Encyclopédie numérique sur les Nations Unies

Patrick Bellot
Marine Campedel

2010D003

Janvier 2010

Département Informatique et Réseaux
Groupe IC2 : Interaction, Cognition et Complexité

Département Traitement du Signal et des Images
Groupe TII : Traitement et Interprétation des Images

ENCYCLOPEDIE NUMERIQUE SUR LES NATIONS UNIES

Rapport rédigé par Télécom ParisTech pour l'IHENU

Auteurs

Patrick Bellot et Marine Campedel

{bellot,campedel}@telecom-paristech.fr

Janvier 2010

Avant-propos

L'objet de ce rapport est de décrire ce que serait l'Encyclopédie idéale, en termes de fonctionnalités pour ses acteurs, et de commencer une étude de faisabilité technique en considérant les technologies actuellement disponibles ou émergentes.

Olivier Auber, responsable du Web du département Informatique et réseaux et consultant en matière de prospective des usages culturels du réseau, a apporté sa contribution au rapport dans le domaine du travail collaboratif.

Samuel Tardieu et Pierre Senellart, enseignants-chercheurs du département Informatique et réseaux, ont apporté leurs concours aux réflexions sur l'Encyclopédie.

Ce rapport reprend des éléments du compte-rendu de la réunion du 18 novembre 2009.

Il analyse les fonctionnalités envisageables pour l'Encyclopédie.

Puis il énumère un ensemble de propositions et d'options techniques non contraignantes dans lesquelles on pourra faire un choix. Ces options couvrent les différents domaines techniques de l'Encyclopédie.

Dans cette étude, on effectuera une distinction technique entre le système de requêtage et de distribution de documents multimédia du système d'édition collaborative. En effet, ces deux systèmes reposent sur des architectures logicielles très différentes.

Le lien entre les deux systèmes est un mécanisme de publication permettant de publier, c'est-à-dire mettre à disposition et indexer, un document validé dans le système d'édition collaborative.

Nous envisageons également une réponse aux appels à projets ANR.

Patrick Bellot et Marine Campedel

Table des matières

1. Introduction

L'Institut des hautes études sur les Nations unies a pour objectif, depuis sa création, d'encourager la recherche sur les Nations unies, le multilatéralisme et la gouvernance mondiale.

L'excellence de ses recherches vise à promouvoir la connaissance du système onusien - sa structure et ses organes, son histoire et sa construction, ses activités opérationnelles ainsi que ses relations diplomatiques.

L'IHENU s'attache également à diffuser ces savoirs tant auprès des experts que du grand public, dans le but de rendre ces connaissances accessibles à tous les citoyens du monde intéressés par l'Organisation mondiale.

Dans cette perspective, l'Encyclopédie numérique sur les Nations unies, projet d'envergure, aspire à la création sur Internet d'un portail de référence sur les Nations unies. Unique en son genre, en sept langues - anglais, français, espagnol, chinois, russe, arabe et portugais -, ce portail, gratuit, visera un public large et hétérogène tels que les chercheurs, les hauts fonctionnaires, les journalistes, les enseignants ou encore les étudiants.

Le principal objectif est la réalisation d'une encyclopédie numérique sur les Nations Unies comprenant un système d'édition collaboratif. Ils doivent être opérationnels début 2013, c'est-à-dire dans 3 ans. Le système d'édition doit être prêt dans 18 mois.

1.1. Identification de besoins - Encyclopédie vs bibliothèque numérique

Le projet de création d'une encyclopédie des Nations Unies ne se réduit pas à une bibliothèque numérique regroupant les documents des Nations Unies, mais a pour objectif de diffuser de la connaissance liée aux Nations Unies.

Le contenu de cette encyclopédie doit avant tout permettre de regrouper des informations d'ordre général sur l'histoire, les métiers, les thématiques traitées par les Nations Unies (en matière de droit ou d'humanitaire par exemple) et permettre également des recherches approfondies de documents. Les ressources documentaires visées sont au nombre de 50000 à 120000. Il s'agit de textes, traités, résolutions (disponibles dans une base de données annexe), de coupure de journaux, d'iconographies, de fichiers audios et vidéos (interviews de personnalités, par exemple). Cette encyclopédie a pour ambition de devenir la source la plus importante d'information sur les Nations Unies. Elle se veut une revue au sens où le site sera un état des lieux des connaissances en matière de gouvernance globale (ce qui a fait l'objet de recherches approfondies et de chantiers scientifiques qu'il reste à mener).

En ce sens l'Encyclopédie est une bibliothèque numérique, dans laquelle des outils de stockage et de récupération d'information seront nécessaires.

Mais l'Encyclopédie se veut également source de connaissance. Elle doit permettre i) de s'adapter à ses usagers qui sont divers (chercheurs, étudiants, journalistes, diplomates, grand public), ii) de mettre en relation différentes sources d'information (internes et externes à l'Encyclopédie), afin de mettre à

jour des éléments d'intérêt (sources de connaissance) pour l'usager, iii) au travers d'une interface conviviale et intuitive.

L'accès à l'encyclopédie est actuellement envisagé sous forme d'un client web. Outre l'encyclopédie « en ligne » sur internet, une plateforme physiquement accessible doit être installée à New York et à Genève, afin de permettre des visites guidées des personnes, au moyen d'écrans plats par exemple.

1.2. Pour qui et par qui ? Identification des acteurs de l'encyclopédie

Nous différencions 3 grandes catégories d'acteurs dans ce projet d'Encyclopédie :

- ✓ Les acteurs techniques : l'IHENU, Télécom ParisTech, d'éventuels partenaires académiques et les entreprises qui développeront la solution technique, ainsi que le personnel qui maintiendra la solution font partie de cette catégorie.

- ✓ les contributeurs de contenu (rédacteurs) : ils appartiennent à des groupes thématiques. Ces groupes doivent produire de manière collaborative des articles, associer des ressources documentaires (multimédia) et prévoir des étapes de révision et de traductions. Dans sa phase initiale, chaque groupe contient 7 à 9 personnes. Une fois l'encyclopédie en ligne, des chercheurs continueront de contribuer au contenu. Un protocole de rédaction doit être mis en place afin de garantir la lisibilité des contenus (tant dans le fond que dans la forme).

- ✓ Les usagers (internautes, visiteurs des plateformes physiques) : le public visé est composé de chercheurs, des étudiants, des journalistes, des diplomates et le grand public, ayant donc des attentes très différentes de cette encyclopédie. Les objectifs de fréquentation du site sont : atteindre 2 millions de visiteurs uniques à la fin de la première année, avoir entre 100 000 et 300 000 visiteurs par mois et près de 5 millions de documents consultés.

1.3. Organisation du projet

Le projet est piloté à l'IHENU par Monsieur Fouad Souak qui a prévu la mise en place de différents groupes de travail et d'un comité de pilotage. Sept groupes sont concernés par le contenu de l'encyclopédie. Ils sont répartis en différentes thématiques. Il y aura également un groupe « technologie et architecture du site » dans lequel Télécom ParisTech participera au niveau du conseil et de la définition du cahier des charges. Ce rapport est une première ébauche.

Une mise en ligne en 2013 est visée. Cette date correspond en outre à l'année où Marseille sera capitale européenne de la culture. Pour les groupes de contenu, 18 mois sont prévus avant de prévoir une interaction avec la plateforme. Ces 18 mois doivent donc permettre la mise en place des outils nécessaires à la mise en ligne des contenus, à leur révision, leur traduction, à l'addition de ressources documentaires et à la navigation.

2. Les fonctionnalités (idéales) de l'Encyclopédie

2.1. Fonctionnalités pour l'équipe technique

Le rôle de l'équipe technique est :

- ✓ de concevoir techniquement l'Encyclopédie : produire un cahier des charges est une première étape nécessaire et importante, qui sera suivie d'une phase de développement, puis de maintenance (évolutive).
- ✓ de gérer le stockage des documents qui forment le contenu de l'Encyclopédie.
- ✓ d'assurer l'accessibilité de l'Encyclopédie.
- ✓ de mettre en place les plateformes physiques à New York et Genève.
- ✓ d'assurer l'évolution technologique (maintenance, correction de bugs, veille et mise en place de nouvelles fonctionnalités).

Afin de faciliter ces actions, l'équipe doit se doter elle-même des fonctionnalités suivantes :

1. Gestion de projet.
2. Veille technologique : cette veille doit permettre de rester en alerte sur les outils, solutions technologiques efficaces afin de rendre l'Encyclopédie plus performante. Ces outils doivent pouvoir être testés préalablement sur une version de l'Encyclopédie qui n'est pas visible des usagers, mais qui doit en être très proche.

   Ceci nécessite la mise en place de critères d'évaluation des performances de la plateforme, vraisemblablement au niveau de chacune des fonctionnalités envisagées. Cela pourrait constituer un verrou technologique pour des fonctionnalités innovantes, telles l'adaptation du contenu de l'Encyclopédie à ses usagers.
3. Développement collaboratif, suivi de versions : versions des logiciels utilisés mais également des développements spécifiques entrepris pour l'Encyclopédie ; mise en place de règles de programmation et de commentaires des codes développés.
4. Solutions de stockage et accessibilité : l'Encyclopédie ayant pour objectif de centraliser des informations dont le nombre devrait croître et ainsi d'attirer de nouveaux usagers, il semble important de se poser des questions en terme de passage à l'échelle, tant en nombre et variété des documents que des utilisateurs. Ceci pourrait constituer un nouveau verrou, en particulier sur des problématiques de fouille de documents multimédia.
5. Gestion du matériel : ceci concerne les plateformes physiques et leur maintenance au quotidien. On peut envisager des fonctionnalités de maintenance à distance, avec des dépannages occasionnels sur place.

6. Gestion des retours d'expérience : les différents acteurs de l'Encyclopédie peuvent faire des retours sur leurs usages (déclarations de bugs, satisfaction/insatisfaction, suggestion, etc.) ; ceux-ci doivent donc être accessibles, lisibles et exploitables par l'équipe technique.

2.2. Fonctionnalités pour les contributeurs de contenu

Les contributeurs de contenu appartiennent aux équipes thématiques. Leur rôle est de définir, rédiger et publier le contenu de l'Encyclopédie. Ils doivent donc disposer des fonctionnalités suivantes :

1. Rédaction collaborative d'articles et suivi de versions (temporelles et linguistiques) : cette fonctionnalité est essentielle pour permettre un travail de groupe, avec des relectures, des révisions et des traductions des articles.

2. Fouille des bases documentaires (multimédia) existantes afin de faire des liens avec les articles en cours de rédaction : « faire des liens » sous-entend que les documents liés sont associés à des actions spécifiques à définir par l'éditeur de l'article (position des médias, taille/résolution des images apparaissant dans l'article, précision d'accès à une vidéo ou à un enregistrement à l'aide d'un clique de souris par exemple, …)

3. Pré-visualisation de l'article mis en ligne avant publication, afin d'ajuster la mise en page du contenu (en particulier entre le texte et les ajouts documentaires).

4. Publication : action de mise en ligne de l'article, avec les documents liés. Ceci pourrait impliquer une phase de vérification par un tiers avant une publication effective.

5. Possibilité de donner un avis sur les outils à disposition, cet avis est traité par l'équipe technique.

2.3. Fonctionnalités pour les usagers de l'Encyclopédie

Les usagers sont donc les utilisateurs finals de l'Encyclopédie. Ils font partie du grand public ou de l'organisation des Nations Unies, ils peuvent être internautes de passage, journalistes, étudiants, chercheurs ou diplomates, … Les scenarios d'exploitation de l'Encyclopédie peuvent donc être très variés et feront l'objet d'un autre rapport. Cependant, ces scénarios s'appuient sur des fonctionnalités communes, présentées ci-après :

1. Navigation libre : l'usager peut « voyager » librement dans l'encyclopédie, c'est-à-dire parcourir à son gré les différentes pages de l'Encyclopédie. Une navigation classique consiste à cliquer sur des liens successifs. Il faut permettre à l'utilisateur de se faire une image mentale consistante du contenu de l'Encyclopédie afin de rendre ses propres stratégies de navigation efficaces. Nous pourrions imaginer une façon innovante de naviguer en présentant une cartographie adaptative du contenu du site, s'affichant avec une granularité adaptée aux actions (clic souris et autres

mouvements de souris) de l'internaute et à l'interface qu'il utilise (type d'écran, téléphone mobile, etc.). Ceci pourrait constituer un verrou intéressant à lever avec l'aide d'ergonomes et de chercheurs spécialisés dans l'interaction homme-machine.

2. Navigation orientée : l'usager peut naviguer selon certains scénarios prédéfinis, par exemple à but pédagogique.

3. Visualisation : l'Encyclopédie contient des documents multimédia et doit donc permettre leur visualisation à partir de tout navigateur. Ceci implique que les documents coûteux au chargement soient présentés sous une forme dégradée (par exemple, présentation des images et vidéos sous forme de vignettes), qui donne ensuite accès à une version complète. Ce mode de visualisation permet en outre de mieux contrôler l'accès à des documents éventuellement payants.

4. Requêtage : l'usager peut consulter l'Encyclopédie avec des attentes précises. Il doit donc pouvoir, à l'aide d'une interface adaptée, exprimer sa requête. Les résultats de ses requêtes peuvent être des objets multimédia, du texte, des liens vers des sources d'information extérieures, … La requête et/ou le résultat de la requête peuvent être conservés en vue d'une réutilisation ultérieure.
   Les modes de requêtage doivent être aussi naturels que possible : l'usage de mots clés écrits est assez classique ; on pourrait imaginer du langage naturel (ce qui reste encore un problème ouvert) écrit ou parlé (avec une reconnaissance de parole), des images similaires à ce qui est recherché, … Ceci peut constituer un verrou intéressant à lever, dès qu'il s'agit d'un requêtage multimédia et sémantique.

5. Téléchargement : certains documents peuvent être simplement visualisés, d'autres peuvent éventuellement être récupérés par l'usager.

6. Adaptation aux utilisateurs : l'Encyclopédie connaît ses usagers (par le moyen d'une modélisation individuelle ou par groupe d'usagers) et leurs comportements ; elle peut donc proposer une vitrine adaptée à chaque utilisateur, tant du point de vue du contenu (ex = suggestion de rubriques d'intérêt) que de la forme (ex=facilité d'accès aux fonctionnalités les plus utilisées).

7. Possibilité de donner un avis sur les outils à disposition, cet avis est traité par l'équipe technique.

8. Moyens de communication et d'expression entre utilisateurs de l'Encyclopédie : l'Encyclopédie propose des forums de discussion et une mise en relation de personne partageant des intérêts thématiques.

2.4. Scenarios de navigation pour les clients

Un site web doit être structuré en fonction d'une stratégie. Des scénarios de navigation doivent être conçus dans le but de conduire l'utilisateur à suivre un chemin dans le site. Cependant il ne faut pas emprisonner le visiteur et le laisser libre de s'en aller ou de changer de rubrique à tout moment. Il sera bien entendu nécessaire de définir les scénarios de navigation des clients de l'Encyclopédie en distinguant les clients libres effectuant des visites occasionnelles des clients professionnels susceptibles de s'inscrire sur le site.

Pour les clients inscrits, il faudra prévoir des fonctionnalités avancées de personnalisation du site et des interfaces. Par exemple :

- Historique des recherches et accès rapide aux documents déjà consultés.
- Accès à des fonctions de recherche plus complexes.
- Fonction de prise de notes associées à un document.
- Lettre d'information mensuelle envoyée par email.
- Communautés virtuelles d'utilisateurs avec blogs thématiques.
- Etc.

Un ensemble de modules présentant des fonctions évoluées accessibles aux utilisateurs inscrits pourrait faire partie d'un projet ANR.

2.5. Design des interfaces utilisateurs

La qualité des interfaces utilisateurs joue un rôle très important dans la fréquentation des sites Web. Une interface peu attrayante ou une ergonomie défaillante et c'est une grande partie des clients qui s'en va. Dans notre cas, une interface utilisateur malvenue ferait fuir le grand public et seuls les professionnels bien obligés en seraient clients.

Il importe de définir l'agencement de la page, son dimensionnement, le type de navigation ainsi qu'une charte graphique.

Parmi les critères pertinents pour une interface utilisateur :

- L'élégance bien sûr.
- La standardisation par rapport aux normes données par le W3C, voir le site en anglais **http://www.w3.org/TR**. Le site de l'IHENU remplit parfaitement ce critère.
- La standardisation par rapport à la mode actuel des sites Web produits par des outils répandus de gestion de site (CMS). Une mise en page trop différente de ces standards peut donner une impression de site mal bricolé. Ici aussi, le site de l'IHENU remplit parfaitement ce critère.
- La légèreté visuelle et technique du site. Il ne faut pas cela mette trop longtemps à se charger parce que l'on a inclût une très jolie partie écrite en Flash. N'oublions pas que le site est international et que ce qui serait acceptable dans les pays du Nord pourrait l'être beaucoup moins dans les pays du Sud où l'infrastructure Internet est moins efficace.

- Il ne faut pas que l'interface soit trop complexe à utiliser pour le grand public. On peut néanmoins proposer des interfaces plus complexes pour les professionnels.
- Cela ne doit pas empêcher quelques interfaces plus lourdes comme des cartes cliquables afin d'appuyer l'image novatrice du site.

La conception de l'interface utilisateur pourrait très certainement faire partie d'un éventuel projet ANR sur l'Encyclopédie. On pourrait, en particulier, inclure des interfaces d'accès pour les personnes qui ont un handicap visuel. Egalement, des interfaces pour les PDA et autres SmartPhones.

3. Recommandation urgente

Le nom du projet et le nom de domaine du site de l'Encyclopédie sont liés. Le nom du projet d'encyclopédie peut être influencé par la disponibilité des noms de domaine. Il faut éviter de choisir un nom de domaine trop long ou trop compliqué. Le nom de domaine doit idéalement être prononçable et avoir une signification.

Il est important de choisir très vite un nom de domaine pour le site et de le déposer au plus vite, avant que l'existence du projet soit largement diffusée, sous toutes ses formes : .com, .org, .net, .biz, .fr, etc. et sous toutes ses variantes.

Certains malins achètent un grand nombre de noms de domaine. Ce procédé, le *grabbing* consiste à prévoir l'achat de noms de domaine de certains organismes ou entreprises et de les acheter avant celles-ci. Les extensions en .com, .net et .org ne sont soumises à aucun contrôle. Avec ce type de pratique, des personnes ont réussi par le passé à revendre à prix d'or des noms de domaine intéressants pour certaines compagnies (leur marque en général). Depuis, la législation a évolué et il est rare qu'un tribunal donne raison à l'escroc. Mais autant éviter les ennuis.

La vérification de disponibilité et le dépôt des noms de domaine pour l'Encyclopédie peuvent être réalisés sur le site http://www.bookmyname.com, entre autres. Cela est gratuit.

4. Gestion électronique de documents (GED)

On distingue plusieurs étapes :

- Acquisition ou production des documents.
- Classement et indexation des documents.
- Stockage des documents.
- Diffusion des documents.

5. Acquisition des documents

On distingue plusieurs types de documents :

- Les anciens documents qui sont en cours de numérisation. On parle de plusieurs millions de documents. La numérisation peut prendre plusieurs formes.
- On peut simplement scanner les documents et obtenir l'image correspondante rangée dans un format standard (PDF par exemple). Dans ce cas, l'indexation du document sera manuelle.
- On peut en plus utiliser un logiciel de reconnaissance optique de caractères (OCR). Dans ce cas, l'indexation du document pourra être automatisée.
- Les nouveaux documents que l'ONU pourrait vouloir mettre en ligne : traités, résolutions, etc. Ceux-ci ont une forme électronique qui pourra être indexée automatiquement.
- Les articles produits grâce au système d'édition collaboratif. Ceux-ci, sous forme numérique, pourront être indexés automatiquement.
- Les documents multimédia : audio, vidéo, photos. Ceux-ci pourront être indexés manuellement ou automatiquement.

Notons que les articles produits par un système d'édition collaboratif de type MediaWiki pourront avoir une forme pour la visualisation rapide (page HTML) et une forme pour la distribution (PDF par exemple) et que cette forme sera obtenue à partir d'une source d'édition unique (en langage Wiki).

6. Classement et indexation des documents

L'opération de classement d'un document, consiste à le prendre globalement en compte et à lui associer des catégories générales. L'indexation a pour objectif, quant à elle, de décrire finement le contenu des documents d'un point de vue signal et/ou sémantique. On distingue l'indexation des documents textuels de celle des documents multimédia (images, vidéo, pages mixtes texte/images). Cette dernière est un sujet de recherche en soi tandis que la première possède des outils largement diffusés.

L'indexation des documents est un problème reconnu des sites encyclopédiques tels que Wikipedia : désambigüisation et gestion des doublons, une structuration floue des relations entre catégories et une indexation plus ou moins sérieuse contrairement aux systèmes documentaires plus traditionnels.

6.1. Le classement humain

Le premier élément de classement et d'indexation est l'humain. Celui-ci va ranger les documents dans une ou plusieurs arborescences (voire des graphes) représentant différents types de classements préalablement définis (temporel, historique, géographique, politique, thématique, etc.). Ces types de classements permettent de proposer à l'utilisateur une recherche de documents par affinage successif.

C'est à l'IHENU de déterminer les arborescences de classement adéquates. C'est un travail particulièrement important et difficile, mais nécessaire. En effet, il permettra de mettre à jour les grands concepts constituant la connaissance disponible sur les Nations Unies, et leurs relations.

6.2. L'indexation humaine

L'humain peut aussi décider des mots-clés associés à un document qui seront pris en compte lorsque l'utilisateur fera une recherche par mots-clés.

Le principal risque de l'indispensable classement humain est le risque d'incomplétude: oubli de classification ou oubli de mots-clés. Un autre risque est que des catégories identiques soient représentées par des mots-clés différents, par exemple « emploi » et « travail ». La solution est alors de proposer aux humains indexant le choix de mots pré-désignés dans des catégories prédéfinies avec le caractère restrictif que cela comporte. Le multilinguisme est également un souci.

ANR : de façon plus ambitieuse, on peut s'inscrire dans le web sémantique et prévoir des ontologies et des formats de description des contenus, de façon à pouvoir exploiter également des raisonneurs logiques dans le requêtage.

L'avantage d'une telle description conceptuelle est qu'elle permet une meilleure adaptation aux différentes langues et termes utilisés par les internautes. Des travaux de recherche sont en cours au laboratoire TSI, sur la construction d'ontologies pour l'annotation d'images satellitaires, en partenariat avec l'entreprise Mondeca.

6.3. Une solution externe : Google

Wikipedia, dans ses premières années, a utilisé Google pour l'indexation de ses documents. Une requête dans Wikipedia était alors transformée en une requête Google. Nous ignorons la nature du lien contractuel entre Google et Wikipedia.

On peut également choisir un autre prestataire que Google pour ce service. Exalead ou PERTIMM, par exemple. Cela pourrait faire partie d'un projet ANR.

6.4. Indexation automatique des documents textuels

Pour pouvoir être automatiquement indexés, les documents doivent être sous forme textuelle.

Cela exclut les simples scans (images) dans le cas des documents plus anciens. Si l'on désire indexer automatiquement les anciens documents, il faut qu'ils soient lus avec un logiciel de reconnaissance optique de caractères (OCR), ce qui est une longue et difficile affaire.

Il est possible de convertir à la forme textuelle la plupart des formats de documents tels que PDF, Word, Excel, etc. Cette conversion n'est pas parfaite, à cause des traits d'union par exemple, mais elle est globalement suffisante pour leur indexation automatisée.

Le logiciel libre Lucene, **http://lucene.apache.org**, est un logiciel d'indexation incrémentale des textes particulièrement efficace et *scalable, ie que ses*

performances ne dépendent pas du nombre de documents traités. Il est utilisé par de nombreux sites tels que ceux d'Apple, de Disney, de FastFind, d'IBM ou encore Wikipedia.

Lucene permet une indexation incrémentale des documents textuels sans nécessiter des ressources très importantes. La taille des données d'indexation représente environ 30% de la taille des données indexées, ce qui est relativement peu.

Lucene propose une gamme très complète de mécanisme de requêtage.

ANR : évaluer la pertinence (*ranking*) des résultats des requêtes dans un contexte, ici celui d'une encyclopédie contenant une part importante de documents à caractère juridique reste un verrou scientifique.

De même, la présentation des résultats sous frome de groupe de documents similaires (*clustering* des résultats) est aussi un problème d'actualité.

Enfin, une présentation sous forme de carte cliquable présentant les liens entre les différents résultats du requêtage est aussi une option.

6.5. Indexation des documents multimédia

On désigne par *document multimédia* les images (animées ou non), les enregistrements sonores, les pages web (qui peuvent contenir intrinsèquement différentes modalités textes/sons/images), etc. L'indexation de ces documents consiste à extraire une représentation du contenu informationnel en vue d'une récupération ultérieure. Ce contenu informationnel peut effectivement être traduit en termes de mots et Google a longtemps fait de l'indexation d'images à l'aide d'une description textuelle de leur contenu : cette description provient du texte alentours dans une page web, ou d'une annotation manuelle (qui présente tous les inconvénients cités précédemment).

Depuis les années 1990, s'est développée une indexation numérique à l'aide d'extracteurs spécifiques : pour le son, on s'intéresse aux évolutions du pitch, du rythme, des enveloppes spectrales, alors que pour les images on s'intéresse aux textures, couleurs et formes. Les représentations numériques reposent sur des modèles mathématiques aujourd'hui bien maîtrisés, dont une collection est normalisée au travers de MPEG7. Le problème de ces représentations numériques est qu'elles ne sont pas directement manipulables par les êtres humains et ne sont pas directement interprétables sémantiquement. Il est nécessaire de prévoir des modes spécifiques de requêtage, classiquement appelés « par l'exemple » : il s'agit alors de chercher des documents à partir d'une (ou d'un ensemble de) image(s)/son(s) exemplaire(s). Ceci introduit la nécessité de comparer (à l'aide de mesures de distances pertinentes) les indexes numériques extraits des documents. Bien souvent des classificateurs automatiques sont également exploités afin de structurer l'ensemble des signatures numériques des documents traités.

Le décalage d'interprétation introduit entre l'information portée par les images (accessible directement à l'œil, difficile à traduire verbalement) et celle portée par les signatures numériques est appelée *fossé sémantique* dans la littérature. De nombreuses approches sont présentées pour franchir ce fossé et l'une, appelée *boucle de pertinence*, exploitée depuis plusieurs années à Télécom

ParisTech avec des bases d'images, est maintenant bien reconnue : il s'agit d'exploiter quelques annotations manuelles que peut facilement faire l'internaute pour valider/invalider les résultats retournés par le système, et produire une sur-couche d'indexation adaptée à l'usager.

ANR : cela implique des méthodes d'apprentissage machine plus ou moins complexes. Il n'existe pas, à notre connaissance, de solution logicielle sur le marché permettant cela. Cependant, dans le cadre du projet Infom@gic (en collaboration avec PERTIMM), nous avons pu tester une approche simple et efficace sur une base de 80 000 images environ.

6.6. Recherche de documents

La recherche de document pourra se faire dans l'une des arborescences définies par l'ONU.

Elle pourra se faire également par mots clés. Différentes stratégies de recherche et d'ordonnancement des résultats peuvent être mise en œuvre. La gamme est très étendue. Ces stratégies peuvent être personnalisées pour un utilisateur. Personnalisation de la forme et du contenu ; adaptation également pendant le requêtage, en prenant en compte les actions de l'internaute et en lui suggérant des pistes pour affiner ses résultats.

6.7. Le multilinguisme

Il est prévu que l'Encyclopédie mettra en œuvre 8 langues, les 7 langues officielles de l'ONU et le Portugais. Si tous les documents sont traduits avant leur publication, il n'y a rien de spécial à faire.

Dans le cas contraire, dans le cas d'une recherche par mots-clés, il faut prévoir de faire une recherche multilingue en faisant appel à la traduction des mots-clés dans les 8 langues de l'Encyclopédie.

Actuellement une solution aux problèmes du multilinguisme est le recours aux ontologies et à la conceptualisation formelle, dans le cadre de ce qui s'appelle le Web Sémantique. Les concepts définis peuvent être reliés à des termes traduits en différentes langues et reliés sémantiquement à d'autres. Cette approche est relativement intéressante et bien développée en langue anglaise. Cependant, il n'est pas certain que les chinois conceptualiseraient de la même façon que nous les domaines couverts par l'Encyclopédie ; par conséquent, il ne s'agit pas simplement d'un problème de traduction de langue mais bien de modes de représentations mentales, ce qui pourrait faire l'objet d'une étude spécifique. Des linguistes et des experts en ontologie pourraient être associés (cf. le projet ANR DAFOE **www.dafoe4app.fr**, dans lequel Telecom ParisTech intervient).

6.8. La solution ANR

Dans le cas d'un projet ANR d'Encyclopédie, l'indexation et la recherche intelligente de document pourraient faire l'objet d'un Work Package soutenu par Exalead, ou PERTIMM que nous pourrions contacter.

7. Diffusion des documents

Dans cette section, nous abordons différents points liés à la diffusion de documents par un site Web.

7.1. Option : visualisation sans distribution

Une possibilité est de donner accès à la visualisation des documents sans pour autant distribuer le document. C'est ce que font parfois les sites de ventes en ligne de journaux et de périodiques. On peut réserver la distribution effective du document à des utilisateurs inscrits par exemple. Tandis que les utilisateurs communs n'auraient accès qu'à la visualisation de ces documents.

Proposer les documents en visualisation n'empêche bien sûr pas le client du site de faire des copies d'écran et de reconstituer une copie du document. Cependant, cette copie ne sera qu'une copie imparfaite et sans valeur. Il est aussi possible d'incruster des marqueurs dans la présentation des documents proposés en visualisation.

Spécialistes de la visualisation et parfois appelés les YouTube du document :

- Scribd (**http://www.scribd.com**), API PHP libre.
- DocStoc (**http://www.docstoc.com**)
- SlideShare (**http://www.slideshare.net**)

Nous retenons Scribd qui dispose d'une API gratuite qui permet d'intégrer le service à diverses technologies, dont des implémentations existent pour PHP, Ruby et .NET (C#). L'aspect technique de la chose semble donc raisonnablement accessible. Reste à examiner juridiquement la License de Scribd API.

Scribd, probablement le plus générique, supporte ainsi par exemple les formats Microsoft Word, Microsoft Excel (**.xls**), Microsoft Powerpoint, texte, Adobe PostScript, OpenOffice, OpenOffice Presentations, OpenOffice Spreadsheets, OpenDocument, Rich Text, JPEG, Portable Network Graphics et GIF. Une fois traités par la plate-forme en ligne, tous sont consultables à l'aide d'un simple lecteur Flash, et donc via le premier navigateur venu.

7.2. Option : signature des documents

La signature numérique est le mécanisme permettant d'authentifier l'auteur d'un document électronique et de garantir l'intégrité du document. Pour cela, les conditions suivantes doivent être réunies :

- Authentique : l'identité du signataire doit pouvoir être retrouvée de manière certaine.
- Infalsifiable : la signature ne peut pas être falsifiée. Quelqu'un ne peut se faire passer pour un autre.
- Non réutilisable: la signature n'est pas réutilisable. Elle fait partie du document signé et ne peut être déplacée sur un autre document.
- Inaltérable : un document signé est inaltérable. Une fois qu'il est signé, on ne peut plus le modifier.

- Irrévocable : la personne qui a signé ne peut le nier.

La signature électronique repose sur un système à base de deux clés, l'une privée que seul détient l'expéditeur du document, et l'autre publique librement accessible sur le web.

- Pour signer un document numérique, on utiliser une clé privée, à laquelle ne correspond qu'une seule clé publique et réciproquement. Seule apparaît la signature attachée au document.
- Le destinataire se procure alors la clé publique correspondant à la clé privée soit directement auprès de l'émetteur, soit par le biais d'un tiers de confiance. Il lui suffit alors de s'assurer que les deux clés correspondent, ce qui lui garantit l'identité de l'émetteur et l'intégrité du document qui lui est parvenu.

La signature de documents peut être mise en œuvre dans le cadre de l'Encyclopédie afin que les clients du site puissent avoir la certitude que le document est bien validé par l'émetteur qui, dans notre cas, est un organisme de référence. Il peut ainsi prouver que le document lui a bien été remis par l'Encyclopédie.

8. Définir les processus collaboratifs d'édition

L'édition collaborative permet à différentes personnes, éventuellement distantes, de produire un document commun destiné à être publié. La première chose à faire est de définir le processus d'édition collaboratif depuis l'idée d'un document jusqu'à sa publication effective. Il s'agit de dégager une logique de groupe. Il est possible qu'il existe plusieurs types de documents éditables et que les processus d'édition diffèrent selon le type de document. Ainsi une courte de note d'actualité pourra être rapidement éditée, soumise à traduction, et publiée par une personne habilitée tandis qu'un rapport plus profond nécessitera le travail coordonné d'une équipe.

En plus du protocole de rédaction, étant donné l'importance de l'ONU, on insistera sur les différents rôles et leurs responsabilités. Par exemple, le rédacteur en chef sera le responsable du document, il organisera l'équipe : arbitres, contributeurs, traducteurs, etc., distribuera les droits, établira les *deadlines*, décidera de la publication, etc.

Dans ce processus, on n'oubliera pas la maintenance du document après publication pour permettre les corrections des erreurs ou les addenda, et même le processus d'élimination du document.

On peut également vouloir imposer des styles (titres, paragraphes, listes, etc.) ainsi qu'un calibrage (longueur des articles, des sections, etc.) selon les types de documents afin de favoriser une certaine homogénéité de l'Encyclopédie.

Cette partie du projet peut représenter un Work Package d'un projet ANR en collaboration avec un laboratoire universitaire spécialisé dans ce domaine. En effet, être capable de gérer un processus collaboratif d'édition (avec par exemple des validations partielles par l'éditeur en chef) et un mécanisme fin de gestion des versions (par exemple, retrouver l'historique d'une phrase ou d'un paragraphe) reste un challenge complexe.

9. Les outils pour le travail collaboratif

La grande famille des outils d'édition collaborative est celle des wikis dont le principal représentant est Mediawiki. Le wiki est un site dynamique qui permet à toute personne disposant d'une connexion d'en modifier les pages via son navigateur Web à l'aide d'une syntaxe simple. Tous favorisent l'auto publication collective et même la structuration a posteriori des contenus. Un wiki propose dans une même interface des contenus, des méta informations sur l'activité éditoriale et des outils de communication.

Etant donné le degré de maturité de ces outils d'édition, il ne nous semble pas utile de chercher une solution plus innovante à ce niveau.

Mediawiki est le logiciel wiki amiral de la fondation Wikimedia et de toute la galaxie wiki. C'est le développement le plus abouti en matière de wiki. Il peut faire à peu près tout, cependant au prix parfois d'une certaine lourdeur de mise en œuvre. Mediawiki est surtout le concentré d'une certaine culture : à savoir l'esprit wiki très particulier qui s'est développé, depuis les premières tentatives de Ward Cunningham, en 1995, à travers une galaxie de logiciels et de sites, tels, dans le monde anglo-saxon, MeatBall et CommunityWiki, ou CraoWiki dans le monde francophone.

Dans un wiki, la hiérarchie, aussi bien des documents que des contributeurs, n'est pas donnée a priori. Elle émerge au fur et à mesure des échanges et peut à tout moment être modifiée, voire basculer. Encore ne peut-on pas parler véritablement de hiérarchie, puisque l'on est dans un véritable hypertexte. Cet aspect demandera à être contrôlé dans notre cas.

L'une des fonctions primordiales des wiki est la fonction *backlinks* (rétroliens) qui permet de lister les documents pointant vers le document courant. Il s'avère de cette fonction est décisive, en particulier pour les documents à caractère juridique. En effet un article de loi, ou bien de traité international, peut être cité par un autre article qui en modifie, confirme ou infirme le sens et la portée. Toute lecture de texte juridique doit se faire à la lumière de cette réalité sous peine de mener à des interprétations fausses ou illusoires, ce qui est parfois le cas chez les rédacteurs eux-mêmes. La première wikification intégrale d'un texte législatif est à notre connaissance celle du Traité établissant une Constitution pour l'Europe, mise en œuvre sur le wiki *notreconstitution.net* en 2005. Ce formalisme wiki a été depuis systématisé par Légifrance.

ANR : un outils de création automatique le liens hypertextes pourrait être proposé dans le cadre d'une encyclopédie à caractère juridique.

Mediawiki met aussi en œuvre une certaine syntaxe d'édition, qui bien qu'assez simple puisqu'elle est devenue couramment pratiquée par les centaines de milliers d'auteurs de Wikipedia, nécessite néanmoins un certain apprentissage même lorsque le wiki est doté d'une interface d'édition WYSIWYG.

Ces caractéristiques propres aux wikis et en particulier à Mediawiki, la culture particulière et le formalisme hypertexte, impliquent de réserver un certain temps d'adaptation, d'accompagnement et de formation mutuelle des membres de l'équipe en charge de l'encyclopédie.

9.1. Modules complémentaires au wiki

Wikimedia ne fonctionne pas seul, à savoir que dans le cas de Wikipedia par exemple, les équipes et les contributeurs ont été amené à concevoir et développer des modules logiciels complémentaires. En particulier :

1. des outils de veille, de monitoring et de visualisation spécifiques permettant de surveiller l'activité des contributeurs ainsi que l'intégrité et la cohérence des pages;

2. des robots remplissant aussi de nombreuses tâches ingrates, telle l'ajout systématique de tags sur des faisceaux de pages, ou bien le contrôle de l'orthographe de certains termes.

3. des outils de communications externes, synchrones ou asynchrones, entre contributeurs: mail, listes de liste de discussion, messagerie instantanée, blogs, micro-blogging, texte ou tableau blanc partagé.

4. des outils cartographie, permettant d'illustrer les articles et de géoréférencer les éléments qui le méritent.

5. des outils de formatage de données brutes, permettant aux utilisateurs de les télécharger au format de son choix.

Cette liste est en perpétuelle évolution, bien entendu.

Dans le cas de l'Encyclopédie, des outils, pour certains similaires devront être mis en œuvre. D'autres plus particuliers devront être conçus pour faciliter l'activité des équipes et leur permettre de se coordonner au mieux.

Les points cités appellent les remarques suivantes:

1. Ces outils de veille, de monitoring et de visualisation qui sont très bénéfiques en terme de productivité, posent néanmoins d'éventuels problèmes de confidentialité et de vie privée pour le membres des groupes de travail. Ils ne peuvent donc être développés et évoluer que sur leur demande et avec leur accord.

2. Les robots sont spécifiques à la plate forme wiki qui sera choisie au final. Si c'est Médiawiki, de nombreux codes sources sont disponibles, dans le cas contraire, il faudra développer.

3. Les outils de communication externes posent des problème d'autonomie du groupe en termes de soumission aux pannes ou aux intrusions. Il conviendra de choisir ces systèmes, voire d'en mettre en place de spécifiques, avec ce soucis en tête.

4. Pour ce qui est de création de cartes (or reprise graphique de documents anciens scannés), il parait illusoire de vouloir développer un Système d'Information Géographqiue (SIG) spécifique. Des partenariats pourraient être conclus avec certains fournisseurs de fonds de carte et de modules de dessin. Il faudra faire un choix entre deux opérateurs aux philosophies très différentes: Google (propriétaire) et OpenStreetMap (Libre). Le deuxième opérateur (OpenStreetMap) permettrait une meilleure autonomie de l'encyclopédie en rendant possible la diffusion des documents depuis son

serveur même. Dans ce cas, un module serveur cartographique serait à étudier en complément.

5. En matière de diffusion de données brutes, de nombreux gouvernements (cf. liste) ont d'ors et déjà des portails (ou des projets de portails) distribuant une variété plus ou moins grande de "datasets". Il conviendra d'étudier en détail les modèles développés par ces Etats et de préciser la typologie des données que l'Encyclopédie de l'ONU pourra proposer de manière complémentaire, ainsi que la licence sous laquelle elle seront fournies.

    - Nouvelle-Zélande : data.govt.nz
    - Australie : data.australia.gov.au
    - USA : data.gov data.gov.uk
    - France : www.apiefrance.com/sections/actualites/vers_un_portail_uniq/

Sur certains de ces outils, un ou plusieurs Work Packages d'un projet ANR pourraient être conçus.

9.2. Un wiki particulièrement intéressant : XWiki

Un outil gratuit et libre a retenu notre attention, il s'agit de XWiki. XWiki est la propriété d'une startup française : **http://www.xwiki.org**. Ce site, bien sûr édité avec XWiki, montre l'étendu, les performances et la qualité de ce système.

L'un des intérêts de XWiki est qu'il possède une interface d'édition WYSIWYG, donc accessible au plus grand nombre des contributeurs possibles. Il permet la publication de documents dans un grand nombre de formats.

L'autre aspect important de XWiki est qu'il est porté par une société commerciale à qui il sera possible de commander des modifications appropriées du système. On pourra par exemple commander à XWiki la mise en œuvre des processus d'édition de la section 8. Ou bien on peut leur commander de mettre en œuvre le contrôle des styles de documents favorisant l'homogénéité de l'Encyclopédie.

D'autre part, il est bien vu d'intégrer une startup telle XWiki dans un projet ANR.

10. Stockage et distribution : la solution externe

Pour gérer la disponibilité d'un service web de distribution, il existe plusieurs techniques :

- louer une prestation complète auprès d'un infocentre ;
- acheter et mettre en œuvre sa propre architecture (*Data Center*, connexion Internet, *Load Balancer*, serveurs, etc.)
- ou faire appel à un *Content Delivery Network* (CDN).

Les CDN sont des architectures réparties de serveurs qui coopèrent afin de mettre à disposition du contenu multimédia volumineux à des utilisateurs. Ces nœuds coopèrent afin de satisfaire les requêtes, leur envoyant le contenu en retour, et déplaçant le contenu de façon transparente (mise en cache,

compression lors du transfert) afin d'optimiser le mécanisme de transmission. L'optimisation peut se traduire par la réduction des coûts de bande passante et l'amélioration de l'expérience utilisateur.

L'utilisation d'une telle plateforme de stockage et de distribution pose cependant des problèmes de confidentialité, de coût et de pérennité.

10.1. Akamai

Akamai (**http://www.akamai.fr**) est un CDN parmi les plus réputés. Il dispose de plus de 56,000 serveurs répartis dans le monde.

De nombreuses entreprises font appel au CDN Akamai (Apple, des journaux, la présidence des USA, etc.).

10.2. Amazon S3

Bon nombre de services utilisent la plateforme CDN Simple Storage Service (S3) d'Amazon Web Services (**http://aws.amazon.com**) pour stocker et distribuer des documents. Voici à titre indicatif les tarifs mensuels pratiqués en 2009 :

Niveau	**États-Unis**	**UE**	**Description :**
0-50 To	$0,150/GB	$0,180/GB	Premiers 50 To de stockage
50-100 To	$0,140/GB	$0,170/GB	50 To suivants de stockage
100-500 To	$0,130/GB	$0,160/GB	400 To suivants de stockage
500+ To	$0,120/GB	$0,150/GB	Au-dessus de 500 To

11. Architecture informatique et réseau

Dans une architecture de services telle qu'envisagée pour l'Encyclopédie, on distinguera les serveurs de distribution et les serveurs d'édition. Il s'agit en fait de deux services différents :

- Les serveurs de distribution sont des serveurs de grande capacité accessibles depuis le monde entier pour les clients de l'Encyclopédie. Ils sont relativement statiques. Ils doivent être capables de supporter de très nombreux utilisateurs.
- Les serveurs d'édition sont des serveurs de travail collaboratif, accessibles uniquement aux équipes d'édition, servant à la préparation de documents et à leur publication. Ils supportent un nombre plus restreint d'utilisateurs.

Cela signifie que l'Encyclopédie a besoin de deux architectures logicielles différentes.

11.1. Serveurs pour l'édition collaborative

Concernant l'édition collaborative d'articles, de rapports, d'analyses, etc., on ne parle plus de milliers d'utilisateurs simultanés. On peut donc utiliser des serveurs classiques de type LAMP, cf. section 11.3.

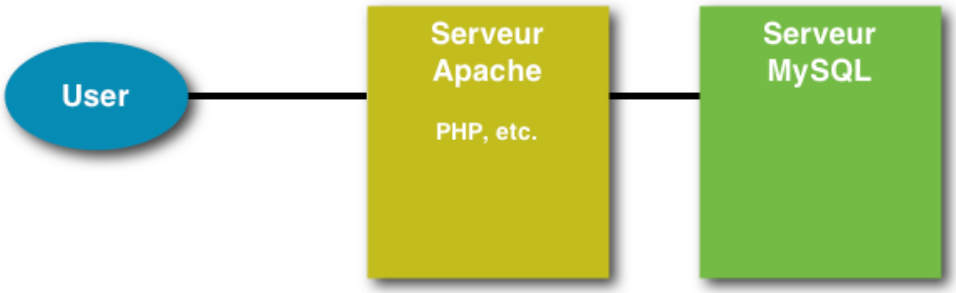
Il est également envisageable de dupliquer ces serveurs afin que chaque domaine d'édition de l'Encyclopédie ait son propre serveur.

11.2. Une architecture de référence pour la distribution

Une architecture de référence est actuellement celle de l'encyclopédie en ligne Wikipedia qui supporte plusieurs dizaines de milliers de connexions par seconde. Cette architecture sert de référence pour les sites de contenu à fort trafic. Wikipedia gère environ 10 millions de documents en une centaine de langues. Ces documents sont continuellement révisés. Elle compte environ 350 serveurs gérés par 6 personnes.

11.3. Architecture LAMP classique

L'architecture LAMP[1] classique est l'architecture d'un site Web utilisant les composants classiques :



Cette architecture ne supporte pas un très grand nombre de connexions ni un très grand débit.

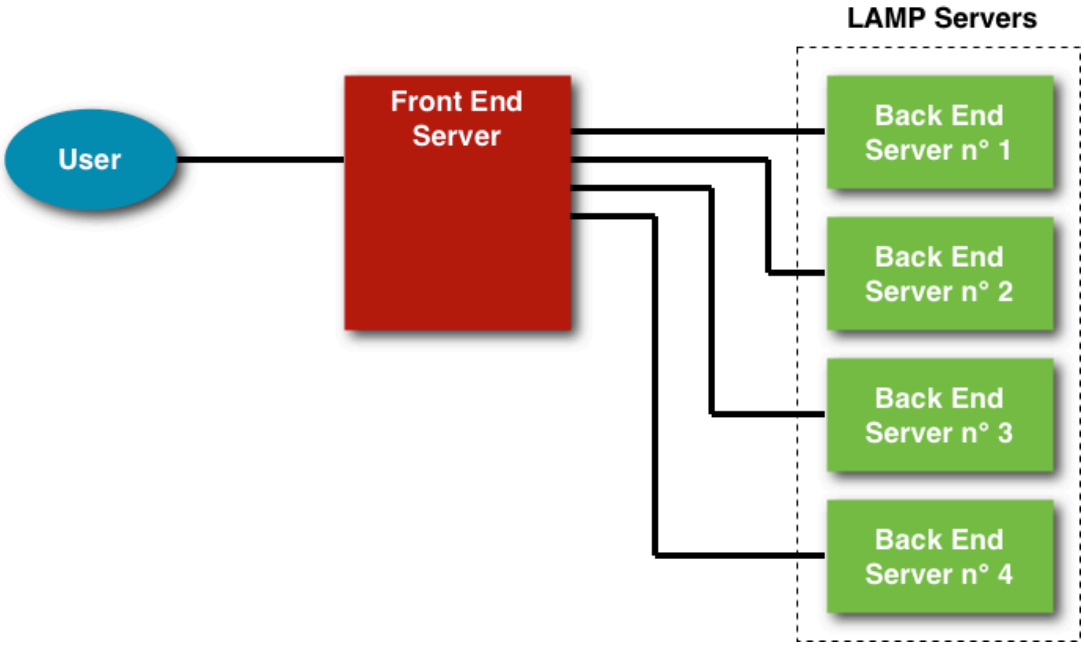
11.4. Architecture Front end – Back end

Afin de rendre les performances suffisantes pour une encyclopédie avec un grand nombre d'utilisateurs, on procède en mettant un ou plusieurs serveurs en *front end* et un plus grand nombre de serveurs de type LAMP en *back end*.

Les serveurs en *front end* sont chargés de recevoir les requêtes des utilisateurs du site et de répartir la charge sur les serveurs en *back end* qui, eux, font effectivement le travail nécessaire. Cela est totalement transparent pour l'utilisateur qui se connecte au serveur de *front end*. Ce dernier fait le travail d'une passerelle entre l'utilisateur (client) et l'un des serveurs de *back end*.

L'ensemble serveur de *front end* et serveurs de *back end* sera classiquement appelé un *cluster*.

[1] LAMP signifie « Linux, Apache, MySQL, PHP » qui sont les logiciels normalement utilisés pour bâtir un serveur.



Ce type d'architecture peut poser des problèmes. On distingue les transactions dites *stateless* de celles dites *stateful*.

Des transactions sont dites *stateless* si le serveur ne garde aucune trace de la transaction à l'exception des logs. C'est le mode de connexion usuel sur un site. Mais dans certains cas, par exemple si l'utilisateur possède un compte, il faut alors gérer une session. Les transactions ne sont plus indépendantes les unes des autres, on dit qu'elles sont *stateful*. Classiquement, c'est le serveur qui conserve des données de session qu'il peut retrouver grâce aux fameux *cookies* qu'il range chez le client.

Dans le cas où un serveur de *front end* est utilisé, il se peut qu'une transaction soit orientée vers un serveur *back end* et la transaction suivante orientée vers un autre serveur *back end*. Ce dernier n'a alors pas les données de session à sa disposition. Nous verrons comment traiter ce problème plus tard.

La solution *front end – back end*, très classique, a l'avantage de pouvoir être mise en œuvre de manière incrémentale au fur et à mesure que l'Encyclopédie prend de l'ampleur. Les logiciels du serveur *front end* et ceux du serveur *back end* peuvent même être installés sur la même machine au tout premiers temps. Puis lorsque le trafic s'accroitra, un véritable serveur *back end* sera introduit, puis un deuxième et ainsi de suite. Il suffira de reconfigurer le logiciel de *front end*.

11.5. Le Content Delivery Network

Le CDN de Wikipedia est composé d'un *cluster* principal et de deux *clusters* secondaires en cache. Le premier est situé en Floride, un autre en Hollande et le troisième en Corée du Sud. Les deux derniers sont en fait des « copies » du premier. Les *clusters* sont appelés des *clusters* régionaux.

Au début de son existence, l'Encyclopédie n'aura peut-être besoin que d'un seul cluster. Mais quand le trafic deviendra important, il sera peut-être nécessaire de créer des clusters régionaux.

11.6. Répartition de charge entre les clusters

Le serveur de *front end* est capable de répartir la charge de travail sur les serveurs de *back end*. La répartition au niveau supérieur de la charge sur les *clusters* régionaux se fait grâce à un système de DNS intelligent.

Le DNS, *Domain Name Server*, est le système qui associe les noms de domaine, par exemple **www.rivf.org**, à l'adresse numérique du serveur sur Internet (adresse IP). C'est cette adresse qui identifie réellement le serveur et qui permet de lui acheminer des requêtes. C'est la résolution de noms de domaine.

Afin de répartir la charge globale sur les différents clusters régionaux, on peut décider de critères géographiques. On peut, par exemple, décider que les clients des deux Amériques seront orientés vers le *cluster* situé aux USA, que les clients d'Afrique et d'Europe seront orientés vers le *cluster* situé en Europe et que les clients de Russie, d'Asie et d'Australie seront orientés vers le *cluster* en Asie.

Ce critère de répartition n'est certainement pas optimal puisqu'il ne prend pas en compte les périodes d'activité forte (journées) et faible (nuits). Il a cependant l'avantage de pouvoir être facilement mis en œuvre à l'aide d'un DNS intelligent appelé GeoDNS.

GeoDNS administre le nom de domaine. Il reçoit des requêtes des clients logiciels ou des serveurs DNS délégués. GeoDNS localise ce client. Est-il en Amérique, en Europe, etc. En fonction de la localisation du client, GeoDNS donnera une réponse différente. Par exemple, si le client est en Europe, GeoDNS répondra par l'adresse IP du *cluster* européen. Ainsi, le server **www.wikipedia.org** ne sera pas physiquement le même selon que le client est en Asie, en Amérique, en Europe, etc. Cette solution a simplement le défaut de ne pas respecter la philosophie de transparence d'Internet.

D'autres solutions existent mais posent certains problèmes :

- Le DNS peut renvoyer une liste d'adresses IP pour un nom de serveur. Ce sont les adresses IP de tous les serveurs équivalents. Alors le client en choisit une au hasard. Cela peut fonctionner mais suppose un client intelligent, ce que l'on ne peut imposer à tous les clients logiciels du DNS. C'est pour cela que les sites proposent parfois aux utilisateurs des sites miroirs (*mirror* sites) pour leurs téléchargements : ils se reposent sur l'intelligence de l'humain qui choisira le site miroir. Supposer l'intelligence des programmes clients DNS ne semble pas raisonnable.
- Une autre solution est d'avoir un seul point serveur d'entrée et que celui-ci fasse une redirection HTTP vers un *cluster* choisi selon un algorithme spécifique ou bien au hasard. Alors le client oublie le point d'entrée et se redirige vers le *cluster* de redirection. Cette solution suppose un logiciel client capable d'interpréter les redirections, c'est le cas des navigateurs mais ce n'est peut-être pas le cas de tous les logiciels clients.

Il nous semble donc que la solution utilisant quelques *clusters* régionaux et une répartition géographique de charge entre les *clusters* via GeoDNS est une solution réaliste. De plus, elle possède l'avantage de n'être nécessaire que lorsque l'Encyclopédie montera en charge et nécessitera plusieurs clusters régionaux.

11.7. Répartition de charge à l'intérieur d'un cluster

Le *front end* est responsable de la répartition de la charge de travail entre les serveurs de *back end*. Un logiciel réputé pour cette fonction est *Linux Virtual Server* (LVS) utilisé par Wikipedia.

LVS supporte le passage à l'échelle et propose les principales stratégies algorithmiques de répartition de charge parmi lesquelles il faudra choisir. Par ailleurs, LVS permet de mettre en œuvre la persistance du lien entre le client et un serveur de *back end* particulier. Quand un client est mis en correspondance avec un serveur de *back end*, un *timer* (compte à rebours) est déclenché. Si le même client fait une nouvelle connexion avant que le *timer* expire, il sera mis à nouveau en correspondance avec le même serveur de back end et le *timer* sera réinitialisé. La *LVS Persistence* permet donc de résoudre le problème posé en fin de section 11.4 concernant la gestion des sessions.

Notons que la capacité de LVS à répartir la charge de travail dans un *cluster* de serveurs est très élevée.

11.8. Mécanismes de cache

Parce que l'actualité peut concentrer les demandes de documents sur un petit sous-ensemble de l'Encyclopédie, il importe de mettre en place un mécanisme de cache. Le mécanisme de cache se trouve entre le client et le serveur *back end*. La première fois qu'une requête est faite, elle est transmise au serveur *back end* mais le mécanisme de cache en stocke le résultat avant de le transmettre au client. Si la même requête est faite par un autre client, elle ne sera pas transmise au serveur car le mécanisme intermédiaire de cache a déjà le résultat en mémoire, le mécanisme de cache peut directement répondre ce résultat.

Un logiciel très connu de mécanisme de cache est Squid. Il est également gratuit et utilisé par Wikipedia.

11.9. Stockage

Les documents seront stockés dans des serveurs de fichiers à haute performance. Les métadonnées des documents seront stockées dans des bases de données MySQL.

11.10. Délivrance des documents

Les pages HTML ainsi que les petits documents textes seront délivrés par un serveur Apache très classique. Apache est l'outil qui supporte les *Content Management System* (CMS).

Les données plus lourdes, fichiers PDF ou document multimédia, seront délivrées par un serveur Lighttpd. Ce serveur, **http://www.lighttpd.net**, qui ne possède pas toutes les fonctionnalités d'Apache, est bien plus efficace. Il est ainsi utilisé par Youtube et Wikipedia pour leurs fichiers multimédia.

La conception et la mise en œuvre d'une architecture extensible pour la délivrance de documents peuvent être incluses dans un projet ANR et confiée à Télécom ParisTech.

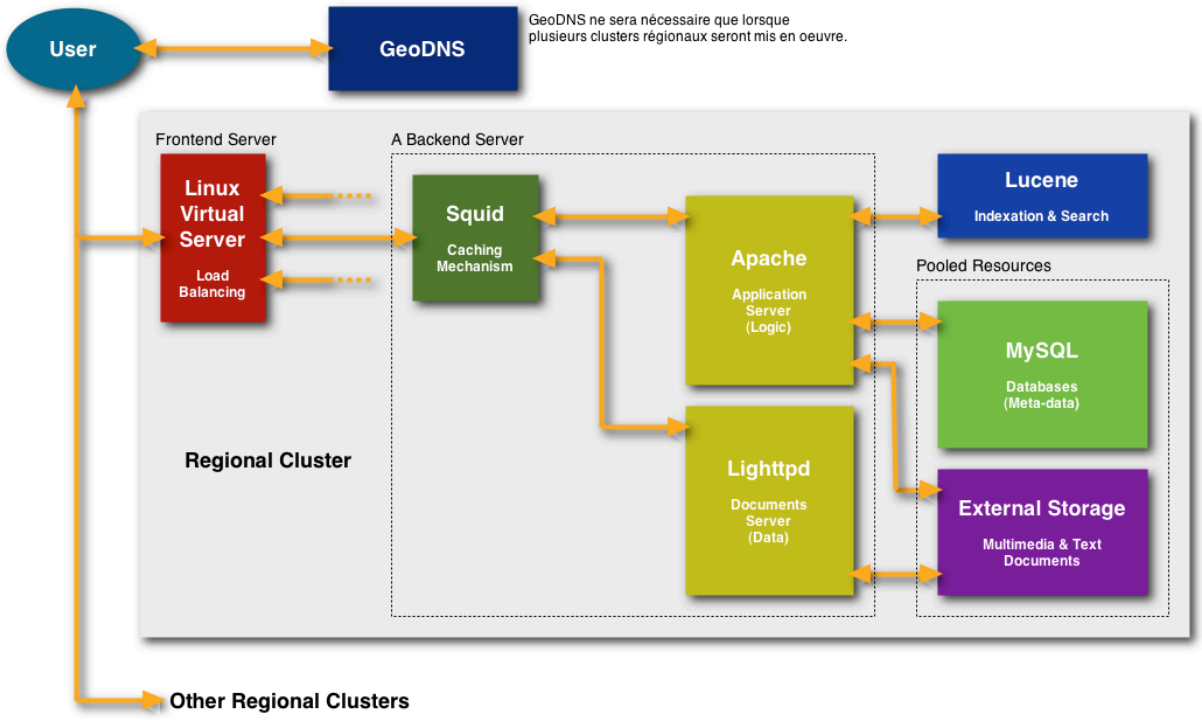
12. Les logiciels libres et gratuits utilisés

Dans l'architecture que nous proposons et qui ressemble beaucoup à celle de Wikipedia, tout est composé de logiciels libres, gratuits et qui ont fait leurs preuves dans différents services en ligne :

- **LVS**, *Linux Virtual Server*, `http://www.linuxvirtualserver.org`, est le logiciel du serveur de *front end* qui permet la répartition de la charge sur les serveurs de *back end*.
- **Squid**, `http://www.squid-cache.org`, est le logiciel des proxys de cache qui sont placés entre le serveur de *front end* et les serveurs de *back end*.
- **Apache**, `http://www.apache.org`, est le serveur Web universellement utilisé possédant d'excellentes performances et supportant beaucoup de logiciels, en particulier MediaWiki.
- **MySQL**, `http://www.mysql.com`, est le serveur de base de données universellement utilisé en conjonction avec Apache.
- **Lighttpd**, `http://www.lighttpd.net`, est un serveur Web léger et particulièrement efficace pour délivrer des contenus multimédia et des contenus lourds, par exemple des documents au format PDF.
- **Lucene**, `http://lucene.apache.org`, est un logiciel d'indexation et de recherche de données textuelles.

13. Dessin de l'architecture de distribution

Cette architecture est inspirée de celle de Wikipedia.

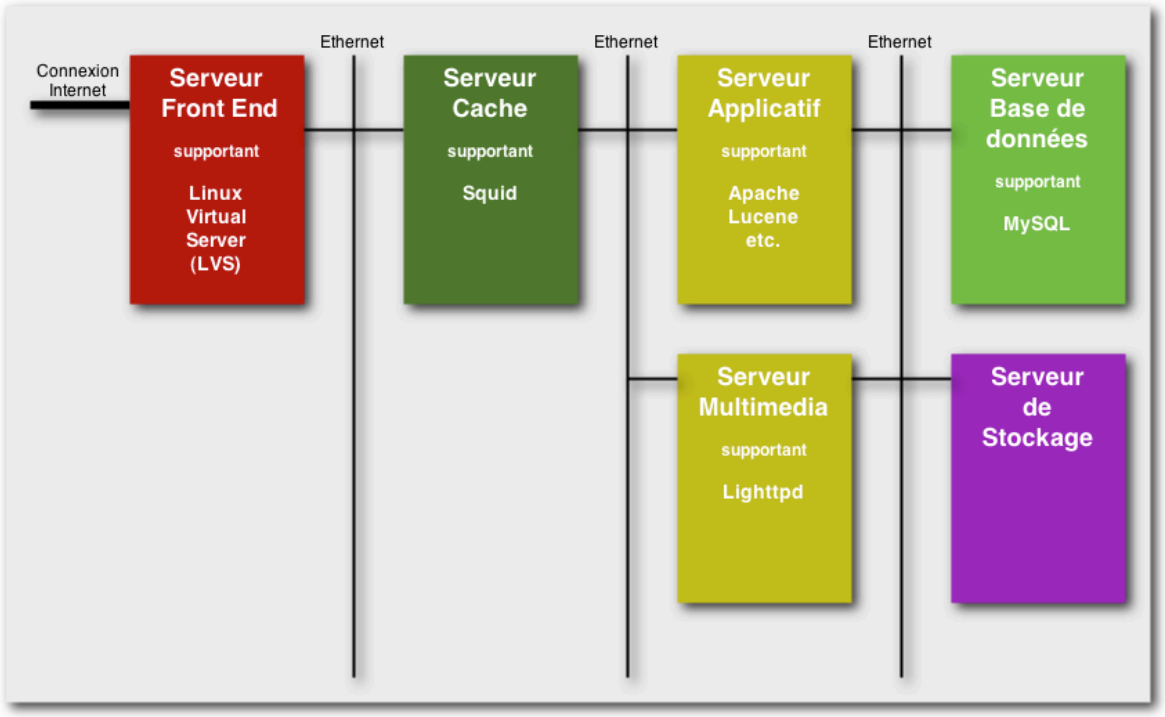


14. Architecture de distribution initiale

Mettre en œuvre l'architecture complète de la section 13 ne sera probablement pas nécessaire dans un premier temps. Nous proposons ci-dessous une architecture initiale qui pourrait convenir dans les premiers temps de l'existence du site de l'Encyclopédie.

Dans cette architecture, il y a 5 PC à profil serveur, c'est-à-dire équipés de processeurs multi cœurs (les cœurs servant en parallèle), de beaucoup de mémoire rapide et d'un disque de taille raisonnable. Pour le serveur de stockage, nous suggérons un serveur spécialisé dans ce travail. Le coût d'un tel dispositif est d'autant plus raisonnable que les logiciels sont gratuits.

Dans le cas d'un projet ANR supportant l'Encyclopédie, on peut imaginer que le dispositif soit hébergé dans un premier temps par Télécom ParisTech.



15. Réalisation

Mettre en place une version initiale de l'Encyclopédie dans 3 ans semble réaliste.

En effet, les logiciels indispensables existent et sont disponibles sur l'Internet. Bien entendu, il faut des ingénieurs pour s'approprier ces logiciels et être capable de bien les configurer.

De plus, une version initiale de l'Encyclopédie ne nécessite pas la mise en œuvre de toutes les fonctions évoquées pour être présentable.

Sans tomber dans le malthusianisme, nous pensons qu'il est préférable de viser un système initial avec des objectifs réalistes que nous sommes sûrs de pouvoir

réaliser plutôt qu'un système idéal mais qui aura immanquablement des problèmes tant l'informatique et les réseaux ne sont pas des sciences exactes.

Donc, il nous paraît essentiel de concevoir l'Encyclopédie de façon modulaire, afin de permettre une flexibilité de ses fonctionnalités dans le temps. Seulement, cette modularité nécessite de définir le cœur du système de façon très stable et réaliste techniquement. Ceci justifie l'usage de technologies actuellement bien reconnues, auxquelles peuvent se greffer des approches innovantes qui se mettront en place progressivement.

16. ANR

L'option de demander un financement partiel à l'ANR est une option envisageable. L'ANR ARPEGE semble convenir dans sa thématique « Infrastructures pour l'Internet, le calcul intensif ou les services ». Nous sommes dans « les services ». Cependant, la date limite de soumission des projets est fixée au 20 février 2010. Cela laisse peu de temps pour contacter Exalead et/ou PERTIMM et/ou XWiki et il faut que celles-ci soient intéressées. La préparation des documents pour l'ANR est une opération lourde qui signifie un travail intensif durant les semaines à venir.

Nous récapitulons ci-dessous les différentes tâches qui pourraient faire l'objet de Work Packages dans un projet ANR et nous donnons d'éventuels partenaires pour la réalisation de ces Work Packages. Cependant, il ne s'agit pas vraiment de verrous à lever, indispensables pour l'ONU, qu'ils soient scientifiques ou technologiques, mais plus d'un travail d'ingénierie et de recherche pour dégager une synthèse conduisant à un modèle d'encyclopédie en ligne.

Dans cette section, nous nous plaçons vis-à-vis de l'ANR. Le point de vue présenté n'est pas obligatoirement notre point de vue en interne mais celui que doit voir l'ANR.

Nous proposons à ARPEGE (deadlline : 20 février) une plateforme d'Encyclopédie professionnelle (pour la distinguer de Wikipedia) comprenant un Content Delivery Network et un système d'édition collaborative de type professionnel.

Nous pourrions insister sur le caractère juridique de l'Encyclopédie afin de justifier des outils spécifiques.

L'ensemble des partenaires possibles : IHENU, Télécom ParisTech, EXALEAD, PERTIMM, XWiki, universités françaises. Sauf les deux premiers, les autres ne savent pas encore que l'on pense à eux.

Des Work Packages (WP) possibles :

WP Gestion du projet :

- responsabilité de l'IHENU.

WP Encyclopédie web site design :

- Partenaires possibles : IHENU, Telecom ParisTech INFRES.
- Tâche : faire une analyse des sites encyclopédiques existants et en tirer un bilan, s'en servir pour définir un site web d'encyclopédie, les scenarios de navigation, les outils de personnalisation de la navigation, etc. Implémentation.
- Verrous : l'accessibilité depuis les PED, le multilinguisme, produire un système « universel » et cohérent.

WP Système d'édition collaborative :

- Partenaires possibles : Télécom ParisTech INFRES , XWiki, un journal en ligne (lemonde.fr ?), une université spécialisée ?
- Tâche : définir le système, les fonctionnalités et les rôles « contributeur », cf. sections 8 et 9, et les implémenter.
- Verrous : produire un système professionnel, « universel » et cohérent, capable de gérer le processus d'édition et d'avoir une gestion fine des versions.

WP Système de classification, d'indexation et de requêtage :

- Partenaires possibles : Telecom ParisTech TSI et INFRES, EXALEAD, PERTIMM.
- Tâche : définir et implémenter le système.
- Verrous : indexation et requêtage multimédia, produire un système « universel », passage à l'échelle, automatisation, ontologies, établissement des liens hypertextes. Tout ceci dans un cadre juridique.

WP Content Delivery Network :

- Partenaires possibles : Telecom ParisTech, universités françaises.
- Tâche : produire une architecture de distribution configurable en fonction de différents paramètres de taille et de contraintes de qualité de service. Proposer et implémenter des outils spécifiques à une encyclopédie (monitoring, log analysis, etc.).
- Verrous : elle doit être accessible disons à un Institut qui a les moyens d'un Institut, ni trop chère ni trop importante. Scalabilité.

WP Use case, validation, dissémination :

- Responsabilité principale : IHENU.
- Tâche : en tant que use case, on fera l'Encyclopédie sur les Nations-Unies. Celle-ci servira à la validation du résultat et à la dissémination des acquis.
- IHENU est responsable d'une tâche importante : apporter des contenus (préexistants ou développés avec le système d'édition collaborative) en nombre suffisant et structurer l'ensemble de ces contenus.

Dépôt légal : 2010 – 1er trimestre
Imprimé à l'Ecole Nationale Supérieure des Télécommunications – Paris
ISSN 0751-1345 ENST D (Paris) (France 1983-9999)

TELECOM ParisTech

Institut TELECOM - membre de ParisTech

46, rue Barrault - 75634 Paris Cedex 13 - Tél. + 33 (0)1 45 81 77 77 - www.telecom-paristech.frfr

Départements INFRES, TSI

© Institut TELECOM -Télécom ParisTech 2010