La place de l'hypertextualité dans le traitement de corpus audiovisuels ou multimédias numériques

Peter Stockinger, P.U.

Séminaire dans le cadre du

Master 2 *Document Numérique Multilingue* à l'Institut National des Langues et Civilisations Orientales (INALCO) pendant l'année universitaire 2006 - 2007

Paris, Institut National des Langues et Civilisations Orientales
2006

Sommaire

I/ Présentation générale

1.1/ Raisons et objectifs

La notion de l'hypertexte ou de l'hypertextualité est utilisée dans des acceptations très variées.

Nous la problématisons dans le contexte de R&D qui est le notre et celui de l'ESCoM. Celui-ci est concerné:

1. la constitution, le traitement et la diffusion de corpus audiovisuels et/ou multimédias numériques

2. et, cela, suivant un cadre théorique et méthodologique d'inspiration sémiotique (ou sémiologique), rhétorique et linguistique.

1.2/ Une définition générale (et banale) de l'hypertexte

L'hypertexte est un lien (discursif ou simplement technique) ou un ensemble de liens entre les parties composant un texte et/ou différents textes composant un « corpus ».

1.2.1/ L'hypertexte comme une pratique, une activité traditionnelle

On a l'habitude de considérer que texte en tant que support de communication organisant, exprimant et véhiculant de l'information (i.e. un "message") est intrinsèquement caractérisé par des liens (des rapports):

1. entre les éléments (thématiques, narratifs, discursifs, d'expression, ...) qui le composent et qui en font un "tout signifiant"
2. ainsi qu'entre ses éléments et des éléments pertinents composant son environnement, son "Umwelt" textuel (i.e. les autres textes avec lesquels il est en rapport).

Ainsi, la problématique hypertextuelle est une problématique propre à toute production textuelle (lato sensu) et ne peut pas être réduite au texte à support numérique.

Exemples :

1. *notes de bas de page* : lien entre un élément d'un paragraphe (d'un titre, …) et une référence bibliographique, une explication, … en bas de page
2. *index (thématique, de noms, …)* : renvoi entre une expression linguistique et la ou les pages ou cette expression est développées
3. *table de matière* : renvoi entre une liste de titres composant un texte et les parties développant chacun des titres
4. *citation* : lien entre le discours développé dans un texte T1 et une partie du discours développé dans le texte T2
5. *rappel, résumé, reformulation, ...:* lien/rapport entre un discours actuel dans la partie textuelle T1 et un autre discours dans la partie textuelle T2 servant au premier de référence

Références théoriques :

- le texte comme un réseau de relations, de rapports (cf. M. Serres) ;
- le texte au sens d'un intertexte (cf. Julia Kristeva, Roland Barthes, Gérard Genette) ;
- le texte-discours comme une production contrainte d'une collectivité de voix (M. Foucault sur l'ordre du discours, 1970)

1.2.2/ L'hypertexte au sens d'une caractéristique principale et centrale du texte numérique

Exemples :

1/ un lien interactif entre un élément textuel (lato sensu) et un autre élément du « même » texte :

- mot – segment textuel ;
- image – segment textuel ;
- image – image ;
- …

2/ un lien interactif entre un élément textuel d'un texte numérique T1 et un texte numérique 2 ou un élément d'un texte numérique 2 :

- mot/image – autre document ;
- mot/image – site web ;
- mot/image – film ;
- …

II/ Questions et interrogations au sujet de la notion « hypertexte/hypertextualité »

2.1/ Quatre grands problèmes

L'hypertexte ou l'hypertextualité posent quatre grands problèmes:

1. celui d'un cadre global de compréhension et de description du phénomène du rapport entre différentes parties d'un texte ou d'un ensemble de textes.

2. celui du développement et de l'utilisation de solutions technologiques appropriées aussi bien pour la conception, la réalisation, la communication et l'appropriation de produits ou prestations hypertextuels.

3. celui de l'intérêt pratique de prestations ou produits hypertextuels dans des contextes d'usage spécifiques et de l'adaptation des habitudes et des routines de travail aux "logiques" spécifiques de ces prestations et produits.

4. celui d'une compréhension plus globale (culturelle et anthropologique) de l'hypertexte en particulier et des technologies le rendant sinon possible au moins efficace dans une perspective de l'évolution des technologies du "symbolique".

Références biblio/webographiques :

– conférence de Michel Serres sur les nouvelles technologies comme externalisation des fonctions cognitives (Interstices, janvier 2006) ;

– entretien avec Fabrice Papy sur l'hypertexte et le document numérique (Paris, ESCoM-AAR, octobre 2004) ;

– conférence de Bruno Bachimont sur l'organisation de la connaissance à l'aide de la technologie hypertextuelle (Paris, Université de Paris VIII - ESCoM-AAR, janvier 2005).

2.2/ Vers une compréhension d'ensemble de l'hypertexte/ de l'hypertextualité

Nous sommes concernés, ici, par le premier problème évoqué ci-dessus. Pour notre travail, le fait que se tisse des liens/des rapports entre des éléments composant un texte ou un ensemble de textes (formant un "textscape") est subordonné au fait que le texte, tout texte:

– sert à la communication d'une information (lato sensu)

- véhicule donc un/des message(s) à l'adresse d'un destinataire "modèle" (au sens d'U. Eco)
- ce qui présuppose un "plan" (i.e. un scénario) de production du texte de la part de son "auteur" (connu, inconnu, individuel ou collectif, anonyme, humain ou non, ...).

Dans ce sens, l'hypertexte ou plutôt la fonction hypertextuelle:

- est une technique (un ensemble de techniques) parmi d'autres techniques composant les ressources (la compétence) nécessaire de l'"auteur" (au sens indiqué ci-dessus)
- pour produire - selon un ("son") plan - un texte
- dont il (l'"auteur") suppose
- qu'il (le texte) soit suffisamment "bon" (approprié, pertinent, ...)
- pour accomplir sa mission, i.e. de communiquer l'information à un destinataire "modèle" - communication qui peut se réaliser selon des registres rhétoriques fort différents (persuasive, didactique, poétique, ...).

La technique ou l'ensemble de techniques concernées par la fonction hypertextuelle est justement de mettre en rapport, selon un plan ou un scénario de production choisi, les éléments textuels concernés de façon telle que l'objectif de communication a une "bonne" chance (une, selon l'"auteur", bonne probabilité) d'être réalisé.

2.3/ Petit aperçu des techniques de production d'un texte

Les techniques les plus centrales qui interviennent dans la production d'un texte sont, on compte:

1. la technique de la sélection de l'information;
2. la technique (rhétorique et discursive) du traitement d'une information sélectionnée;
3. la technique de l'expression d'une information sélectionnée et traitée rhétoriquement (via de langages et de médias d'expression);
4. la technique de l'organisation spatiale ou temporelle de l'information véhiculée par un texte;
5. la technique de la "matérialisation" sur un support physique choisi de l'information véhiculée par un texte;
6. la technique de la communication (diffusion) de l'information véhiculée par le texte;
7. la technique de la conservation, de la mise à jour et de l'archivage de l'information véhiculée par le texte.

Références biblio/webographiques :

- conférence de Peter Stockinger sur la sémiotique et le document virtuel personnalisable présentant notamment les principaux plans sémiotiques pour apprehender le document ou le texte (ENST Brest, juin 2002)
- livre de Peter Stockinger sur les nouveaux produits d'information contenant des chapitres qui sont consacrés à l'organisation structurale d'un document ou texte (Paris, Hermes Science Publications 1999)
- livre de Peter Stockinger sur l'organisation structurale du document audiovisuel décrivant les principaux plans du document/texte audiovisuel (Paris, Hermes Science Publications 2003)

2.4/ Technique narrative et fonction hypertextuelle

Une technique particulièrement importante dans la production d'un texte selon un plan ou un scénario "auteur" est celle de la narration:

- non pas au sens obligatoirement restreint (i.e. au sens de "raconter une histoire")
- mais au sens d'un "enchaînement syntagmatique" (A.J. Greimas) des éléments "de contenu" (i.e. des informations sélectionnées) qui composent un texte.

Cet enchaînement narratif peut être :

- plus ou moins stéréotypé et faire partie de la "routine" de la production/ appropriation des informations (cf. les organisations syntagmatiques typiques d'un guide touristique, d'un reportage, d'un article de journal, d'un cours, ...) ;
- plus ou moins linéraire ou, au contraire, posséder des trames plus complexes faisant intervenir des "histoires parallèles", des "histoires dans l'histoires", des "digressions", etc. ;
- plus ou moins libre ou conditionné comme c'est le cas dans les jeux ou certains cours;
- etc.

On voit bien qu'une partie très importante de la problématique hypertextuelle - sinon toute la problématique - peut être parfaitement reformulée en terme de l'organisation narrative (i.e. syntagmatique) du texte ou d'un ensemble de texte formant - en analogie au paysage géographique - le paysage textuel (le "textscape") d'un acteur social.

Référence webographique :

- document enligne de Jean Clement consacrée à la question de l'hypertextualité comme discours, comme une forme discursive (Univ. de Paris VIII- Saint Denis, 1995)

2.5/ Résumé: trois notions à éclaircir

Pour résumer, la définition d'un cadre de compréhension globale de la problématique hypertextuelle en vue de son traitement motivé et aussi explicite que possible, renvoie:

1. à la compréhension sémantique de l'hypertexte, du lien hypertextuel ;

2. à la compréhension narrative de la fonction hypertextuelle dans un texte ou un ensemble de textes ;

3. et enfin à une définition plus explicite de ce qu'on peut entendre sous le terme "texte" ou encore "textualité" (numérique) au sens d'un support de communication véhiculant un message pour un destinataire "modèle".

III/ Notre contexte de R&D

3.1/ Le programme Archives Audiovisuelles de la Recherche

Dans le cadre de notre programme Archives Audiovisuelles de la Recherche (mis en place par nous-même en 2001 à la Fondation Maison des Sciences de l'Homme), nous poursuivons une politique de constitution et d'exploitation d'un patrimoine scientifique audiovisuel en sciences humaines et sociales - patrimoine enligne et, pour le moment, publiquement accessible.

Cette politique se concrétise en un ensemble d'activités de production, de la description-indexation, du montage et de la publication enligne d'événements scientifiques sur le site portail de notre programme AAR.

Toutes ces activités sont encadrées et progressivement améliorées à l'aide de partenariats scientifiques et techniques et de projets de R&D français et Européens tels que, par exemple, le projet français SAPHIR et le projet Européen LOGOS.

Références:

- document enligne présentant d'une manière plus systématique le programme AAR (Paris, ESCoM, septembre 2005)
- film institutionnel du programme AAR réalisé par Christian Veits et Richard Fillon (Paris, ESCoM, septembre 2005)
- plaquette officielle du programme AAR (Paris, ESCoM, 2005 - 2006)
- organigramme du programme AAR (Paris, ESCoM, 2006 - 2006).

3.2/ Les trois principales activités

Afin de mieux comprendre la notion de l'hypertexte ou de l'hypertextualité, nous allons donc le replacer dans notre contexte de travail à l'ESCoM - contexte caractérisé par les trois axes suivants:

1. constitution, gestion et diffusion de patrimoines numériques audiovisuels de la recherche
2. travail en et sur la communication spécialisée (notamment : scientifique, pédagogique, de « vulgarisation », journalistique, …)
3. travail, plus particulièrement, sur les différents genres pertinents à la compréhension du discours scientifique : entretien, conférence, séminaire, cours, exposé, témoignage, reportage, argumentaire, débat, …

Référence webographique :

- hyper-documentation enligne présentant en détail les activités de l'ESCoM et du programme AAR (Paris, ESCoM, septembre 2005).

3.3/ Les corpus audiovisuels

La plus grande partie du fonds audiovisuel généré et géré par notre programme AAR, est composé d'entretiens, de conférences, de cours et de séminaires, ... filmés, numérisés et diffusés enligne sur le web.

Autrement dit, il s'agit ici d'un corpus de documents/textes audiovisuels doublement spécifiques:

1. il s'agit essentiellement d'un discours oral fixé sur image qui organise, exprime et véhicule le message contrairement, par exemple, aux documents/textes audiovisuels appartenant au genre "documentaire audiovisuel" où le message est organisé, exprimé et véhiculé par trois plans constitutifs: image - son - voix off (du narrateur);
2. il s'agit essentiellement de documents/textes audiovisuels caractérisés par des genres de discours qui font partie de la communication scientifique et/ou didactique.

Il s'agit donc d'un corpus homogène quoique déjà assez important d'un point de vue quantitatif (environ 2800 heures de vidéos enligne, à l'heure actuelle) et ouvert (s'y ajoute, chaque mois, entre 20 et 40 heures de vidéos). Enfin, ce corpus homogène est complété par un corpus en revanche plutôt eclaté d'environ 250 heures de vidéos composé de tournages d'expositions scientifiques et d'événements à caractère culturel ou social, de documentaires, de reportages, de témoignages, etc.

3.4/ Portails et sites web concernés

L'ensemble du corpus est accessible via le site portail de notre programme Archives Audiovisuelles de la Recherche.

Trois portails thématiques sont actuellement en construction et constituent les "pilotes" de nos actuelles activités de R&D dans le domaine de la republication de corpus audiovisuels afin de les adapter à des contextes d'usage spécifiques et de les ouvrir à un "marché" intrinsèquement multilingue et multiculturel:

1. portail thématique "Diversité Culturelle et Linguistique" (DLC);
2. portail thématique "Patrimoine Culturel des Minorités et des Peuples Indigènes" (PCI);
3. portail thématique "Mondialisation et Diversité Culturelle" (MDC)

Voici encore un autre site important pour notre propos:

– "Guanacache. Les eaux de la soif". Site consacré au patrimoine du peuple Huarpe en Argentine proposant un accès aux informations sous forme d'un vidéo-livre interactif, d'une encyclopédie thématique, d'une version allemande, de dossiers pédagogiques, etc.

IV/ Notre problématique de R&D

4.1/ Les quatre problématiques principales

Voici les quatre problématiques qui motivent nos travaux de R&D à l'ESCoM et dans le cadre du programme des AAR:

1. Problématique I: permettre une localisation aussi précise que possible de l'information cherchée dans un corpus audiovisuel et/ou multimédia
2. Problématique II: contribuer à une bonne appropriation et exploitation du « contenu » du corpus
3. Problématique III: ouvrir le contenu spécialisé à un public multilingue et culturellement différent
4. Problématique IV: valoriser les fonds audiovisuels et/ou multimédias de nos sites.

4.2/ La localisation de l'information dans un corpus audiovisuel et/ou multimédia

Cette problématique concerne, d'une manière générale, la satisfaction d'un besoin (intérêt, désir, ...) cognitif de l'utilisateur (lecteur, spectateur, ...) d'une archive (bibliothèque, espace d'information, ...). Ce besoin (intérêt, désir, ...) s'exprime sous forme d'une absence, d'un manque, ... de connaissances ou encore, d'une manière moins catégorique, sous forme d'un motif de clarification, de "correction", de "mise à niveau", etc.

Ceci dit, la satisfaction d'un besoin (intérêt, désir, ...) d'information ne peut pas être réduite à la simple localisation d'une information supposée pertinente (i.e. répondant au besoin, intérêt, ... exprimé). Elle doit être comprise plutôt dans un contexte plus large d'une "veille d'information" comprenant:

1. la localisation de l'information à proprement parler dans un espace d'information (web, bibliothèque, archive, ...);
2. l'exploration du contexte potentiellement pertinent d'une information localisée;
3. l'évaluation (la hiérarchisation préférentielle, i.e. appréciative) des informations pouvant répondre à un besoin (intérêt, désir, ...).

Il existe des "solutions" typiques pour répondre (d'une manière plus ou moins satisfaisante) à un tel besoin ou intérêt d'information:

- moteurs de recherche;
- technologies relevant du *data* (ou *text*) *mining* et du "*knowledge management*";
- catalogues thématiques (thesaurus, ontologies, ...) faisant référence à un standard (tel que le Dewey, en bibliothéconomie) ou non;

– sites à thème fonctionnant comme des points d'accès à un ensemble de ressources pertinentes (au thème) et souvent classifiées en quelques catégories pragmatiquement saillantes.

Les difficultés de base que se posent à une localisation pertinente d'une information dans le fonds des AAR, sont:

– l'existence d'un corpus de ressources audiovisuelles quantitativement déjà assez important (environ 2300 heures de vidéos qui augmentent à raison de 25 à 35 heures de vidéos par mois);
– le fait que le contenu référentiel (les "domaines de connaissance") est très diversifié et éclaté (il couvre en effet tout l'éventail de la recherche en sciences humaines et sociales);
– le fait que ce contenu est monolingue (i.e. qu'il s'exprime en une langue);
– le fait surtout - et finalement - que la vidéo est un document linéaire impliquant la possibilité qu'une information recherchée ne peut être abordée et traitée après une période relativement longue pendant laquelle sont traitée, dans la même vidéo, d'autres questions.

En faisant abstraction de la problématique de l'information monolingue (cf. ci-après), trois "solutions" semblent s'imposer mais chacune possède ses limites motivant ainsi la mise en place et le financement de très grands programmes de R&D aussi bien en France qu'en Europe (cf. à ce propos le site de l'ANR consacré à la problématique du corpus en SHS; le site du Programme RIAM en France et le site *Information and Communication Technologies* de la Communauté Européenne):

1. La transcription du discours oral en données textuelles (technologie "*speech to text*") en tant que procédure préalable pour une recherche et une localisation de l'information dans le flux audiovisuel (i.e., pour être plus précis, sonore) lui-même. La qualité des résultats de la transcription semble être très variable dépendant directement de paramètres tels que le débit, l'accent, le timbre, l'émotivité, le vocabulaire/le langage utilisé, etc.

2. Les technologies *data* ou *text mining* présupposent, dans le cadre des archives audiovisuelles, les résultats de la transcription du discours oral en données textuelles. Elles proposent des analyses statistiques de la fréquence de chaînes de caractère (termes, en cas d'un traitement terminologique des résultats "*speech to text*") et/ou de collocation de pairs de chaînes de caractère ("termes"). Les résultats peuvent être soit interprétés par un expert humain soit encore subir un deuxième cycle de traitement sous forme d'une analyse linguistique robuste et simple des chaînes de caractère et de leur classification thématique. Ainsi, dans le cas idéal, les informations développées dans un corpus audiovisuel sont transcrites en données textuelles, analysées au niveau statistiques et linguistiques et présentées sous forme de thèmes (topiques) représentant les connaissances du corpus. Les thèmes eux-mêmes peuvent être définis sous formes de réseaux ou graphes conceptuels ce qui permet donc, en principe, une forme de raisonnement

à partir d'une requête exprimée par l'utilisateur (raisonnement pouvant s'exprimer, par exemple, sous forme de conseil, d'aide à l'exploration, d'adaptation de chemins d'exploration, etc.). Mais, dans la pratique, on est encore assez loin de cette situation idéale ...

3. L'annotation et l'enrichissement des ressources numériques à l'aide de méta-données, i.e. de descriptions de leur "contenu" suivant un modèle ou schéma et répondant à des contextes d'utilisation (par exemple: le schéma LOM - *Learning Object Modelling* - pour un contexte d'enseignement/d'apprentissage; le schéma DC - *Dublin Core* - pour un contexte bibliothécaire; le schéma MPEG 7 - *Moving Picture Experts Group 7* - pour un contexte d'usage de stockage, de gestion et d'utilisation/d'exploitation de données audiovisuelles, etc.). Outre le fait qu'il n'existe pas encore des schémas de description prenant en compte réellement la structure d'un texte (i.e. sa sémiotique), c'est surtout le travail "manuel" de description-indexation qui est énorme et n'a donc, tel quel, peu de chance à être utilisé systématiquement sur des grands corpus audiovisuels (mais aussi textuels stricto sensu). D'où le recours aux technologies *text mining* et *knowledge management* comme des outils permettant d'automatiser certains aspects de la description (segmentation, indexation, ...) thématique du contenu d'un corpus audiovisuel.

Références (au sujet de la transcription discours oral - donnée textuelle stricto sensu):

- Pour plus d'informations sur le domaine de la transcription du discours oral (de la langue parlée) en donnée textuelle, cf. *Groupe du Traitement du Langage Parlé* du LIMSI/CNRS;
- *MediaSpeech*, une technologie développée par la société *VECSYS* (et le LIMSI-Groupe du Traitement du Langage Parlé) pour l'indexation automatique de corpus audiovisuels;
- La société *Sail Labs Technology* proposant des solutions de transcription de contenu multimédia, de l'organisation de ce contenu en des catégories thématiques et aussi des aides à la traduction (semi-automatique) des mots clés extraits lors du processus de la transcription.

Références (au sujet du data ou text mining et du knowledge management):

- A consultez le site *Cogitant* du *LIRMM* (CNRS - Université de Montpellier III) et du *LERIA* (Université d'Angers) proposant une suite de logiciels pour la création et utilisation de graphes conceptuels - suite de logiciels utilisés dans les deux projets Saphir et Logos;
- A consultez le site de la société *ARISEM* proposant ses solutions technologiques *Kaliwatch* ou *Sprint*;
- Anette Béguin-Verbrugge (Univ. de Lille III): document enligne sur l'intérêt du KM pour le SIC (systèmes d'information et de communication)

Références (au sujet de la description de ressources audiovisuelles):

- Excellent compte rendu d'une journée *AFNOR/BNF* (juin 2005) sur le bon usage des métadonnées dans la gestion/utilisation des ressources électroniques;
- Présentation très claire du standard *LOM fr* (i.e. qui est une adaptation au contexte français) sur le site *EducNet*;
- Présentation succincte de *MPEG 7* par Wikipédia (pour avoir des informations détaillées sur les différents schémas MPEG 7: *MPEG 7 Overview* de Leonardo Chiariglioni)

4.3/ L'appropriation du "contenu" d'un corpus audiovisuel

L'appropriation de l'information désigne cette pratique très complexe qui met en scène l'"utilisateur" (i.e. le lecteur, spectateur, ...) d'une ressource numérique (audiovisuelle, textuelle, ...) qui exploite celle-ci selon à la fois:

1. ses besoins, intérêts, désirs, objectifs
2. et ses compétences, ses références culturelles, ses valeurs, etc.

L'appropriation recouvre à la fois la lecture, l'interprétation, la réécriture (mentale ou physique) d'une ressource numérique, son adaptation (au profil - culturel - de l'utilisateur), sa re-utilisation plus ou moins "fidèle" ou libre par l'utilisateur, etc.

En d'autres termes, l'utilisateur - ou, comme on dit plutôt en sémiotique, le destinataire - devient, en s'appropriant une ressource numérique, lui-même auteur d'une nouvelle ressource qui peut être (mais ne doit pas obligatoirement l'être):

- plus ou moins similaire à la ressource d'origine;
- une "simple" construction mentale de l'utilisateur sans acquérir le statut d'une ressource (numérique, analogique) extériorisée et plus ou moins pérenne;
- à son tour une ressource source pour des nouveaux processus d'appropriation débouchant sur la création (simplement mentale et/ou physique) de nouvelles ressources.

Cette activité d'appropriation de ressources cognitives a fait l'objet de différentes approches - sociolinguistique, sémiotique, rhétorique, cognitiviste, informatique, ... - du discours. Pratiquement parlant, elle renvoie:

- au fait déjà mentionné plusieurs fois de la republication de données audiovisuelles (textuelles, ...) sous forme de genres de publication particuliers adaptés à des pratiques ("routines sociales") spécifiques,
- ainsi qu'à celui de la mise à disposition à des communautés d'utilisateurs d'outils et de procédures leurs permettant de republier selon leurs besoins et intérêts des corpus audiovisuels (textuels, ...).

Ces deux points constituent les principaux enjeux des deux projets de R&D déjà cités: *Saphir* et *Logos*.

Références:

- A consultez notre enseignement dispensé à l'INALCO (Paris) pendant les années 2003, 2004 et 2005 ainsi qu'à l'Université de Cuyo à Mendoza (Argentine) sur la republication de corpus audiovisuels ainsi que les réalisations de dossiers thématiques enligne par les étudiants ayant suivi cet enseignement.
- *Semiotic video processing and personalised publishing.* de Peter Stockinger (Conférence donnée dans le cadre d'un workshop organisé par le projet Européen Chiron à Paris 2005)

4.4/ La communication d'une information dans un contexte multilingue et multiculturel

Le "contenu" du fonds audiovisuel des AAR est un contenu monolingue (comme c'est le cas de pratiquement toutes les archives audiovisuelles dans le monde). Le problème qui se pose ici est, bien sûr, comment rendre accessible un contenu monolingue à des locuteurs qui:

- soit ne connaissent pas la langue "source" utilisée dans une vidéo diffusée sur le web,
- soit ne possèdent que trop peu de compétence linguistique dans la langue "source" pour pouvoir évaluer le contenu proposé par rapport à leurs attentes.

Toute en gardant le principe d'une "*lingua franca*" telle que l'anglais (ou plutôt une forme simplifiée et "internationalisée" de l'anglais), il s'agit ici:

- de produire des aides linguistiques (pas obligatoirement des traductions, stricto sensu) permettant aux intéressés potentiels d'évaluer au moins grosso modo le contenu d'une vidéo produite dans une langue qu'ils ne connaissent pas (ou peu)
- et, de nouveau, de mettre à la disposition des communautés d'utilisateurs des outils simples d'annotations linguistiques de ressources numériques (audiovisuelles, textuelles, ...).

Ces aides linguistiques peuvent être accompagnées d'annotations d'ordre culturel (relatif au contenu, à l'auteur, ...) et constituer ainsi des solutions pragmatiques de médiation ("interculturelle") d'un contenu "culturellement marqué" tel que celui proposé par les AAR.

Références:

- A consulter notre enseignement consacré à la republication de corpus audiovisuels en contexte multilingue et multiculturel (Lisbonne, septembre 2006) et les premières réalisations de versions bilingues enligne d'entretiens ou de conférences par les étudiants ayant suivi cet enseignement.

4.5/ La valorisation d'un corpus audiovisuel

Par valorisation, on entend la "mise en valeur" des prestations d'un espace d'information comme, par exemple, les prestations en terme de contenu scientifique de la part des AAR.

Une opération de valorisation ne doit pas être comprise dans un sens étroitement commercial - elle est de nature profondément politique dans la mesure où elle doit rendre visible et "tangible" la spécificité et la "valeur ajoutée" d'un espace d'information tel que celui des AAR.

Or, une partie importante de la communication de valorisation d'un fond tel que celui des AAR repose sur l'identification des segments qui peuvent l'aider dans cette tâche: segments possédant un impact émotionnel ou psychologique fort, segments proposant une vue d'ensemble sur une question donnée, segments capables d'éveiller la curiosité de l'utilisateur, etc.

Cette problématique est, actuellement, encore peu explorer dans le cadre de nos activités de R&D mais constitue un des axes dans le cadre du projet Saphir. Via une analyse et indexation préalable d'un corpus audiovisuel, il s'agit d'y isoler des segments "valorisant" et de les publier ensuite sous forme de "bandes d'annonce" soit sur le site des AAR lui-même soit sous forme d'annonces publicitaires à distance sur d'autres sites.

V/ Localisation et exploration d'une information dans un corpus audiovisuel – exemple et scénario « idéal »

5.1/ L'hypertextualité et l'exploration d'un corpus audiovisuel

Dans le contexte des AAR au sens d'un SIC (système d'information et de communication), l'hypertextualité est une fonction particulière. Rappelons que la localisation d'information comprise comme une activité particulière d'une veille d'information recouvre:

- l'identification ou la localisation stricto sensu d'une information pertinente (par rapport aux besoins ou intérêts, désirs, ... d'un utilisateur/visiteur);

- l'exploration du ou des contextes potentiellement pertinents pour "approfondir" la compréhension et la connaissance d'une information localisée;

- l'évaluation (la hiérarchisation préférentielle) des ressources "contenant" une information cherchée par rapport à leur intérêt pour l'utilisateur.

- La fonction hypertextuelle est concernée surtout par la deuxième et aussi, d'une manière peut-être moins évidente, par la troisième étape dans l'activité de la veille.

5.2/ Exemple: veille d'information sur le portail thématique audiovisuel "Patrimoine Culturel des minorités et des peuples Indigènes" (PCI)

Comme la plupart des SIC "standard", aussi le PCI propose des outils de base pour accéder à son fonds audiovisuel:

- un moteur de recherche multilingue;
- un catalogue thématique représentant grossièrement sous forme de grands taxèmes les connaissances du domaine de référence;
- un catalogue "collections";
- un index des noms d'auteurs ayant contribués à la constitution du fonds.

5.3/ Le moteur de recherche: état actuel et perspectives

Le moteur de recherche, développé par Francis Lemaitre (ESCoM-FMSH), s'appuie notamment sur une indexation systématique, structurée, normalisée et multilingue du fonds audiovisuel PCI:

- *systématique*: chaque événement scientifique ainsi que chaque "partie" d'un événement scientifique du PCI fait objet d'une indexation;
- *structurée*: l'indexation par mots clés suit un cadre méthodologique et thématique que nous avons spécifié et développé dans le cadre d'une sémiotique textuelle (pour des références, cf. ci-après);
- *normalisée*: le processus de l'indexation suit une procédure établie (priorité aux entités nommées; consignes grammaticales et orthographiques; ...);
- *multilingue*: la plupart des index sont traduits obligatoirement en anglais et espagnol ainsi que d'une manière assez récurrente en allemand, polonais, russe, italien, ...

Ceci dit, il subsiste, actuellement, d'importantes limitations:

- l'aspect multilingue de l'indexation n'est pas encore réellement mis en valeur par le moteur de recherche;
- l'indexation (aussi bien des fichiers que de n'importe quel segment) se fait d'une manière exclusivement manuelle (cf. les enjeux de l'intégration des technologies "speech to text", du "text mining" et du "knowledge management");
- tout en étant "bridés", les index sont produits librement - il n' y a pas recours à un "vocabulaire contrôlé" ;
- tout en étant classifiés dans les grands taxèmes de l'analyse thématique, les index devraient être classifiés (d'une manière automatique ou moins interactive) en des catégories plus spécialisées (correspondant au catalogue thématique du PCI);
- enfin, il manque au moteur de recherche actuel un composant - même très simple - d'analyse linguistique permettant de "traduire" la requête d'un utilisateur vers le vocabulaire composant l'index même si cette requête ne reproduit pas exactement tel ou tel "mot" composant l'index (cf. la distinction nécessaire à maintenir entre le langage interne de description-indexation qui est un métalangage et les langages propres aux utilisateurs et à l'aide desquels ils expriment leurs besoins, intérêts, désirs d'information - langages d'utilisateurs qui doivent être "traduits" vers le langage de description interne et vice versa).

5.4/ Le catalogue thématique: état actuel et perspectives

Le catalogue thématique propose un accès via une liste structurée de thèmes représentant les connaissances du domaine PCI. Malgré sa forme actuellement encore simple et statique, il s'agit ici d'un outil - potentiellement - très puissant.

Chacun des thèmes composant le catalogue propose une liste d'une ou de plusieurs documents audiovisuels de longueur assez variable. C'est le cas, par exemple, du thème "Hygiène et santé" qui pointe actuellement (en décembre 2006) sur 5 événements scientifiques dont chacun est publié sous forme d'un "vidéo-livre

interactif" composé d'un ensemble de "chapitres vidéos" (cf. par exemple, le vidéo-livre de l'entretien avec Gregory Maskarinec sur le chamanisme au Népal).

Ceci dit, contrairement à ce qui est proposé actuellement, un enjeu central pour une version ultérieure du catalogue thématique est que le lien hypertextuel qui a son origine dans un thème spécialisé (i.e.: Hygiène et santé), ne pointe pas sur le document audiovisuel en entier (i.e. le vidéo-livre contenant l'entretien avec Gregory Maskarinec) mais directement sur le « passage » ou, plutôt les passages où sont développées les informations relatives au thème choisi.

Cela présuppose la segmentation et l'indexation du flux audiovisuel en séquences thématiquement pertinentes - un travail "manuel" assez lourd exigeant les outils nécessaires ainsi qu'une bonne représentation des connaissances (exprimée, par exemple, sous forme de "patterns", de "schémas" thématiques appelés techniquement "graphes conceptuels").

Les thèmes composant la liste du catalogue actuel sont indépendants les uns des autres - ils ne sont pas "connectés" entre eux. Cela a des conséquences importantes sur l'activité de la veille d'information dans le corpus audiovisuel.

Par exemple, un thème à lui seul donnant accès à une liste de passage ne peut représenter que trop partiellement une information recherchée – l'information recherchée est mieux représentée par plusieurs thèmes "ensemble" (par exemple : au lieu de "pratiques d'hygiène et de santé", plutôt "pratiques d'hygiène et de santé" chez les peuples indigènes habitant la région des Andes").

Pour réaliser cet objectif, les thèmes doivent donc être définis sous forme de configurations, de "patterns" ou "schémas" que sont des modèles de connaissances d'un domaine. Techniquement, ces configurations ou schémas forment des graphes conceptuels. Un graphe conceptuel est un objet constitué de n thèmes, de relations (unaires, binaires, ...) entre ces thèmes et de contextes d'enchâssement faisant en sorte qu'un graphe entier peut être "encapsulé" en un autre graphe. Enfin chaque thème - correspondant grosso modo à la notion "concept" dans la théorie des graphes conceptuels (J. Sowa) - possède obligatoirement un champ "référent" qui est composé, dans notre cas, soit par les mots clés provenant de l'indexation soit par le time code localisant un graphe et les mots clés sur le flux audiovisuel.

Former des graphes conceptuels représentant des modèles de connaissances du domaine PCI, présuppose que :

- tous les thèmes soient définis dans une ontologie référentielle (i.e. une ontologie spécifiant le vocabulaire du langage pour décrire le domaine de connaissances PCI);
- toutes les relations soient définies dans une ontologie relationnelle (i.e. une ontologie spécifiant le vocabulaire du langage pour décrire comment se contracte les différents thèmes pour former des "modèles de connaissances".

La possibilité d'utiliser des graphes conceptuels en veille d'information pour un corpus tel que celui du PCI, permettrait donc d'exprimer une requête, une demande d'information qui sollicite le recours à plusieurs thèmes. Mais l'utilisation des graphes conceptuels permet également:

- La hiérarchisation de passages (segments) audiovisuels (textuels) selon le degré de proximité entre une demande formulée par l'utilisateur (et "traduite" en un graphe conceptuel) et son contenu (également "traduite", voire plutôt: indexer en graphe conceptuel). Même si aucun passage ne se trouve dans le corpus qui développe exactement l'information recherchée certains passages peuvent développer des informations plus ou moins proches. Par exemple, même s'il n'y a pas, dans le corpus PCI, de passages relatifs aux pratiques d'hygiène chez les indigènes Huarpe on y trouve, néanmoins, des passages sur les pratiques d'hygiène chez les indigènes Mapuche et aussi des passages relatifs à d'autres pratiques chez les indigènes Huarpe, ...).

- (Via les relations entre les thèmes et quelques règles de production de graphes) la construction "dynamique" de contextes potentiellement pertinents pour pouvoir explorer et évaluer une information localisée dans un segment du corpus PCI. Par exemple : le passage sélectionné « exemples de la pratique d'hygiène et de santé chez les indigènes Huarpe », peut-il être complété par d'autres passages proposant, des informations sur ce type de pratique chez d'autres indigènes dans la même région ou ailleurs ou encore des informations sur d'autres pratiques chez les indigènes Huarpe, etc.?

Ainsi, dans une version ultérieure (pour le PCI, un premier prototype sera disponible en septembre 2007), le catalogue thématique construira dynamiquement, à partir de la requête d'un utilisateur, les contextes potentiellement pertinents soit pour identifier une information recherchée soit pour explorer, à partir d'une information identifiée et localisée, son univers sémantique.

5.5/ Le catalogue "collections (des vidéos)": état actuel et perspectives

Le catalogue "collections (des vidéos)" présente aussi, dans sa version actuelle, un outil très simple et approximatif d'accès au fonds audiovisuel et de localisation d'une information pertinente dans ce fonds.

Ceci dit, l'hypothèse qui "se cache" derrière cet outil d'accès, c'est celle que toute ressource audiovisuelle composant le fonds du PCI peut être caractérisée par un, voire plusieurs genres de discours (entretien et conférence, certes, mais aussi témoignage, biographie, récit de découverte, chronologie d'événements, description rhétorique - portrait, étopée, prosopopée -, homélie, etc.).

Il s'agit donc à la fois d'examiner les principaux genres de discours qui organisent le corpus PCI et de spécifier une ontologie de genres de discours qui, formellement, fonctionne de la même façon que l'ontologie référentielle citée ci-dessus.

En "croisant" les thèmes interprétés par leurs référents (les mots clés provenant de l'indexation) avec la liste des genres de discours, des requêtes ou encore des guides de requête/d'exploration du genre "récits de découverte de peuples indigènes à ..."/"tableaux de moeurs de peuples ..."/portraits de grands ethnologues ..."/ etc. deviennent possibles. On pourra également filtrer les requêtes par rapport aux actes de discours "de base" tels que "définition", "description - explication - exemple", synthèse, etc.

Références:

– hyper-documpentation enligne présentant en détail les activités de l'ESCoM et du programme AAR (Paris, ESCoM, septembre 2005).

VI/ L'appropriation de l'information dans un fond audiovisuel – un scénario

6.1/ Le vidéo-livre interactif sur le portail PCI

L'appropriation (la lecture, la compréhension, …) d'une information se fait, dans le cadre du PCI et de tous les autres sites appartenant à l' « anneau » AAR, sous forme d'une consultation libre et donc, selon l'utilisateur, plus ou moins systématique, de sites web dynamiquement générés (à partir des données contenues dans la base de données) pour un « événement scientifique » (un entretien, une conférence, un séminaire, un documentaire, …).

L'élément central et principal y est la vidéo de l'événement. Cette vidéo est publiée en analogie au genre « livre ». D ans ce sens on parle du « vidéo-livre interactif », d'une vidéo consultable via une "table de matière" (un sommaire) interactif dont les titres renvoient au passages vidéos concernés (à la manière de la table de matière dans un livre « classique » qui renvoie aux chapitres du livre).

Prenons l'exemple d'un entretien réalisé à la FMSH à Paris en novembre 2005 avec l'historien bolivien René Arze sur les guerres dans l'histoire de la Bolivie et le rôle déterminant dans ces guerres joué par les combattants indigènes.

Cet entretien a été filmé dans son intégralité et enregistré sur des casettes DV. Une partie a été ensuite ré-numérisée (en format *avi*) et, puis, découpée, montée, indexée et, enfin, compressée et publiée en trois formats différents sous forme d'un site web dynamique accessible à partir d'ici.

Le site web dynamique (i.e. généré à partir des données relatives à cet entretien et stockées dans la base des données des AAR) ressemble, métaphoriquement parlant, à une sorte de livre.

La page d'accueil du site dédié à l'entretien avec René Arze fait allusion à la page de couverture d'un livre où on trouve d'une part des informations signalétiques relatives à l'événement et d'autre part une très brève présentation du contenu de l'entretien ainsi que du chercheur lui-même.

Sur la partie gauche de la page d'accueil, on trouve le sommaire, la table des matières de ce vidéo-livre" interactif enligne (vidéos, curriculum vitae, photos d'archives, ...). En cliquant sur les différents intitulés du sommaire, on accède aux "chapitres" correspondants du "vidéo-livre".

Le "chapitre" le plus central est évidemment le chapitre donnant accès aux différentes parties de l'entretien avec l'historien bolivien. Ces différentes parties composent, pour rester avec la métaphore livresque, les différents sous-chapitres du chapitre "vidéo". Ces sous-chapitres sont organisés essentiellement selon un point de vue thématico-chronologique, c'est-à-dire qu'un sous-chapitre :

- possède une certaine unité d'un point de vue du contenu;
- est positionné par rapport aux autres sous-chapitres en respectant l'ordre temporel du déroulement de l'interview.

Il s'agit ici, en d'autres mots, d'un genre particulier de publication (ou de republication) d'un événement scientifique (i.e. d'un entretien). Il constitue actuellement la seule clé offerte aux visiteurs du site PCI (et de tous les autres sites faisant partie de l'anneau AAR) pour s'approprier (lire, interpréter, ...) le contenu des ressources audiovisuelles.

Pour diversifier les possibilités d'appropriation, un travail de R&D important est actuellement entrepris portant sur la problématique de la republication de corpus audiovisuels (et, plus largement, multimédias).

Références:

- Pour plus d'informations sur le genre "vidéo-livre interactif", cf. le cours enligne "The interactive hypermedia book" (IHB) dedicated to the Huarpe civilisation" (Lisbonne, septembre 2006)

6.2/ La republication de corpus audiovisuels

Cette problématique concerne l'utilisation d'un même corpus audiovisuel (multimédia) pour des genres de publication différents où un genre de publication est supposé répondre à des contextes d'usage particulier, voir à des profils spécifiques des lecteurs des documents audiovisuels.

Par exemple, le corpus audiovisuel du site portail PCI peut servir à la fois à l'enseignement (secondaire ou supérieur), à l'apprentissage "libre" ou personnel, à la recherche, à un certain type de journalisme, à l'éveil de la reconnaissance et du respect de la diversité des cultures et des mœurs, à des usages professionnels comme, par exemple, au personnel travaillant pour les ONG, et ainsi de suite.

A ces types d'usages correspondent grosso modo et plus ou moins bien des genres de publication tels que des cours, des documentaires, des reportages, des vidéo-lexiques, des "high-lights", des débats virtuels, des versions plus ou moins libres et dans différentes langues, etc.

Les principaux enjeux ici sont les suivants:

1. description et modélisation des genres de publication et mise en place d'une sorte de "grammaire de production" de modèles de publication audiovisuelle et/ou multimédia;

2. identification des principales activités et tâches composant le processus de la republication;
3. développement d'un environnement technique, d'un outil informatique de republication de corpus audiovisuel.

Ces trois enjeux constituent une partie essentielle des deux projets de R&D déjà mentionnés Saphir et Logos.

6.3/ L'exemple du site « Guanacache. Les eaux de la soif »

Le site "Guanacache. les eaux de la soif" (titre d'un livre du journaliste et auteur argentin Gregorio Manzur) est dédié à la valorisation et diffusion de l'héritage culturel lato sensu du peuple Huarpe - un peuple indigène vivant dans des conditions extrêmement difficiles dans le nord de la région de Mendoza en Argentine.

Ce site proposera des informations sur le peuple Huarpe et son monde via plusieurs genres de publication:

- "highlights"
- dossiers thématiques ou encyclopédie thématique;
- vidéo-lexique;
- dossiers pédagogiques;
- documentaires virtuels;
- versions multilingues;
- clips vidéos pour "nomades (mobile phone, ipod, psp, ...)
- etc.

Il s'agit d'un chantier de travail dont l'objectif est de développer localement un ensemble de services d'information (scientifique, didactique, ludique, professionnelle, multilingue, ...) qui doivent être disponibles pour le site portail PCI dans son ensemble (première version prévue pour le mois de septembre 2007).

6.4/ La place de l'hypertextualité dans le processus de la republication

La problématique hypertextuelle dans le contexte de la republication de corpus audiovisuels est concernée notamment:

1. par la mise en place de syntagmes ou encore par l'intégration syntagmatique d'un ensemble de segments audiovisuels en un "tout cohérent";
2. par l'"enrichissement" d'un contenu donné sous forme de commentaires ("annotations") ou encore sous forme de mise en relation du contenu avec d'autres ressources d'information.

Dans le premier cas, la problématique hypertextuelle est concernée plus particulièrement par la dimension narrative dans la republication de corpus

audiovisuels. Tout genre textuel (cours, documentaire, dictionnaire, livre, reportage, ...) possède son ou ses plans plus ou moins stéréotypés d'intégration selon lesquels des parcours de lecture, d'appropriation sont soit offerts soit imposés aux lecteurs (visiteurs, apprenants, ...).

Dans le deuxième cas, il s'agit de transformer un texte (un segment, un paragraphe, ...) donné et appartenant au corpus "de départ" en une "nouvelle" version - nouvelle version qui est plus ou moins proche ou, au contraire, éloignée de la version initiale mais qui n'est rien d'autre que le résultat partiel du processus de republication.

Dans les deux cas, la fonction hypertextuelle permet, si elle est bien contrôlée, d'avancer sur le chemin de la republication d'un corpus donné.

VII/ Localisation et exploration d'une information dans un corpus audiovisuel – exemple et scénario « idéal »

7.1/ « Travailler avec un corpus audiovisuel »

"Travailler avec un texte" (dans notre cas: avec un document audiovisuel) veut dire donner au visiteur/utilisateur d'un site tel que celui des AAR la possibilité de le "manipuler" à ses fins, i.e. selon ses besoins et ses intérêts.

L'exemple vraisemblablement le plus parlant est celui de la "table de travail" traditionnelle du chercheur, de l'enseignant ou de l'étudiant - table de travail qui constitue un espace à la fois physique et social de l'acquisition et de la production de connaissances. Concrètement parlant, le travail avec un texte s''inscrit toujours dans un projet quelconque dans le cadre duquel le texte constitue soit une ressource, un moyen pour atteindre un but quelconque soit un but en soi.

Toujours est-il que le travail avec (et autour) du texte se décompose en un ensemble d'activités typiques telles que:

- l'enregistrement et le classement (personnel) d'une (nouvelle) ressource textuelle;
- la prise de note (i.e. l'extraction des parties textuelles particulières pertinentes pour le projet de l'utilisateur);
- la production de résumés, de synthèses, de compte rendus, ... d'une partie textuelle;
- la production de commentaires sur telle ou telle partie en relation avec le projet;
- la vérification de telle ou telle partie textuelle quant aux informations fournies (sous forme, par exemple, d'une recherche de nouvelles informations, d'informations complémentaires, ...);
- l'ajout de références, de "liens" à d'autres ressources avec comme but de rendre plus clair le propos d'une partie textuelle traitée, de le "relativiser", de le critiquer, de le compléter, etc.;
- la traduction (telle quelle, résumant, sous forme d'une version plus ou moins libre, ...) de telle ou telle partie textuelle;
- l'enregistrement et le classement (personnel) de toutes ces activités de "traitement d'information textuelle";
- la mise à jour de s résultats de telle ou telle activité spécifique de "traitement d'information";
- et, enfin, la publication (lato sensu) de ces activités (ou d'une partie de ces activités) sous forme d'une simple réunion de pages manuscrites ou

imprimées à usage purement personnel, sous forme d'un modèle préétabli de republication, voire sous forme d'une véritable publication originale.

7.2/ Site portail et collaboratoire – l'espace numérique de travail

Comme on le sait, toutes ces activités de travail sur et autour d'un texte peuvent être réalisées soit par une personne seule, soit par un groupe de personnes travaillant ensemble pour réaliser un projet commun.

Peu importe s'il s'agit d'un travail individuel ou d'un travail collectif, fournir ce genre de services aux visiteurs/utilisateurs d'un site portail tel que celui des AAR, présuppose de passer d'une conception de type "portail d'information/de connaissances" à une conception de type "collaboratoire" ou encore "espace numérique de travail".

Un collaboratoire ou un espace numérique de travail réunit donc, très généralement parlant, d'une part des ressources numériques d'information (sous forme de documents écrits, audiovisuels, ... organisés en des bibliothèques électroniques) et d'autre part des services et outils/instruments nécessaires pour effectuer un travail tel que celui présenté ci-dessus.

Références:

- l'ouvrage collectif Portails et collaboratoires pour l'enseignement et la recherche (coordonné par P. Stockinger); Paris Hermes Science Publications 2003.

7.3/ Vers un service « d'intérêt général » (« community service ») des AAR

Dans le cadre concret du Programme AAR, cela veut dire qu'un visiteur/ utilisateur enregistré devra avoir la possibilité:

- d'obtenir son propre espace personnel de travail (i.e. sa "table de travail");
- de se constituer son propre corpus audiovisuel (et, plus généralement, multimédia) de travail;
- de créer son propre cadre de classement (son propre "thesaurus", voire sa propre "ontologie");
- de réaliser à distance (!) l'ensemble des activités citées ci-dessus (découpage/segmentation des vidéos, annotations, enrichissement, traduction, ...);
- de procéder à distance (!) à la publication personnalisée des résultats de ses activités;
- de publier les résultats de ses activités sur un site personnel accessible uniquement à ceux identifiés par le travailleur.

En comparant ces exigences avec le processus de republication d'un corpus audiovisuel afin de permettre une meilleure appropriation des connaissances contenues dans le corpus, alors on peut aisément se rendre compte qu'il s'agit ici essentiellement de passer d'un modèle "centralisé" d'exploitation d'un corpus source à un modèle totalement décentralisé:

- dans le cadre du modèle "centralisé, c'est une institution, un service qui - comme traditionnellement l'éditeur ou le producteur/metteur en scène - règle pour le lecteur/le destinataire visé le processus et les modalités de l'appropriation de l'information;
- dans le cadre du modèle "décentralisé", c'est la communauté concernée (dans notre cas: les enseignants, les étudiants, les chercheurs, ...) qui prennent en main leurs besoins et intérêts d'appropriation sous forme d'un travail adapté sur et autour du texte.

C'est ainsi qu'un portail tel que celui des AAR peut se transformer en un service d'intérêt général qui non seulement fournit des informations ou des connaissances à une communauté donnée mais permet également à celle-ci de l'utiliser à ses propres fins.

Références:

- le projet OPALES (programme RIAM: 2001 - 2004), coordonné par INA Recherche, ayant eu comme but à définir un outil de travail personnalisé sur des corpus audiovisuels;
- la conférence de P. Stockinger "Digital audiovisual archives in humanities" (Chania, Grèce 2003) expliquant les objectifs du projet OPALES.

VIII/ Le cadre théorique

8.1/ La sémiotique du texte

Comme déjà signalé dans la présentation générale du cours, les activités de R&D autour de la publication/republication de corpus audiovisuels (et, plus généralement, multimédias) et plus spécifiquement autour de l'hypertextualité, font référence à un cadre théorique et méthodologique constitué par l'analyse du discours et la sémiotique du texte.

Le texte est le support nécessaire à toute communication pour véhiculer, faire circuler une information entre les membres d'une communauté de communication.

Les approches en sciences du langage qui s'occupent de l'étude du texte l'étudient d'une manière analogue à la linguistique (structurale) des langues comme un objet qui

- possède une organisation autonome interne
- s'inscrit dans et se positionne par rapport à un environnement (culturel, social, technique et naturel).

Il existe toute une diversité de traditions et courants (linguistiques, rhétoriques, sémiotiques, cognitifs, ...) qui s'occupe du phénomène "texte" dans sa globalité ou en en privilégiant tel ou tel aspect particulier.

Références:
- analyse (sémio-linguistique) du discours (P. Charaudeau; D. Maingueneau)
- la sémiotique de la langue (Michael Halliday);
- la théorie de la structure rhétorique (William Mann, Sandra Thompson, ...)
- la "critical discourse analysis" (Norman Fairclough, ...);

8.2/ Le texte comme entité structurale

Le texte entendu au sens d'une entité structurale est supposé de posséder une organisation interne spécifique qu'il partage avec un ensemble d'autres textes faisant partie du même type ou - comme on dit - genre.

Afin de pouvoir expliciter et décrire plus systématiquement l'organisation interne d'un texte (d'un corpus de textes), on fait appel à un ensemble de plans constitutifs dont les suivants:

- plan thématique du texte: sélection et thématisation des informations traitées et développées dans le texte (le corpus de textes);
- plan narratif lato sensu: intégration syntagmatique des informations sélectionnées et traitées dans le texte (le corpus de texte);
- plan rhétorique et discursif: choix d'un acte ou genre de discours pour parler d'une information sélectionnée et thématisée dans le texte (le corpus textuel);
- plan de l'expression: choix d'un langage et de medium d'expression dans le texte (le corpus textuel);
- plan de l'organisation formelle et physique de l'information traitée dans le texte (le corpus textuel);
- plan du support physique et technique de la "matérialité" (R. Chartier) du texte (du corpus textuel).

Chaque plan constitue l'objet d'importants travaux à la fois théoriques et empiriques impossibles à résumer ici en quelques lignes.

Ceci dit, une connaissance opérationnelle des différents plans est indispensable pour les différents secteurs des technologies de l'information et de la connaissance.

Références:

- livre de Peter Stockinger, Le document audiovisuel; Paris, Hermes Science Publications 2003

8.3/ L'environnement pertinent du texte

D'une manière analogue à l'acteur social, un texte possède son "monde vie", c'est-à-dire qu'il est:

- participe à des activités et pratiques sociales;
- interagit et communique avec d'autres textes en formant un "textscape", une "communauté de textes" plus ou moins typique, plus ou moins stable ou, au contraire, éphémère;
- se "nourrit" et nourrit d'autres textes en constituant ainsi un élément dans l'évolution textuelle et culturelle.

Description et Modélisation des relations intertextuelles sont à la fois théoriquement et empiriquement nécessaires :

- pour comprendre le sens des rapports hypertextuels dont est pourvu un document numérique
- et pour un projet de publication/republication d'un document numérique (audiovisuel).