes monolingues, bilingues et multilingues de données terminologiques correspondant au corpus de la liste ISO 2382-36:2008.

36 Technologies de l'information pour l'apprentissage, l'éducation et la formation

36.01 Termes généraux

36.01.01
apprentissage
acquisition de connaissances, d'habiletés ou d'attitudes

36.01.02
formation
développement d'habiletés ou de connaissances à l'aide d'**activités d'apprentissage** procédurales centrées sur une application spécifique

36.01.03
apprentissage sur le Web
apprentissage en ligne utilisant les technologies Web et Internet

Liste monolingue en français

36 Information technology for learning, education and training	**36 Technologies de l'information pour l'apprentissage, l'éducation et la formation**
36.01 General terms	**36.01 Termes généraux**
36.01.01 **learning** acquisition of knowledge, skills or attitudes	**36.01.01** **apprentissage** acquisition de connaissances, d'habiletés ou d'attitudes
36.01.02 **training** development of skills and/or understanding through procedurally defined **learning activities** focused on a specific application	**36.01.02** **formation** développement d'habiletés ou de connaissances à l'aide d'**activités d'apprentissage** procédurales centrées sur une application spécifique
36.01.03 **web-based learning** **on-line learning** that uses web technologies and Internet-based technologies	**36.01.03** **apprentissage sur le Web** **apprentissage en ligne** utilisant les technologies Web et Internet

Liste bilingue anglais/français

36 Information Technology for Learning, Education, and Training.	36 Technologies de l'information pour l'apprentissage, l'éducation et la formation	36 تكنولوجيا المعلومات للتعليم والتربية والتكوين	36 학습, 교육, 그리고 훈련을 위한 정보 기술
36.01 36.01 General terms	36.01 36.01 Termes généraux	36.01 مصطلحات عامة	36.01 36.01 일반개념
36.01.01 **learning** (preferred term) acquisition of understanding, knowledge, skills and/or Attitudes	36.01.01 **apprentissage** (terme préféré) acquisition de connaissances, développement d'habiletés et d'attitudes	36.01.01 (preferred term) تعلّم (terme admis) تلقّم عملية اكتساب الفرد معلومات أو طرائق جديدة للسلوك، أو مهارات جديدة في العمل المهني	36.01.01 학습 이해, 지식, 기술, 및/또는 태도의 습득
36.01.02 **education** (preferred term)	36.01.02 **éducation** (terme préféré)	36.01.02 n, (preferred term) تربية	36.01.02

Liste multilingue anglais/français/arabe/coréen

L'édition des listes doit aussi permettre l'édition des index par langue.

Index alphabétique français

A

F

G

Index alphabétique français

L'application doit permettre aussi de publier en plusieurs formats texte : texte plat et enrichi.

6.3. Les éléments de gestion et d'administration (droits et profils d'accès)

Malgré son caractère libre et ouvert, l'application doit permettre :

- une sécurisation des données : accès par profil (admin/expert/visiteur…),
- une gestion des responsabilités d'auteur [contributeur, éditeur terminologique…], institution, NB ou LO,
- une gestion des versions d'une entrée terminologique ou de partie d'une entrée (les facettes multilingue d'une entrée terminologique obligent à distinguer ces différents niveaux dans la mesure où une nouvelle contribution linguistique doit être maitrisée par des experts de la langue en question)

7. Les ressources existantes

Pour aider à la réalisation de cette première phase du projet, nous mettons à la disposition des développeurs des ressources complémentaires qui retracent l'expérience acquise en la matière. Des projets précédents ont été élaborés pour aboutir aux résultats escomptés. Le projet a du être interrompu pour plusieurs raisons. Voici un état de l'existant comme ressources et outils pouvant aider à maîtriser les contours du projet et à accélérer sa réalisation.

- Deux types de DTD XML avec des feuilles de style ont permis de développer une application terminologique comme celle prévue par ce cahier des charges. Le premier type de DTD est générique. Il concerne la saisie et l'édition des données terminologiques dans les deux langues de base ; anglais et français. Le deuxième type de DTD et feuille de style associé a été défini pour les entrées terminologiques dans d'autres langues, en l'occurrence l'arabe et le coréen. Ces DTD et leurs feuilles de styles associées sont opérationnelles sous n'importe quel éditeur XML. Elles ont été testées sous le logiciel XMLmind. Un document technique présentant cette expérience est donné en ressources complémentaire.
- Les textes des normes nécessaires au développement de l'application figurent aussi parmi les ressources complémentaires. Ils peuvent être soumis à la demande en fonction des besoins exprimés par les développeurs.
- Des données terminologiques normalisées sont aussi disponibles. Il s'agit des listes en anglais et en français qui font l''objet de la norme ISO 2382-36 et des listes terminologiques en langue arabe et coréenne qui ont fait l'objet de l'expérience mentionnée sous XMLmind. Des listes dans d'autres langues comme le berbère ou le malgache sont également disponibles. Toutes ces ressources peuvent faire l'objet d'une expérimentation pendant la phase de développement de l'application à produire.

8. Calendrier et phasage

La demande faite prend en compte des contraintes calendaires biannuelles lourdes (celles des Plénières du SC36 tous les 6 mois en mars et septembre) et celles des projets parallèles entamés ave des partenaires dans les délégations du SC36. La crédibilité du projet exige l'établissement du calendrier de réalisation suivant :

Phase 1 du projet de création d'un outil terminologique		
Opération		Echéance
	- Analyse de *Greenstone*	15 juin 2010
	- Prototypage normalisé de l'application	
	- Développement des DTD (structure) et des CSS (interface)	15 juillet 2010
	- Développement des modèles de produits (listes)	30 juillet 2010
	- Saisie et test de données (Entrée/Sortie)	15 aout 2010
	- Validation et installation sur serveur	30 août 2010

9. Conclusion

Ce cahier des charges est partiel. Il concerne la première phase de production d'un outil de développement terminologique multilingue conforme aux normes internationales en vigueur.
Les phases suivantes du projet seront définies en fonction des résultats obtenus de la phase actuelle et des retours d'expérience observés pendant la mise en application de ce projet.
Parmi les objectifs techniques à atteindre sur le long terme (en phases successives) :

- Le développement de modules de fouille de texte pour extraction de données terminotiques des collections des données,
- Le développement de passerelles avec des éditeurs de graphes conceptuels,
- La production d'ontologies de domaine pour le e-Learning,
- La conception de modules sémantiques conformes aux normes du W3C et au langage OWL pour la production de *Topic maps*.

10. Pièces jointes

Comme annoncé dans le point 7 relatif à l'analyse de l'existant, ce cahier des charges est accompagné des ressources électroniques suivantes (toutes compressées dans le fichier « ressources_complémentaires.rar ») :

	Nom de fichier	**Titre**	**Contenu**
1.	ISO_IEC_2382-36.pdf	La norme ISO 2382-36	La forme des listes normalisées des termes et des définitions e-Learning du SC36
2.	ISO_IEC_10241.pdf	La norme ISO 10241:1992	Les spécifications de présentation des listes terminologiques multilingues
3.	Multilingual_terms_project.doc	Quadrilingual set of E-Learning terminology conform to SC36 WG1 N0123 (ISO/IEC 2382-36-2)	Liste quadrilingue de termes et définition e-Learning produites par la délégation AUF sous XMLmind
4.	IsoGen2382sc36.doc	IsoGen 2382sc36: Xml edition	Caractéristiques techniques du modèle IsoGen2382 développé pour Cratago en 2007 sous XMLmind et Genetrix
5.	IsoGen2382sc36.rar	divers	Fichiers de l'application « IsoGen2382sc36 » développée sous XMLmind pour produire les listes terminologiques e-Learning multilingues (Voir fichier n° 4 pour les modalités d'installation et les fichiers du dossier « doc » dans cette archive RAR pour l'installation des DTD et CSS pour l'arabe et le coréen).