Périodique électronique étudiant

‹cursus›

de l'École de bibliothéconomie
et des sciences de l'information
de l'Université de Montréal.

Volume 4, numéro 1 (automne 1998)

ISSN 1201-7302

Comité de rédaction

Comité de lecture

Dans ce numéro

Page d'accueil de Cursus

Page d'accueil de l'EBSI

Dernière mise à jour : mars 1998

Cursus vol. 4 no 1 (automne 1998) - Comité de rédaction

Page d'accueil de ce numéro

Cursus vol. 4 no 1 (automne 1998) - Comité de lecture

Page d'accueil de **ce numéro**

Périodique électronique étudiant

<cursus>

de l'École de bibliothéconomie
et des sciences de l'information
de l'Université de Montréal.

Volume 4, numéro 1 automne 1998)

ISSN 1201-7302

Comité de rédaction

Comité de lecture

Dans ce numéro

Tout en suivant de près le projet du Resource Description Framework (RDF), l'objet de notre recherche consiste, à partir d'un ensemble restreint de documents photographiques, à intégrer aux 15 éléments de description prévus dans le Dublin Core les informations prescrites par les Règles pour la description des documents d'archives (RDDA). Pour ce faire, nous avons suivi la démarche suivante. Après avoir intégré les données des RDDA au Dublin Core à l'aide du Nordic Metadata Template, nous avons mis en place un site web temporaire. Ce site a été indexé puis interrogé à l'aide de moteurs de recherche. Les objectifs de ce projet sont de deux ordres. D'une part, il s'agit de vérifier les avantages que peuvent représenter les métadonnées dans la gestion des archives photographiques. En effet, dans la mesure où de plus en plus d'archives photographiques sont disponibles sur le web, les métadonnées ne deviendraient-elles pas un outil idéal de gestion de ces archives? Cet outil permettrait autant de décrire et de diffuser les

documents d'archives que de les repérer ou de les gérer. D'autre part, il s'agit de mettre en évidence l'importance que prennent les normes de description des documents d'archives dans l'élaboration des métadonnées pour les documents HTML.

Véronique Parenteau, "L'analyse de textes littéraires assistée par ordinateur: une introduction"

Les analyses littéraires sont généralement qualitatives. Des outils informatiques jumelés à des méthodes statistiques peuvent toutefois permettre de faire des analyses quantitatives des textes littéraires. L'étude de données "brutes" comme les mots, les syllabes et la ponctuation sert de base à des comparaisons qui peuvent notamment aider à identifier la paternité d'un texte, à distinguer les imitations des oeuvres authentiques, à comprendre les motifs rythmiques dans des vers et à saisir comment un auteur a contribué à l'évolution du langage. Tous ces types d'analyses nécessitent cependant une intervention humaine. L'analyse de textes littéraires assistée par ordinateur a ses limites et bien des experts en littérature ne croient pas à son utilité.

Kumiko Vézina, "Survol du monde de l'indexation de l'image"

Le domaine de l'indexation des images est encore relativement jeune si on le compare à celui de l'indexation textuelle mais il est en pleine effervescence. À partir de la fin des années 70, de nombreuses institutions voulurent mettre en valeur leurs collections visuelles. Ils mirent donc sur pied des projets de thésaurus et de classification de grande envergure destinés à mieux organiser leurs images. Toutefois, ces projets soulevèrent de nombreux problèmes vis-à-vis l'indexation du matériel visuel et moussèrent l'intérêt académique dans le domaine de la bibliothéconomie et des sciences de l'information. Ce survol présente donc les premiers grands projets d'organisation d'images, les systèmes actuels d'indexation des images utilisant de nouvelles techniques comme thésaurus visuels et langages de description pictorielle, les problèmes reliés à ce domaine et les recherches à faire.

Page d'accueil de Cursus
Page d'accueil de l'EBSI

Dernière mise à jour : mars 1998

LES MÉTADONNÉES COMME OUTIL DE GESTION DES ARCHIVES PHOTOGRAPHIQUES

par

Yvon Lemay

Cursus vol. 4 no 1 (automne 1998)

Cursus est le périodique électronique étudiant de l'École de bibliothéconomie et des sciences de l'information (EBSI) de l'Université de Montréal. Ce périodique diffuse des textes produits dans le cadre des cours de l'EBSI.

ISSN 1201-7302

C. élec. : cursus@ere.umontreal.ca
URL : http://www.fas.umontreal.ca/ebsi/cursus/

Droits d'auteur

L'auteur

Yvon Lemay a terminé sa maîtrise en bibliothéconomie et en science de l'information (option archivistique), à l'École de bibliohéconomie et des sciences de l'information (EBSI) en avril 1998. Il est actuellement cinémathécaire aux Services documentaires Information de la Société Radio-Canada. Il fait partie d'une équipe qui travaille sur les films de nouvelles dans le cadre du projet des Archives des réseaux français.

Ce texte a été réalisé à l'hiver 1998 dans le cours *Recherche individuelle* (BLT-6850) sous la direction de James Turner. Nous aimerions remercier le professeur Turner pour l'aide qu'il nous a accordée ainsi que la direction de Parcs Canada, canal de Lachine pour nous avoir permis d'utiliser le Relevé photographique du canal de Lachine et de reproduire l'une des photographies dans notre texte. Nous aimerions également remercier les responsables du laboratoire d'informatique de l'EBSI, en particulier Rabah Djanati, pour leur précieux soutien technique.

Le texte suivant a été produit dans le cadre du cours BLT 6850, Recherche individuelle, sous la direction de M. James Turner.

Table des matières

INTRODUCTION

Tout en suivant de près le projet du Resource Description Framework (RDF) qui est présentement mené par le World Wide Web Consortium, l'objet de notre recherche consiste, à partir d'un ensemble restreint de documents photographiques, à intégrer aux 15 éléments de description prévus dans le Dublin Core (soit la zone <meta des documents HTML) les informations prescrites par les *Règles pour la description des documents d'archives* (RDDA).
Les objectifs que nous poursuivons, en effectuant ce projet pilote, sont de deux ordres. D'une part, il s'agit de vérifier les avantages que peuvent représenter les métadonnées dans la gestion des archives photographiques. En effet, dans la mesure où de plus en plus d'archives photographiques sont disponibles sur le web, les métadonnées ne deviendraient-elles pas un outil idéal de gestion de ces archives? Cet outil permettrait autant de décrire et de diffuser les documents d'archives que de les repérer ou de les gérer. D'autre part, il s'agit de mettre en évidence l'importance que prennent les normes de description des documents d'archives dans l'élaboration des métadonnées pour les documents HTML.
Pour mener à bien ce projet, nous avons suivi la démarche suivante. Après avoir intégré les données des RDDA au Dublin Core à l'aide du Dublin Core Metadata Template, nous avons mis en place un site web temporaire. Ce site a été indexé puis interrogé à l'aide de moteurs de recherche afin de vérifier les avantages ainsi que les inconvénients des métadonnées dans la gestion des archives photographiques.
Évidemment, la portée d'un tel projet pilote est forcément limitée. Il faudrait, entre autres, examiner plus en détail la manière d'intégrer les RDDA au éléments du Dublin Core. Toutefois, nous osons croire que, malgré ses limites, ce projet aura permis de montrer comment les archivistes peuvent et doivent aujourd'hui exercer leur rôle. D'ailleurs, comme le soulignait David Bearman, la discipline de l'archivistique a toujours été de caractère méta-informationnelle puisque, par définition, elle consiste à recueillir de l'information sur les systèmes d'information. Raison de plus de continuer à le faire.

LES MÉTADONNÉES ET L'ARCHIVISTIQUE

À partir de la fin des années 1980, et notamment avec la conférence sur les recherches en matière de documents électroniques organisée par la National Historical and Records Commission (NHRC) en 1991, la question des métadonnées est devenue un sujet de préoccupation de plus en plus important dans le domaine des archives. Et pour cause.
Face à l'environnement électronique, il est dorénavant nécessaire de faire une distinction entre la structure imposée par le logiciel et l'information contenue dans un document. Cette dernière constitue les données alors que les règles de structuration et de présentation des données, telles qu'imposées et déterminées au sein des logiciels constituent les métadonnées, c'est-à-dire l'information au sujet de l'information.
Ainsi, au plan archivistique, les métadonnées représentent de nombreux avantages. Comme le fait remarquer David Wallace, elles permettent: a) d'identifier et de préserver le contexte du document; b) de préserver la structure du document et du logiciel; c) de produire et de conserver toutes les informations relatives à l'évaluation et à l'arrangement des documents; d) d'effectuer la gestion du cycle de vie des documents; e) de préserver et/ou d'assurer la migration de la fonctionnalité des systèmes; f) de créer des inventaires des ressources disponibles dans les milieux (Wallace, 1993, 88). Bref, les métadonnées donnent la possibilité aux archivistes de jouer pleinement leur rôle et d'assumer la gestion des documents électroniques.
Il existe actuellement une très grande variété de métadonnées. Il suffit de consulter le site "Digital Libraries: Metadata

Resources" que l'International Federation of Library Associations and Institutions (IFLA, 1998) consacre à la question des données sur les données pour s'en convaincre. Aussi la question qui se pose est: faut-il laisser laisser libre cours à l'initiative de tout un chacun en la matière ou, au contraire, faut-il chercher à développer et à imposer des normes qui seraient communes à tous? Il semble que, de plus en plus, les divers intervenants du domaine de l'information préoccupés par cette question soient convaincus du bien-fondé d'opter pour la deuxième option, notamment en ce qui concerne les documents HTML que l'on retrouve sur le Web.

LES MÉTADONNÉES ET LE DUBLIN CORE

"The World Wide Web has changed both the creation, distribution, storage and retrieval and presentation of information. The amount of information and networked resources produced are increasingly growing on the WWW. This means that there will be difficulties for the end-user to search, browse and navigate to the relevant information" (Preben, 1998)

Entre mars 1995 et mars 1997 a eu lieu une série de quatre ateliers réunissant des experts internationaux de différentes disciplines dans le but de développer "*a core element set that provides adequate data for Web resource discovery and is simple for authors and content managers to create and maintain*" (Weibel et Lagoze, 1997, 176).

Le résultat de leurs travaux, connu sous le nom de "Dublin Core Metadata", soit le nom de la ville (Dublin, Ohio) où s'est tenu la première rencontre, comprend 15 éléments de description. Il est à noter que, suite au troisième atelier ("Image Metadata Workshop"), il a été convenu de développer un modèle permettant de tenir compte des ressources tant textuelles que visuelles.

Voici la liste des 15 éléments de description du Dublin Core ainsi qu'un aperçu du leur contenu. Ces informations proviennent du document "Dublin Core Metadata Element Set: Reference Description" disponible sur le Web (http://purl.oclc.org/metadata/dublin_core/) :

1. Title
   Label: title
   The name given to the resource by the creator or publisher.
2. Author or Creator
   Label: creator
   The person or organization primarily responsible for creating the intellectual content of the resource. For example, authors in the case of written documents, artists, photographers, or illustrators in the case of visual resources.
3. Subject and Keywords
   Label: subject
   The topic of the resource. Typically, subject will be expressed as keywords or phrases that describe the subject or content of the resource. The use of controlled vocabularies and formal classification schemas is encouraged.
4. Description
   Label: description
   A textual description of the content of the resource, including abstracts in the case of document-like objects or content descriptions in the case of visual resources.
5. Publisher
   Label: publisher
   The entity responsible for making the resource available in its present form such as a publishing house, a university department, or a corporate entity.
6. Other Contributor
   Label: contributor
   A person or organization not specified in a creator element who has made significant intellectual contributions to the resource but whose contribution is secondary to any person or organization specified in a creator element (for example, editor, transcriber, and illustrator).
7. Date
   Label: date
   The date the resource was made available in its present form. Recommended best practice is an 8 digit number in the

form YYYY-MM-DD as defined in http://www.w3.org/TR/NOTE-datetime, a profile of ISO 8601. In this scheme, the date element 1994-11-05 corresponds to November 5, 1994. Many other schema are possible, but if used, they should be identified in an unambiguous manner.

8. Resource Type
   Label: type
   The category of the resource, such as home page, novel, poem, working paper, technical report, essay, dictionary. For the sake of interoperability, type should be selected from an enumerated list that is under development in the workshop series at the time of publication of this document. See http://sunsite.berkeley.edu/Metadata/types.html for current thinking on the application of this element
9. Format
   Label: format
   The data format of the resource, used to identify the software and possibly hardware that might be needed to display or operate the resource. For the sake of interoperability, format should be selected from an enumerated list that is under development in the workshop series at the time of publication of this document.
10. Resource Identifier
    Label: identifier
    String or number used to uniquely identify the resource. Examples for networked resources include URLs and URNs (when implemented). Other globally-unique identifiers,such as International Standard Book Numbers (ISBN) or other formal names would also be candidates for this element in the case of off-line resources.
11. Source
    Label: source
    A string or number used to uniquely identify the work from which this resource was derived, if applicable. For example, a PDF version of a novel might have a source element containing an ISBN number for the physical book from which the PDF version was derived.
12. Language
    Label: language
    Language(s) of the intellectual content of the resource. Where practical, the content of this field should coincide with RFC 1766. See: http://ds.internic.net/rfc/rfc1766.txt
13. Relation
    Label: relation
    The relationship of this resource to other resources. The intent of this element is to provide a means to express relationships among resources that have formal relationships to others, but exist as discrete resources themselves. For example, images in a document, chapters in a book, or items in a collection. Formal specification of relation is currently under development. Users and developers should understand that use of this element is currently considered to be experimental.
14. Coverage
    Label: coverage
    The spatial and/or temporal characteristics of the resource. Formal specification of coverage is currently under development. Users and developers should understand that use of this element is currently considered to be experimental.
15. Rights Management
    Label: rights
    A link to a copyright notice, to a rights-management statement, or to a service that would provide information about terms of access to the resource. Formal specification of rights is currently under development. Users and developers should understand that use of this element is currently considered to be experimental.

Nous aurons l'occasion d'apporter des précisions quant au contenu de ces éléments lorsque nous aborderons la façon d'incorporer les métadonnées du Dublin Core au document HTML. Pour l'instant, nous aimerions souligner le fait que, mis à part certains éléments qui sont reliés à des listes en développement, le contenu de la plupart des 15 éléments de description du Dublin Core n'est pas à proprement parler contrôlé. Au mieux, suggère-t-on le recours à certains outils existant pour ce faire. D'où par conséquent l'intérêt d'utiliser les normes qui ont été développées au fil des ans afin de décrire les documents d'archives.

LES RÈGLES POUR LA DESCRIPTION DES DOCUMENTS D'ARCHIVES (RDDA)

Publié en 1990 par le Bureau canadien des archives, les *Règles pour la description des documents d'archives* (RDDA), comportent trois principaux volets: a) les zones de description, b) les niveaux de description et c) les vedettes.

Les zones de description

Au nombre de neuf, les zones de description sont les suivantes:

1. Zone du titre et de la mention de responsabilité
2. Zone de l'édition
3. Zone des précisions relatives au support
4. Zone des dates de création, de diffusion, de publication, etc.
5. Zone de la collation
6. Zone de la collection
7. Zone de la description des documents d'archives
8. Zone des notes
9. Zone du numéro normalisé et des modalités d'acquisition

Il est à remarquer que le contenu de chacune de ces zones varie en fonction de la catégorie de documents à décrire. Par exemple, le chapitre 4 sur les documents iconographiques donne des indications concernant la description "des documents d'archives prenant la forme de tableaux, de dessins, de gravures, de photographies ou d'images produites par divers procédés d'illustration visuelle" (Règles, 4.0A1).

Les niveaux de description

Selon les *Règles pour la description des documents d'archives*, "*la description d'un fonds d'archives consiste en des descriptions qui représentent le fonds comme un ensemble organique et dynamique. Ce dernier est composé de séries qui peuvent elles-mêmes être constituées de dossiers qui, à leur tour, contiennent des pièces. Chacune de ces parties fait (ou peut faire) l'objet d'une description. Il en résulte plusieurs descriptions qui doivent être reliées hiérarchiquement entre elles afin de rendre compte de l'organisation du fonds*" (Règle 1.0A1).
Cette relation hiérarchique est particulièrement évidente dans la zone 7, la zone de la description des archives. Au niveau du fonds ou de la série, la description comprendra soit une histoire administrative concise de la personne morale, soit une notice biographique concise de la personne physique qui est responsable de la création de l'unité archivistique en question, un historique de la conservation ainsi que des informations sur les fonctions ou activités qui ont été à l'origine de la création soit du fonds, soit de la série (Portée et contenu).
Au niveau du dossier ou de la pièce, les règles demandent de faire un historique de la conservation correspondant à l'unité archivisitique à décrire (lorsque justifié) et de donner, dans la section "Portée et contenu", les informations sur les fonctions et les activités qui ont été à l'origine de la création du dossier ou de la pièce.

Les vedettes

Enfin, les *Règles pour la description des documents d'archives* prévoient trois types d'accès : accès de provenance, accès d'auteur et les autres accès indépendants du sujet. Par accès de provenance, il s'agit "*des accès aux noms du créateur du fonds et de ses séries si les noms du créateur de la série diffèrent de ceux du créateur du fonds*" (Règle 21.1). Par accès d'auteur, il s'agit au niveau de la série, du dossier et de la pièce "*des accès aux noms des auteurs identifiés dans la zone du titre et de la mention de responsabilité*" (Règle 21.5). Pour les autres accès indépendants du sujet, mentionnons seulement l'accès aux noms apparaissant dans la zone du titre et de la mention de responsabilité (Règle, 21.9).

Pour bien comprendre comment ces règles s'appliquent, prenons à titre d'exemple les trois niveaux de description que nous avons utilisés pour rendre compte du *Relevé photographique du canal de Lachine*, le document qui nous servira à mener ce projet pilote.

Exemple 1

Il s'agit de la description du *Relevé photographique du canal de Lachine* dans son ensemble. Il est à noter que, dans cette description, nous avons considéré le relevé comme étant une série d'un fonds d'archives. La présentation correspond à la forme la plus courante. Dans ce type de présentation, la première partie contient les informations des zones 1 à 6, la deuxième partie, les informations de la zone 7 et et la troisième partie celles des zones 8 et 9, là ou les vedettes n'apparaissent pas dans ce type de présentation.

RELEVÉ PHOTOGRAPHIQUE / Parcs Canada, Canal de Lachine. - 1997. - 523 photographies: coul.; 10x15 cm + plans.

HISTOIRE ADMINISTRATIVE:
Fermé à la navigation en 1970, le canal de Lachine a été transféré à Parcs Canada par le ministère des Travaux publics en 1978. Depuis cette date, il fait partie du réseau des sites historiques canadiens. D'une longueur de quatorze kilomètres, le canal de Lachine qui relie le Vieux-Montréal au lac Saint-Louis est un élément patrimonial majeur de la région montréalaise. Les cyclistes, piétons et skieurs qui empruntent la piste cyclable longeant le canal peuvent y admirer des bâtiments et des écluses qui ont été témoins de toute l'évolution industrielle de Montréal.

En prévision d'importants travaux de restauration devant débuter au printemps 1998, travaux qui mèneront à la réouverture du canal à la navigation de plaisance au tournant du XXIe siècle, le canal de Lachine a fait réaliser au cours de l'été 1997 un relevé photographique de ses installations ainsi que des bâtiments avoisinant. L'objectif de ce relevé était double : conserver un témoignage de l'état actuel du canal et constituer un outil de référence pour la réalisation des travaux.

PORTÉE ET CONTENU:
Les 523 photographies du relevé sont réunies dans trois cahiers selon six secteurs qui, à l'exception des secteurs 1 et 3, sont subdivisés en trois sections:

- **Secteur 1 (Lachine)**
- **Secteur 2 (Rockfield) : rive nord, piste cyclable - rive sud**

- **Secteur 3 (Ville Saint-Pierre/Lasalle) :**
    - **3.1 (jusqu'à la passerelle des Trinitaires) : rive nord - piste cyclable - rive sud**
    - **3.2 (jusqu'à Ville Lasalle et le quartier Saint-Paul) : rive nord - piste cyclable - rive sud**
    - **3.3 (jusqu'au pont de Côte Saint-Paul) : rive nord - piste cyclable - rive sud**
- **Secteur 4 (Saint-Henri/Côte Saint-Paul) : rive nord - piste cyclable - rive sud**
- **Secteur 5 (Pointe Saint-Charles/Petite-Bourgogne) : rive nord -piste cyclable - rive sud**
- **Secteur 6 (Vieux-Port de Montréal) : rive nord - piste cyclable - rive sud.**

Chaque photographie est accompagnée d'une légende comprenant le nom (du lieu représenté), la localisation (c'est-à-dire l'endroit où a été prise la photographie) et, dans plusieurs cas, un numéro de référence qui correspond aux fiches techniques du document *Inventaire et évaluation des ressources culturelles canal de Lachine* produit par la firme Archémi en 1995. Un plan d'ensemble ainsi qu'un plan de chacun des six secteurs provenant du document de consultation publique *L'avenir du lieu historique national du Canal-de-Lachine* (1996) ont été incorporés aux cahiers.

NOTES:
Le titre est basé sur le contenu du document.
Les négatifs sont placés à la fin du troisième cahier.

Exemple 2

Le deuxième exemple nous situe au niveau suivant de description. Ce niveau correspond à un dossier plutôt qu'à une sous-série comme cela devrait être normalement le cas. Nous avons choisi d'éliminer ce niveau de description afin de permettre à l'utilisateur d'avoir accès plus rapidement aux photographies en format réduit qui sont situées au bas de la description.

SECTEUR 5 (POINTE SAINT-CHARLES/PETITE-BOURGOGNE) : PISTE

CYCLABLE / Parcs Canada, Canal de Lachine. - 1997. - 52 photographies : coul.; 10 x 15 cm.

PORTÉE ET CONTENU:
Les 52 photographies de cette section du secteur 5 couvre la piste cyclable du canal de Lachine de la passerelle Atwater (à l'ouest) jusqu'au pont Wellington (à l'est).

NOTES:
Titre basé sur le contenu du document.
Les photographies sont accompagnées d'une légende.
Toute reproduction totale ou partielle des photographies est interdite sans le consentement préalable de Parcs Canada (lachine-mtl@pch.gc.ca).

PHOTOGRAPHIES:

- **NO 29 Pont des Seigneurs - Rue des Seigneurs, au-dessus de l'écluse 3 - 5OC05C**
- **NO 30 Écluse 3 (sud) - À l'est du pont des Seigneurs - 5OC02N**
- **NO 43 Piste cyclable - Lampadaire C25, rive sud, vue est**
- **NO 47 Canal de fuite (nord) - Au nord du canal actuel, à l'angle des rues des Seigneurs et Basin - 5VI03N**

Exemple 3

Le dernier niveau correspond à la description d'une pièce. Dans ce cas, la description est précédée de l'image photographique en grand format (comparativement au format réduit du niveau précédent).

PONT DES SEIGNEURS / Chantal Lavallée. - 1997. - 1 photographie : coul. ; 10 x 15 cm + légende.

PORTÉE ET CONTENU:
Cette photographie est la 29e du secteur 5 (Pointe Saint-Charles/Petite-Bourgogne) : piste cyclable du *Relevé photographique du canal de Lachine*.

NOTES:
Titre officiel propre.
Légende : Nom : Pont des Seigneurs, Localisation : Rue des Seigneurs, au-dessus de l'écluse 3, Numéro de référence: 5OC05C
Le numéro de référence correspond à une fiche technique du document *Inventaire et évaluation des ressources culturelles canal de Lachine* produit par la firme Archémi en 1995.
Ressources culturelles no 103 sur le plan du document de consultation publique *L'avenir du lieu historique national du Canal-de-Lachine* (1996).

LES RDDA et le Dublin Core

Sachant en quoi consistent les *Règles pour la description des documents d'archives* (RDDA) et comment elles s'appliquent, il reste maintenant à savoir de quelle façon elles peuvent s'intégrer aux 15 éléments de description du Dublin Core. Pour ce faire, dressons d'abord, à l'aide d'un tableau, un parallèle entre les 15 éléments du Dublin Core et les neuf zones des RDDA, puis, à partir d'un exemple, voyons pratiquement comment cette intégration peut s'effectuer .

Tableau

1. **Title**	• **zone 1**
2. **Author or Creator**	• **zone 1**
3. **Subject and Keywords**	• **accès (vedettes)**
4. **Description**	• **zone 7; zone 5 ; zone 3**
5. **Publisher (du document électronique)**	• **zone 2**
6. **Other Contributor**	• **zone 8**
7. **Date (de création du document électronique)**	• **xxxxxxxxxxxxxxxxx**
8. **Resource Type (liste en développement)**	• **xxxxxxxxxxxxxxxxx**
9. **Format (à partir d'une liste en développement)**	• **xxxxxxxxxxxxxxxxx**
10. **Resource Identifier (URL, ISBN)**	• **zone 9**
11. **Source (du document)**	• **zone 8; zone 4; zone 2**
12. **Language**	• **zone 8 (notes)**
13. **Relation (dans ou à d'autres documents)**	• **relations hiérarchiques; zone 8; zone 6**
14. **Coverage (liste en développement)**	• **xxxxxxxxxxxxxxxxx**
15. **Rights management**	• **zone 8**

Comme l'indique ce tableau, mis à part ceux dont le contenu est déterminé soit par des listes en développement, soit par une norme ISO, les éléments de description prévus par les *Règles pour la description des documents d'archives* sont en mesure de s'intégrer aux 15 éléments du Dublin Core. Et l'on pourrait même ajouter que leur caractère normalisé en font des outils tout désigné pour cette tâche. À ce propos, dans le document de présentation des 15 éléments du Dublin Core, l'on tenait à signaler que "to promote global interoperability, a number of the element descriptions may be associated with a controlled vocabularies for the respective element values" (Dublin Core, 1997).

Bien sûr, ce n'est là qu'une première ébauche de correspondance qui devra être davantage élaborée. Néanmoins, malgré le caractère encore exploratoire de ce parallèle entre les éléments prescrits par les RDDA et ceux du Dublin Core, le degré de concordance que l'on constate à ce stade-ci est suffisamment élevé pour nous permettre de croire qu'il est légitime de procéder à un tel essai d'intégration.

D'ailleurs, il est à remarquer que cette mise en parallèle entre le Dublin Core et les *Règles pour la description des documents d'archives* indique que ces dernières devraient, dans l'avenir, tenir compte des éléments 7, 8, 9, 10 et 15 du Dublin Core dans leur politique de description. Espérons que ce sera le cas, notamment dans le chapitre sur les documents informatiques qui doit venir s'ajouter sous peu à ceux déjà disponibles.

Mais si le contenu des RDDA peut manifestement s'intégrer aux éléments du Dublin Core, quelle forme cela prendra-t-il concrètement? Pour le savoir, reprenons le dernier exemple concernant la description d'une pièce du *Relevé photographique du canal de Lachine*. À noter qu'en ce qui concerne l'élément 3 "Subject and Keywords", nous avons ajouté, comme le suggère le guide sur *La Gestion des documents photographiques au gouvernement du Canada* réalisé par les Archives nationales du Canada, les descripteurs du *Library of Congress Thesaurus for Graphic Materials* (LCTGM) aux vedettes prévues par les RDDA. Pour ce qui est des éléments 5, 6 et 10, nous avons pris en considération le contexte actuel du projet pilote.

Exemple

1. Pont des Seigneurs
2. Lavallée, Chantal
3. RDDA= Parcs Canada, Canal de Lachine; Lavallée, Chantal
   LCTGM= Canals--Canada--Montréal--1997; Bridges--Canada--Montréal--1997
4. Photographie : coul. ; 10 x 15 cm + légende. PORTÉE ET CONTENU: Cette photographie est la 29e du secteur 5 (Pointe Saint-Charles/Petite-Bourgogne) : piste cyclable du *Relevé photographique du canal de Lachine*. NOTES: Titre officiel propre. Légende : Nom : Pont des Seigneurs, Localisation : Rue des Seigneurs, au-dessus de l'écluse 3, Numéro de référence: 5OC05C .
5. École de bibliothéconomie et des sciences de l'information, Université de Montréal
6. Lemay, Yvon
7. 1998-02-19
8. Relevé photographique
9. Image numérisée (35,2 Ko) en format JPEG
10. URL: http://tormade.ere.umontreal.ca/~lemayyv/projet/
11. Parcs Canada, Canal de Lachine. *Relevé photographique*. Montréal, 1997, 3 cahiers, 523 photographies
12. français
13. Le numéro de référence correspond à une fiche technique du document *Inventaire et évaluation des ressources culturelles canal de Lachine* produit par la firme Archémi en 1995. Ressources culturelles no 103 sur le plan du document de consultation publique *L'avenir du lieu historique national du Canal-de-Lachine* (1996).
14. Canada, Québec (Province), Montréal, 1997
15. Toute reproduction totale ou partielle des photographies est interdite sans le consentement préalable de Parcs Canada (lachine-mtl@pch.gc.ca).

MISE EN PLACE DES ÉLÉMENTS DU PROJET PILOTE

Générer et incorporer les métadonnées

Comment trouver des façons simples d'intégrer les métadonnées aux documents HTML "without requiring additional tags or changes to browser software, and without unnecessarily compromising current practices for robot collection of data"? (Weibel, 1998). **La solution sur laquelle les responsables du Dublin Core se sont entendus est la suivante : 1) encoder les métadonnées dans les étiquettes META des documents HTML à raison d'un élément par étiquette et utiliser autant d'étiquettes que nécessaire, 2) donner la référence URL des règles de description qui ont été utilisées pour décrire le document, en l'occurence celles du Dublin Core et 3) utiliser autant que possible des sources connues afin d'établir le contenu des éléments de description.**
Dans le but d'aider les utilisateurs à créer des métadonnées selon le modèle du Dublin Core, le Nordic Metadata Project a développé et rendu accessible sur le Web un gabarit qui en gère automatiquement la syntaxe. Il suffit de saisir le contenu de 15 éléments de description du Dublin Core dans les

espaces appropriés du Dublin Core Metadata Template et, une fois l'opération complétée, les métadonnées sont générées aussitôt en format HTML selon les normes prescrites. Tout ce qu'il reste à faire est d'en copier le résultat et de l'incorporer au document.

Exemple

Voici le traitement opéré par le Dublin Core Metadata Template du Nordic Metadata Project à partir des métadonnées de la 29e photographie de la section de la piste cyclable du Secteur 5 du *Relevé photographique du Canal de Lachine* (exemple précédent). Nous avons ajouté les balises de début et de fin de la zone HEAD pour en signaler l'emplacement exact dans le document HTML.

```
<HEAD>

<TITLE> Pont Des Seigneurs

<META NAME="DC.Date.X-MetadataLastModified"
CONTENT="(SCHEME=ISO8601) 1998-02-19"

<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/
dublin_core_elements#date"



<META NAME="DC.Title" CONTENT="Pont des Seigneurs"

<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/
dublin_core_elements#title"

<LINK REL=SCHEMA.rdda REFERENCE="ISBN: 0969079745"



<META NAME="DC.Creator.PersonalName" CONTENT="Lavall%E9e,
Chantal"
```

<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements#creator"

<LINK REL=SCHEMA.rdda REFERENCE="ISBN: 0969079745"

<META NAME="DC.Subject" CONTENT="Parcs Canada, Canal de Lachine"

<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements#subject"

<LINK REL=SCHEMA.rdda REFERENCE="ISBN: 0969079745"

<META NAME="DC.Subject" CONTENT="Lavall%E9e, Chantal"

<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements#subject"

<LINK REL=SCHEMA.rdda REFERENCE="ISBN: 0969079745"

<META NAME="DC.Subject" CONTENT="(SCHEME=TGM1) Canals--Canada--Montr%E9al--1997"

<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements#subject"

<META NAME="DC.Subject" CONTENT="(SCHEME=TGM1) Bridges--Canada--Montr%E9al--1997"

<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements#subject"

<META NAME="DC.Description" CONTENT="Photographie : coul. %3B 10 x 15 cm + l%E9gende.

PORT%C9E ET CONTENU: Cette photographie est la 29e du secteur 5

(Pointe Saint-Charles/Petite-Bourgogne): piste cyclable du Relev%E9 photographique

du canal de Lachine. NOTES: Titre officiel propre. L%E9gende : Nom : Pont des Seigneurs,

Localisation : Rue des Seigneurs, au-dessus de l'%E9cluse 3, Num%E9ro de r%E9f%E9rence :

5OC05C"

<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements#description"

<LINK REL=SCHEMA.rdda REFERENCE="ISBN: 0969079745"

<META NAME="DC.Publisher" CONTENT="%C9cole de biblioth%E9conomie et des sciences

de l'information, Universit%E9 de Montr%E9al"

<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements#publisher"

<LINK REL=SCHEMA.rdda REFERENCE="ISBN: 0969079745"

<META NAME="DC.Contributor.PersonalName" CONTENT="Lemay, Yvon"

<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements#contributor"

<LINK REL=SCHEMA.rdda REFERENCE="ISBN: 0969079745"

<META NAME="DC.Date" CONTENT="(SCHEME=ISO8601) 1998-02-19"

<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements#date"

<META NAME="DC.Type" CONTENT="Image.Photograph"

<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements#type"

<META NAME="DC.Format" CONTENT="(SCHEME=IMT) text/html"

<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements#format"

< REL=SCHEMA.imt HREF="http://sunsite.auc.dk/RFC/rfc/rfc2046.html">

<META NAME="DC.Format" CONTENT="(SCHEME=IMT) image/jpeg"

<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements#format"

<LINK REL=SCHEMA.imt HREF="http://sunsite.auc.dk/RFC/rfc/rfc2046.html"

<META NAME="DC.Identifier" CONTENT="http://tormade.ere.umontreal.ca/%7Elemayyv/projet/"

<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements#identifier"

<META NAME="DC.Source" CONTENT="Parcs Canada, Canal de Lachine. Relev%E9

photographique. Montr%E9al, 1997, 3 cahiers, 523 photographies"

<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements#source"

<LINK REL=SCHEMA.rdda REFERENCE="ISBN: 0969079745"

<META NAME="DC.Language" CONTENT="(SCHEME=ISO639-1) fr"

<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements#language"

<META NAME="DC.Relation" CONTENT="Arch%E9mi, Inventaire et %E9valuation des ressources

culturelles canal de Lachine, 1995"

<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/
dublin_core_elements#relation"

<LINK REL=SCHEMA.rdda REFERENCE="ISBN: 0969079745"

<META NAME="DC.Relation" CONTENT="Parcs Canada, L'avenir du
lieu historique national

du Canal-de-Lachine, 1996"

<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/
dublin_core_elements#relation"

<LINK REL=SCHEMA.rdda REFERENCE="ISBN: 0969079745"

<META NAME="DC.Coverage" CONTENT="Canada, Qu%E9bec
(Province), Montr%E9al, 1997"

<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/
dublin_core_elements#coverage"

<META NAME="DC.Rights" CONTENT="Toute reproduction est
interdite sans le consentement

pr%E9alable de Parcs Canada (lachine-mtl@pch.gc.ca)"

<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/
dublin_core_elements#rights"

<LINK REL=SCHEMA.rdda REFERENCE="ISBN: 0969079745"

</HEAD

Le Dublin Core Metadata Template s'ajuste aux besoins de l'utilisateur, c'est-à-dire que le gabarit est conçu de sorte qu'il est possible d'ajouter de nouvelles cases à celles qui sont données implicitement pour chacun des éléments. Règle générale, le nombre de caractères disponibles est suffisant bien que nous avons dû, dans le cas du 15e élément, ajuster le contenu à l'espace prévu par le gabarit. Mais, une fois les métadonnées générées, rien n'empêche l'utilisateur d'apporter les corrections qu'il juge nécessaire.
En plus de générer automatiquement les métadonnées en format HTML selon la convention établie par les membres du Dublin Core, le Dublin Core Metadata Template permet à l'utilisateur d'avoir accès au contenu des listes en développement (lorsqu'elles sont disponibles) ainsi qu'à diverses sources en ligne pour lui venir en aide dans l'indexation et la classification des documents. Comme l'indique les résultats générés par le gabarit, la référence est ensuite indiquée dans l'étiquette en question (ex.: TGM1 pour *LC Thesaurus for Graphic Materials 1 : Subject Terms*).
Il est à noter qu'en ce qui concerne les éléments 8 (Type) et 9 (Format), nous avons suivi les directives des listes en développement et que pour l'élément 14 (Coverage), comme ces directives n'étaient pas disponibles, nous avons dû opter pour une formule de notre choix. En plus de la référence URL au Dublin Core, nous avons ajouté, comme le suggère le Nordic Metadata Project, la référence aux RDDA (en l'occurence le numéro ISBN) chaque fois que cela nous semblait pertinent.

Mise en place, validation et inscription du site temporaire

Le site temporaire s'intitule "Rélevé photographique du canal de lachine: expérience sur les RDDA et le Dublin Core". Les 22 fichiers qu'il contient ont été transférés dans notre compte d'usager à l'adresse suivante: http://tornade.ere.umontreal.ca/~lemayyv/projet/. Le site comprend:

- **une page d'accueil précisant qu'il s'agit d'un site temporaire mis en place dans le but de mener une expérience sur les RDDA et le Dublin Core à partir du *Relevé photographique du canal de Lachine* (et métadonnées)**
- **une description du *Relevé photographique du canal de lachine* (et métadonnées)**

- **une description du secteur 5 : rive nord (et métadonnées)**
- **une description du secteur 5 : piste cyclable (et métadonnées)**
- **une description du secteur 5 : rive sud (et métadonnées)**
- **les descriptions de 11 photographies (et métadonnées)**
- **un plan détaillé du secteur 5 du canal de Lachine**
- **5 fiches d'Archémi correspondant aux numéros de référence.**

Une fois validé par le "HTML Validation Service Response" (http://valsvc.webtechs.com/) et les corrections nécessaires apportées, nous avons inscrit le site temporaire auprès de 9 moteurs ou répertoires de recherche en faisant appel au service gratuit offert par la firme Submit It (http://www.Submit-it.com). Ces moteurs ou répertoires sont :

- Alta Vista
- Excite
- AOL NetFind
- Magellan
- Infoseek
- HotBot
- Lycos
- Open Text
- Web Crawler

De plus, nous l'avons inscrit auprès du moteur de recherche HotMeta Search Engine. "*HotMeta is a metadata search engine. It fetches web-accessible documents from the Internet, indexes the metadata extracted from them and provides the user interface to search on this metadata*" (MetaWeb Project, 1998).
Après une semaine, et maintes vérifications auprès des différents moteurs ou répertoires de recherche, le site a été répertorié uniquement par Alta Vista, Infoseek et HotBot. En ce qui concerne HotMeta Search Engine, l'insuccès auprès de ce moteur de recherche semble s'expliquer par le fait que, pour l'instant, les seuls sites qu'il indexe sont des sites australiens. C'est dommage car ce moteur de recherche est parmi l'un des rares à être spécialement conçu en fonction des 15 éléments de description du Dublin Core. D'autres existent, comme le Nordic Web Index, mais ils permettent seulement d'interroger et non d'ajouter de nouveaux sites comprenant les éléments de description du Dublin Core.

LES MÉTADONNÉES À L'HEURE DE L'INTERROGATION

Sur le Web

Quelle est la politique des moteurs de recherche en ce qui a trait aux métadonnées? Prenons l'exemple d'Alta Vista, le chef de file incontesté en ce domaine. *"In the absence of any other information, Alta Vista will index all words in your document (except for comments), and will use the first few words of the document as a short abstract. It is however possible for you to control how your page is indexed by using the META tag to specify both additional keywords to index, and a short description. Let's suppose your page contains:*

- *<META name="description"*

content="We specialize in grooming pink poodles."

- *<META name="keywords" content="pet grooming, Palo Alto, dog"*

AltaVista will then do two things: It will index both fields as words, so a search on either poodles or dog will match. It will return the description with the URL. In other words, instead of showing the first couple of lines of the page, a match will look like the following: Pink Poodles Inc. We specialize in grooming pink poodles. http://pink.poodle.org/ - size 3k - 29 Feb 96. AltaVista will index the description and keywords up to a limit of 1,024 characters" (Alta Vista, 1998). De prime abord, une telle politique semble des plus favorables à l'interrogation de pages Web à l'aide des métadonnées issues des 15 éléments de description du Dublin Core. Or malheureusement, dans la pratique, il n'en est rien comme nous avons pu le constater lors de l'interrogation de notre site temporaire par différents moteurs de recherche. Qu'il s'agisse d'Alta Vista, Infoseek ou HotBot, les résultats ont été les mêmes. En mode "Advanced Search", chacun de ces moteurs répertorie le site de façon très précise. En effet, la requête "Relevé photographique du canal de Lachine" donne comme résultat : 1 document.

Infoseek found 1 page containing at least one of these words: "relevé photographique du canal de Lachine"
Relevé photographique du canal de Lachine: expérience sur les RDDA et ...
RELEVÉ PHOTOGRAPHIQUE DU CANAL DE LACHINE: EXPÉRIENCE SUR LES RDDA ET LE DUBLIN CORE. Ce site est temporaire. Il vise à mener une expérience dans le cadre d'un cours ...
98% http://tornade.ere.umontreal.ca/~lemayyv/projet/ (Size 4.2K) Document date: 19 Mar 1998

Hot Bot
look for all the words
Relevé photographique du canal de Lachine
Returned: 1 match.
Breakdown: relevé: 11654, photographique: 5109, du: 2406933, canal: 248877, de: 14479400, lachine: 6235
1.Relevé photographique du canal de Lachine: expérience sur les RDDA et le Dublin Core
100% RELEVÉ PHOTOGRAPHIQUE DU CANAL DE LACHINE: EXPÉRIENCE SUR LES RDDA ET LE DUBLIN CORE Ce site est temporaire. Il vise à mener une expérience dans le cadre d'un cours (BLT-6850: Recherche individuelle) du programme de maîtrise de l'École de...
http://tornade.ere.umontreal.ca/~lemayyv/projet/, 4277 bytes, 20Mar98

Compte tenu de la multitude de sites que l'on retrouve sur le Web, le résultat est certes fort appréciable. Mais, par ailleurs, aucun des trois moteurs de recherche ne permet une interrogation plus détaillée. Autant Alta Vista qu'Infoseek ou Hot Bot ne semblent aller plus loin que la zone du titre qui apparaît dans le haut de l'écran du logiciel de navigation. Autrement dit, ils ignorent complètement les éléments de description du Dublin Core.

Hot Bot
look for all the words
+META +name=DC.Publisher +Content=École de bibliothéconomie et des sciences de l'information
Sorry-- your search yielded no results.
You might want to revise your search query and try again...

Infoseek found no results for:
"META NAME=DC.Contributor Content=Lemay, Yvon"

Dans le but de vérifier cet état de fait et de tenter, si possible, de contourner cet obstacle, nous avons ajouté des mots-clés (keywords) à toutes les pages de notre site déjà munies des métadonnées du Dublin Core. Peine perdue! Cet ajout n'a fait aucune différence lors de l'interrogation.

Infoseek found no results for: "keywords=Parcs Canada, Canal de Lachine"

Hot Bot
look for all the words
keywords=Parcs Canada, Canal de Lachine
Sorry-- your search yielded no results.

Est-ce que ce résultat négatif est dû à notre façon d'inscrire les métadonnées, soit <META NAME=Keywords Content=XXXau lieu de <Keywords=XXX? Si oui, cela viendrait corroborer notre hypothèse.

En mode local

Ce constat à propos des métadonnées et des moteurs de recherche n'aurait pas été complet sans une tentative d'interrogation en mode local. Pour ce faire, nous avons téléchargé "AltaVista Personal 97", une version offerte gratuitement par la firme Alta Vista à partir de la page de recherche avancée du moteur de recherche. À la requête "Relevé photographique du canal de Lachine", ce n'est plus 1 mais bien 12 documents qui ont été répertoriés cette fois par Alta Vista, à savoir:

1. PS5N14.HTM
   CARTONS RECYCLÉS DE MONTRÉAL INC. (ENTREPÔTS BLACKSMITH SHOP) / Chantal Lavallée. - 1997. - 1 photographie : coul.; 10 x 15 cm + légende. PORTÉE ET CONTENU: Cette photographie est la 14e du sec
   a:\PS5N14.HTM - 98-04-05
2. PS5N15.HTM
   CARTONS RECYCLÉS DE MONTRÉAL INC. (ENTREPÔTS BLACKSMITH SHOP) / Chantal Lavallée. - 1997. - 1 photographie : coul.; 10 x 15 cm + légende. PORTÉE ET CONTENU: Cette photographie est la 15e du sec
   a:\PS5N15.HTM - 98-04-05
3. PS5N13.HTM
   CARTONS RECYCLÉS DE MONTRÉAL INC. / Chantal Lavallée. - 1997. - 1 photographie : coul.; 10 x 15 cm + légende. PORTÉE ET CONTENU: Cette photographie est la 13e du secteur 5 (Pointe Saint-Charles
   a:\PS5N13.HTM - 98-04-05
4. PS5S28.HTM
   REDPATH SUGAR REFINERY / Chantal Lavallée. - 1997. - 1 photographie : coul. ; 10 x 15 cm + légende. PORTÉE ET CONTENU: Cette photographie est la 28e du secteur 5 (Pointe Saint-Charles/Petite-Bo
   a:\PS5S28.HTM - 98-04-05
5. PS5S27.HTM
   REDPATH SUGAR REFINERY / Chantal Lavallée. - 1997. - 1 photographie : coul. ; 10 x 15 cm + légende. PORTÉE ET CONTENU: Cette photographie est la 27e du secteur 5 (Pointe Saint-Charles/Petite-Bo
   a:\PS5S27.HTM - 98-04-05
6. PS5S26.HTM
   REDPATH SUGAR REFINERY / Chantal Lavallée. - 1997. - 1 photographie : coul. ; 10 x 15 cm + légende. PORTÉE ET CONTENU: Cette photographie est la 26e du secteur 5 (Pointe Saint-Charles/Petite-Bo
7. PS5S29.HTM

REDPATH SUGAR REFINERY / Chantal Lavallée. - 1997. - 1 photographie : coul. ; 10 x 15 cm + légende. PORTÉE ET CONTENU: Cette photographie est la 29e du secteur 5 (Pointe Saint-Charles/Petite-Bo
a:\PS5S29.HTM - 98-04-05

8. PS5P25.HTM
PONT DES SEIGNEURS / Chantal Lavallée. - 1997. - 1 photographie : coul. ; 10 x 15 cm + légende. PORTÉE ET CONTENU: Cette photographie est la 29e du secteur 5 (Pointe Saint-Charles/Petite-Bourgo
a:\PS5P25.HTM - 98-04-05
9. PS5P32.HTM
ÉCLUSE 3 (NORD) / Chantal Lavallée. - 1997. - 1 photographie : coul.; 10 x 15 cm + légende. PORTÉE ET CONTENU: Cette photographie est la 32e du secteur 5 (Pointe Saint-Charles/Petite-Bourgogne)
a:\PS5P32.HTM - 98-04-05
10. PS5P47.HTM
CANAL DE FUITE (NORD) / Chantal Lavallée. - 1997. - 1 photographie : coul.; 10 x 15 cm + légende. PORTÉE ET CONTENU: Cette photographie est la 47e du secteur 5 (Pointe Saint-Charles/Petite-Bour
a:\PS5P47.HTM - 98-04-05
11. PS5P43.HTM
PISTE CYCLABLE / Chantal Lavallée. - 1997. - 1 photographie : coul.; 10 x 15 cm + légende. PORTÉE ET CONTENU: Cette photographie est la 43e du secteur 5 (Pointe Saint-Charles/Petite-Bourgogne)
a:\PS5P43.HTM - 98-04-05
>
12. INDEX.HTM
RELEVÉ PHOTOGRAPHIQUE DU CANAL DE LACHINE: EXPÉRIENCE SUR LES RDDA ET LE DUBLIN CORE Ce site est temporaire. Il vise à mener une expérience dans le cadre d'un cours (BLT-6850: Recherche individuelle
a:\INDEX.HTM - 98-04-05

Une différence pour le moins notable mais qui n'est pas due cependant à la présence de métadonnées. En effet, pour vérifier si, en mode local, Alta Vista tenait compte des métadonnées, nous avons effectué deux requêtes à partir du nom de la photographe ayant pris les images du *Relevé photographique du canal de Lachine*. Dans un cas, nous avons inscrit son nom tel qu'il apparaît dans le texte, soit Chantal Lavallée. Dans l'autre, nous l'avons inscrit tel qu'il apparaît dans les métadonnées, soit Lavallée, Chantal. Les résultats ne laissent aucun doute:

- "Chantal Lavallée"
About 11 documents match your query.
- "Lavallée, Chantal"
AltaVista Search cannot find any documents that match your query.
Enter a new query, try a different combination of words or phrases, or read the search examples in Help
- "Meta Name=DC.Creator.Personal Name Content=Lavallée,Chantal"
AltaVista Search cannot find any documents that match your query.
Enter a new query, try a different combination of words or phrases, or read the search examples in Help

Bien sûr, nous réalisons que les expériences que nous avons menées dans le cadre de ce projet pilote sont très limitées et qu'il serait nécessaire d'effectuer une recherche plus poussée sur les métadonnées et les moteurs de recherche afin de mieux identifier la ou les sources de problèmes. Néanmoins, malgré la portée limitée de nos résultats, il est légitime de se poser la question suivante: est-ce que l'inclusion de métadonnées, selon le modèle du Dublin Core, dans un document HTML est une opération vraiment utile? Avant de répondre à cette question, examinons le projet du Resource Description Framework (RDF) mené par le World Wide Web Consortium.

LE DUBLIN CORE ET LE RDF

Le Resource Description Framework (RDF) est une spécification présentement (je rappelle

au lecteur que nous sommes à l'hiver 1998) développée par le World Wide Web Consortium qui vise à établir une infrastructure pour le traitement des métadonnées. Parmi les nombreuses applications envisagées, le RDF pourra servir : "*in resource discovery to provide better search engine capabilities; in cataloging for describing the content and content relationships available at a particular Web site, page, or digital library; by intelligent software agents to facilitate knowledge sharing and exchange; in content rating; in describing collections of pages that represent a single logical "document"; for describing intellectual property rights of Web pages, and in many others*"(W3C, "Introduction to RDF Metadata").

L'objectif général du RDF "*is to define a mechanism for describing resources that makes no assumptions about a particular application domain, nor defines the semantics of any application domain. The definition of the mechanism should be domain neutral, yet the mechanism should be suitable for describing information about any domain*" (W3C, "RDF Model and Syntax"). À cette fin, le RDF utlise le XML (eXtensible Markup Language) comme langage. Issu du SGML (Standard Generalized Markup Language), la norme ISO de structuration des documents, le XML est un langage qui, en plus d'intégrer Unicode, la norme mondiale de conversion de tous les formats d'alphabet, permet de créer autant de balises que nécessaire (Lubkov, 1997). En effet, à la différence du HTML (Hypertext Markup Language), lui aussi un dérivé du SGML, le XML ne fait que décrire une syntaxe pour le balisage. Par conséquent, "*the names of the tags are not set in concrete and authors can "invent" them as appropriate. Whereas HTML provides a fixed repertoire of named tags - P for 'paragraph', H1 for 'heading 1' and so on, XML simply spells out the rules for using angle brackets and other notation to specify a mark-up language of your own design. The tag names and what they actually mean are left for you to choose as appropriate for your application*" (W3C, "XML Activity"). Le RDF comprend 3 éléments de base: 1) des noyaux (nodes) 2) des attributs (attributes) et 3) des valeurs (values). "*Nodes can be any web resources (pages, servers, basically anything for which you can give a URI), even other instances of metadata. Attributes are named properties of the nodes, and their values are either atomic (text strings, numbers, etc.) or other resources or metadata instances. In short, this mechanism allows us to build labeled directed graphs*" (W3C, "Introduction to RDF Metadata"). La réunion de ces éléments donne lieu à des assertions (assertions) contenues dans des "series" (serializations) qui sont précédées d'instructions (Processing Instructions) permettant d'identifier la source des abréviations utilisées dans les assertions. Par exemple, l'affirmation selon laquelle "John Smith is the Author of the document whose URL is http://www.bar.com/some.doc" s'exprimerait comme suit:

<?namespace href="http://docs.r.us.com/bibliography-info as="bib"?
<?namespace href="http://www.w3.org/schemas/rdf-schema" as="RDF"?
<RDF:serialization
<RDF:assertions href="http://www.bar.com/some.doc"
<bib:authorJohn Smith</bib:author
</RDF:assertions
</RDF:serialization

Du moins en octobre 1997, car depuis ce temps la syntaxe du RDF a grandement évolué. Dorénavant, c'est-à-dire d'après la version d'octobre 1998 (W3C, "*Resource Description Framework (RDF) Schemas*"), cette affirmation s'exprimerait plutôt ainsi:

<rdf:RDF
<xmlns:rdf="http://www.w3.org/schemas/rdf-schema"
<xmlns:bib="http://docs.r.us.com/bibliography-info" <rdf:Description about="http://www.bar.com/some.doc"
<bib:authorJohn Smith</bib:author
</rdf:Description
</rdf:RDF

Selon les derniers états de la syntaxe, l'exemple du *Relevé photographique du canal de Lachine* montrant les métadonnées du Dublin Core en format HTML prendrait donc la forme suivante (forme qui, il est important de le souligner, reste à être confirmée par les responsables du Dublin Core):

<rdf:RDF
<xmlns:rdf="http://www.w3.org/TR/WD-rdf-syntax/"
<xmlns:dc= "http://purl.oclc.org/metadata/dublin_core/" <rdf:Description about="http://tornade.ere.umontreal.ca/~lemayyv/projet/ps5p25.htm"
<dc:TitlePont des Seigneurs</dc:Title
<dc:Creator.PersonalNameLavallée, Chantal</dc:Creator.PersonalName
</RDF:assertions
<dc:Subject
<rdf:Bag
<rdf:liCanals--Canada--Montréal--1997</rdf:li
<rdf:liBridges--Canada--Montréal--1997</rdf:li
</rdf:BAG
<dc:schemeTGM1</dc:scheme
</dc:Subject
<dc:DescriptionPhotographie: coul.; 10x15 cm + légende. PORTÉE ET CONTENU: Cette photographie est la 29e du secteur 5 (Pointe Saint-Charles/Petite-Bourgogne): piste cyclable du *Relevé photographique du canal de Lachine*. NOTES: Titre officiel propre. Légende: Nom: Pont des Seigneurs, Localisation: Rue des Seigneurs, au-dessus de l'écluse 3, Numéro de référence: 5OC05C.</dc:Description
</rdf:Description
</rdf:RDF

Et ainsi de suite pour chacun des 15 éléments de description du Dublin Core. À noter que le RDF est conçu pour supporter des relations binaires. Par conséquent, pour exprimer plus de deux relations, il est nécessaire de les décomposer en autant de sous-ensembles, comme on peut déjà le constater avec l'élément *"Subject"* dans notre exemple. Le Dublin Core semble donc parfaitement adaptable au RDF et la chose n'est pas surprenante comme en fait foi cette déclaration tirée de l'un des documents de travail du World Wide Web Consortium: "*One obvious application for RDF is in the description of web pages. This is one of the basic functions of the Dublin Core [DC] initiative. The Dublin Core is a set of 15 elements believed to be broadly applicable to describing web resources to enable their discovery. The Dublin Core has been a major influence on the development of RDF. An important consideration in the development of the Dublin Core was to allow simple descriptions, but also to provide the ability to qualify descriptions in order to provide both domain specific elaboration and descriptive precision*" (W3C, "RDF Schemas").
Mais quelles sont les chances que cette nouvelle norme en matière de métadonnées et que le nouveau langage sur lequel elle s'appuie connaissent une large diffusion sur le World Wide Web? Très bonnes, semble-t-il, si l'on se fie à la stratégie de développement utilisée par le Consortium. "*The RDF working group - the W3C vehicle for crafting new standards - includes representatives from key companies and organizations: Netscape, Microsoft, IBM, Nokia, OCLC, etc. The interest from the large web browser vendors gives us hope that large scale deployment of tools which understand about RDF will take place; this in turn should lead to the widespread adoption of RDF on the web*"(W3C, "Introduction to RDF Metadata").

CONCLUSION:
LES MÉTADONNÉES COMME OUTIL DE GESTION

Quels avantages le Dublin Core représente-t-il dans la gestion des archives photographiques? Les normes de description des documents d'archives ont-elles un rôle important à jouer dans l'élaboration du contenu des métadonnées? Qu'en est-il de ces deux objectifs qui motivaient notre recherche?
De prime abord, l'on serait porté à dire que les métadonnées issues des 15 éléments de description du Dublin Core sont à toutes fins pratiques inutiles dans la gestion des archives photographiques et qu'il serait préférable d'investir son temps et ses énergies ailleurs. En effet, comme nous avons pu le constater, la présence des éléments du Dublin Core n'a changé en rien les

résultats obtenus lors du repérage du site temporaire par les différents moteurs de recherche. Alors pourquoi s'embarrasser de données sur les données puisque, dans leur format actuel, elles sont ignorées par les moteurs de recherche?
Malgré les apparences, il serait prématuré de tout abandonner maintenant, particulièrement après ce que nous venons de voir à propos du RDF. Non seulement son développement a été influencé par le Dublin Core mais, compte tenu du rôle joué par le World Wide Web Consortium dans le milieu informatique et de l'implication des principales compagnies et organisations dans le dossier, il permettra au Dublin Core de disposer enfin de tous les moyens nécessaires pour s'imposer.
Ce geste serait d'autant plus prématuré, qu'entre temps, il est toujours possible de mettre en place des solutions permettant de tirer profit des métadonnées du Dublin Core. Par exemple, il suffit de jouer le jeu des moteurs de recherche et de les faire précéder par les trois types de métadonnées (title, keywords et description) qu'ils prennent en considération. Une autre solution et qui, croyons-nous, serait encore plus profitable, du moins dans la perspective d'une interrogation en mode local, consiste en ceci : il suffirait que la description qui accompagne un document photographique adopte non pas le format de présentation des RDDA mais celui du Dublin Core. De la sorte, comme le moteur de recherche indexe tous les mots contenus dans un document HTML, il serait possible de simuler, si l'on peut dire, une interrogation selon les 15 éléments du Dublin Core. Et rien n'empêche de combiner les deux solutions.
Bref, nous sommes persuadé que ceux qui persisteront à incorporer les métadonnées du Dublin Core à leurs documents HTML disposeront dans l'avenir d'outils de gestion des archives photographiques qui feront l'envie de tous. Description détaillée de l'image et de son environnement. Meilleur repérage tant sur le Web qu'en mode local. Production de listes permettant de gérer l'organisation des documents photographiques. Possibilité d'établir des liens entre les photographies provenant de différentes sources tout en préservant leur contexte d'origine. Sans compter que ces outils développés en fonction de normes seront mieux à même de s'adapter à l'évolution de l'environnement électronique.
Quant à la question des normes de description, la réponse est plus nette. À la suite de cette recherche, il est clair que les RDDA, par leur caractère normalisé, s'avèrent un élément essentiel dans l'élaboration du contenu des métadonnées. Bien sûr, il est nécessaire d'adapter le format de ces règles à celui du Dublin Core mais, comme nous avons pu le constater, cette transposition ne porte nullement atteinte à leur intégrité.
Mais pour faire en sorte que les RDDA deviennent la référence sur le Web en matière de métadonnées de documents d'archives, la communauté archivistique devra entreprendre diverses actions. Pour que ces règles puissent s'imposer, elle devra entre autres: 1) repenser les RDDA en fonction du contexte électronique, 2) donner accès à une version électronique en ligne des RDDA, 3) développer un gabarit du même genre que le Dublin Core Metadata Template permettant de générer et d'intégrer des métadonnées aux documents d'archives et 4) rendre disponible le chapitre des RDDA sur les documents électroniques.
Nul doute, l'archivistique est et doit, plus que jamais, être une discipline méta-informationnelle.

BIBLIOGRAPHIE

- ALTA VISTA (Page consultée le 26 mars 1998) "*Alta Vista Search: Adding a URL*" [En ligne]. Adresse URL: http://altavista.digital.com/av/content/addurl_meta.htm
- ARCHIVES AND MUSEUM INFORMATICS (1996). "CNI/OCLC Workshop on Metadata for Networked Images". *Archives and Museum Informatics*, vol. 10, no 3, 1996, p. 270-276.
- ARCHIVES NATIONALES DU CANADA (1993). *La gestion des documents photographiques au gouvernement du Canada*. Ottawa, Archives nationales du Canada, 1993, 43 p.
- BUREAU CANADIEN DES ARCHIVISTES (1990). *Règles pour la description des documents d'archives*. Ottawa, Bureau canadien des archivistes, 1990.
- BEARMAN, DAVID (1996). "Possible Contributions of the Reference Model of Metadata Required for Evidence to a Reference Model of Metadata Required for Image Description". *Archives and Museum Informatics*, vol. 10, no 3, 1996, p. 295-302.
- BEARMAN, DAVID (1996). "Developments in Metadata Management Framework". *Archives and Museum Informatics*, vol. 10, no 2, 1996, p. 185-188.
- BEARMAN, DAVID; DUFF, WENDY (1996). "Grounding Archival Description in the Functional Requirements for Evidence". *Archivaria*, no 41, printemps 1996, p. 275-303.
- BECKER, HANS ET AL. (Page consultée le 12 février 1998). "*Dublin Core Element : Coverage*" [En ligne]. Adresse URL: http://www.sdc.ucsb.edu/~mary/coverage.htm

- BESSER, HOWARD (1997). "Images Databases : The First Decade, the Present, the Future". *Digital Image Access and Retrieval*. P. Bryan Heidorn; Beth Sandore (éds). Urbana-Champaign, University of Illinois, 1997, p. 11-28.
- COX, RICHARD J. (1997). "Electronic Systems and Records Management in the Information Age : An Introduction". *Bulletin of the American Society for Information Science*, vol. 23, no 5 (juin-juillet 1997), p. 7-9.
- DANIELS JR., RON; IANELLA, RENATO; MILLER, ERIC (Page consultée le 23 avril 1998). "*Expressing the Dublin Core in the Resource Description Framework : Suggestions based on an early examination of the problem*" [En ligne]. Adresse URL: http://www.acl.lanl.gov/~rdaniel/RDF/DC/ExpDC_2.html
- DUBLIN CORE (Page consultée le 22 janvier 1998). "*A Bibliography of Materials Related to the Dublin Core*" [En ligne]. Adresse URL: http://www-diglib.stanford.edu/diglib/pub/dublin.html
- DUBLIN CORE (Page consultée le 17 novembre 1997). "*Dublin Core Metadata Element Set: Reference Description*" [En ligne]. Adresse URL: http://purl.oclc.org/metadata/dublin_core/
- INTERNATIONAL FEDERATION OF LIBRARY ASSOCIATIONS AND INSTITUTIONS (Page consultée le 22 janvier 1998). "*Digital Libraries: Metadata Resources*" [En ligne]. Adresse URL: http://www.ifla.org/II/metadata.htm
- LUBKOV, MICHEL (1997). "SGML, HTML, XML, des normes pour les documents". *Archimag*, no 107 (septembre 1997), p. 30-31.
- META WEB PROJECT (Page consultée le 26 mars 1998). "*HotMeta Broker Control Panel*" [En ligne]. Adresse URL: http://flare.dstc.edu.au:8017/control1.html
- META WEB PROJECT (Page consultée le 26 mars 1998). "*HotMeta Search Engine*" [En ligne]. Adresse URL: http://www.dstc.edu.au/RDU/MetaWeb/test/hotmeta.html
- NORDIC METADATA PROJECT (Page consultée le 22 janvier 1998). *Dublin Core Metadata Template* [En ligne]. Adresse URL: http://www.lub.lu.se/cgi-bin/nmdc.pl
- NORDIC WEB INDEX (Page consultée le 30 avril 1998). *Module de recherche en anglais* [En ligne]. Adresse URL: http://nwi.lub.lu.se/?lang=en
- PARKER, ELISABETH BETZ (1987). *LC Thesaurus for Graphic Materials : Topical Terms for Subject Access*. Washington, D.C., Library of Congress, 1987, 591 p.
- PREBEN, HANSEN (Page consultée le 22 janvier 1998). "*User Guidelines for Dublin Core Creation*" [En ligne]. Adresse URL: http://www.sics.se/~preben/DC/DC_guide.html
- SUBMIT IT (Page consultée le 9 février 1998). "*Submit It!® Announcement Services*" [En ligne]. Adresse URL: http://www.Submit-it.com
- WALLACE, DAVID (1996). "Managing the Present : Metadata as Archival Description". *Archivaria*, no 39, printemps 1995, p. 11-21.
- WALLACE, DAVID, A. (1993). "Metadata and the Archival Management of Electronic Records: A Review". *Archivaria*, no 36, automne 1993, p. 87-110.
- WEB DEVELOPPER'S VIRTUAL LIBRARY (Page consultée le 2 avril 1998). "*META Tagging for Search Engines*" [En ligne]. Adresse URL: http://WWW.Stars.com/Search/Meta/Tag.html
- WEBTECHS (Page consultée le 2 avril 1998). "*WebTechs HTML Validation Service*" [En ligne]. Adresse URL: http://valsvc.webtechs.com/
- WEIBEL, STUART L.; LAGOZE, CARL (1997). "An Element Set to Support Resource Discovery : The State of the Dublin Core, January 1997". *International Journal on Digital Libraries*, vol. 1, no 2 (septembre 1997), p. 176-186.
- WEIBEL, STUART L. (Page consultée le 22 janvier 1998). "*A Proposed Convention for Embedding Metadata in HTML*" [En ligne]. Adresse URL: http://www.oclc.org:5046/~weibel/html-meta.html
- W3C (Page consultée le 20 avril 1998). "*Introduction to RDF Metadata*" [En ligne]. Adresse URL: http://www.w3.org/TR/NOTE-rdf-simple-intro"
- W3C (Page consultée le 20 avril 1998). "*Press Release: W3C Issues XML 1.0 as a Recommandation*" [En ligne]. Adresse URL: http://www.w3.org/Press/1998/XML10-REC"
- W3C (Page consultée le 20 avril 1998). "*XML Activity*" [En ligne]. Adresse URL: http://www.w3.org/XML/Activity.html"
- W3C (Page consultée le 20 avril 1998). "*Resource Description Framework (RDF) Schemas*" [En ligne]. Adresse URL: http://www.w3.org/TR/WD-rdf-schema/"
- W3C (Page consultée le 5 décembre 1997). "*Resource Description Framework (RDF) Model and Syntax*" [En ligne].

Adresse URL: http://www.w3.org/TR/WD-rdf-syntax/

- W3C (Page consultée le 17 novembre 1997). "*Resource Description Framework (RDF)*" [En ligne]. Adresse URL: http://www.w3.org/Metadata/RDF/

L'analyse de textes littéraires assistée par ordinateur: une introduction

par

Véronique Parenteau

Cursus vol. 4 no 1 (automne 1998)

Cursus est le périodique électronique étudiant de l'École de bibliothéconomie et des sciences de l'information (EBSI) de l'Université de Montréal. Ce périodique diffuse des textes produits dans le cadre des cours de l'EBSI.

ISSN 1201-7302

C. élec. : cursus@ere.umontreal.ca
URL : http://www.fas.umontreal.ca/ebsi/cursus/

Droits d'auteur

L'auteure

Après des études de premier cycle en Études littéraires à l'Université du Québec à Montréal, Véronique Parenteau a obtenu sa maîtrise à l'EBSI avec le profil "Analyse de l'information et bases de données" au printemps 1998. Durant l'été 1997, elle a travaillé à la création du Répertoire de sites Web de référence du Québec (http://www2.biblinat.gouv.qc.ca/wgraphie/intro.htm) de la Bibliothèque nationale du Québec. Depuis juin 1998, elle est bibliothécaire pour l'entreprise CEDROM-SNi.

Le texte suivant a été produit dans le cadre du cours BLT 6271, Recherche en analyse documentaire,

sous la direction de Mme Michèle Hudon.

Pour joindre l'auteure : parenteauv@cedrom-sni.qc.ca

Table des matières

"The statistical analysis of a literary text can be justified by the need to apply an objective methodology to works which for a long time may have received only impressionistic and subjective treatment. Hesitation by literary scholars and mistrust of such a blatantly quantitative approach may be alleviated by choosing the least contestable mode of analysis, namely that of counting" (Holmes, 1994, p.87).

Introduction

Certains diront que, s'il y a un domaine que l'informatique n'a pas encore envahi, c'est bien la littérature, cet art très ancien que seul le cerveau humain peut produire, lire et comprendre. Comment un ordinateur pourrait-il intervenir dans l'analyse d'un texte issu de l'imagination d'un homme ou d'une femme? Littérature et informatique ne semblent pas pouvoir faire très bon ménage.

Pourtant, il existe un certain nombre de spécialistes - pour la plupart, des non littéraires - qui misent sur cet outil pour les assister dans leurs analyses littéraires. Il s'agit bien d'une "assistance", puisque

l'ordinateur ne peut pas analyser une oeuvre littéraire comme le ferait un chercheur humain. Il faut évidemment une bonne dose d'interprétation dans l'analyse littéraire et c'est une opération que même le plus évolué des logiciels connus ne parviendra pas à accomplir.

La grande "nouveauté" apportée par l'informatique aux études littéraires, c'est une façon d'analyser les textes quantitativement, statistiquement. Ce sont des méthodes qui permettent de traduire en chiffres, en tableaux, en graphiques, des données textuelles qui font généralement l'objet d'analyses qualitatives.

Qu'est-ce que l'informatique peut apporter à l'étude de la littérature? Qu'est-ce que l'analyse statistique permet de découvrir, de comprendre au sujet des oeuvres littéraires? Qu'est-ce qu'un ordinateur peut faire de plus qu'un être humain en ce domaine? Quels usages les chercheurs font-ils de cet outil qu'est l'ordinateur? Quelles sont les limites de l'informatique dans l'étude de la littérature?

C'est à ces questions que ce texte veut répondre en proposant un état de la question sur l'analyse de textes littéraires assistée par ordinateur. Il ne s'agit donc pas de décrire des logiciels d'analyse de texte ni de comprendre le fonctionnement technique et statistique de l'analyse de texte par ordinateur. Il s'agit plutôt de comprendre comment l'ordinateur peut être utile à l'analyse littéraire et quelles sont ses limites.

La première partie de cet exposé vise à définir brièvement en quoi consiste l'analyse de textes littéraires assistée par ordinateur (i.e. leur analyse statistique) et, plus particulièrement, la stylométrie qui est une façon de quantifier le style. La seconde partie présente les procédés selon lesquels les logiciels informatiques peuvent mesurer le style d'un texte littéraire, les variables dont ils tiennent compte. Ensuite, plusieurs exemples concrets d'analyses statistiques sont présentés afin d'illustrer les usages que les spécialistes font de l'ordinateur dans leurs recherches en littérature. La section suivante montre la complémentarité entre le travail humain et celui de la machine. Enfin, la cinquième et dernière partie est consacrée aux limites de l'analyse de textes littéraires par ordinateur et à la synthèse de différentes critiques formulées à l'endroit de cette façon d'étuder la littérature.

1. L'analyse statistique de la littérature

Dans le domaine des études littéraires, on rencontre surtout des analyses qualitatives d'oeuvres, de courants, de genres. Dans le cas d'études visant à déterminer la paternité d'un texte, par exemple, on se tourne généralement vers l'opinion des experts en littérature sur le style et les subtilités de l'usage du langage, du vocabulaire et de la grammaire. Arriver à un consensus des opinions, voilà le problème de la plupart des domaines impliquant une grande part "d'intuition" humaine et d'"expérience", comme c'est le cas en études littéraires. Les méthodes quantitatives et statistiques d'analyse des données pourraient avoir beaucoup à offrir aux sciences humaines, dont font partie les études littéraires (Lowe et Matthews, 1995). Elles peuvent apporter des informations supplémentaires qui sont quantifiables.

La majorité des analyses de textes littéraires impliquant l'informatique utilisent la stylométrie comme moyen d'analyse. Il s'agit en fait d'une forme de quantification du style. Le style d'un auteur est ce qui tend à distinguer son écriture entre toutes.

"Recognizing, for example, that even a writer who flaunts an abstruse vocabulary will also need to use many mundane words, stylisticians regard style as a general predisposition toward a particular mode of expression rather than an invariant habit or constant" (Sigelman et Jacoby, 1996, p.11).
Chaque texte se définit par un ensemble de caractéristiques statistiques, mesurables. Si plusieurs oeuvres d'un auteur comportent les mêmes caractéristiques, l'auteur fait un usage récurrent d'un style particulier. Si l'oeuvre d'aucun autre auteur ne possède les mêmes caractéristiques, on peut dire que son style est unique (Holmes, 1994).

La stylométrie - aussi appelée "statistique stylistique" - est l'application des méthodes mathématiques pour extraire des mesures quantitatives d'un texte (Lowe et Matthews, 1995). Les données sur lesquelles se penche la stylométrie, ce sont les mots. Ils sont la matière brute de cette science. Selon David I. Holmes, aucun stylométriste n'est encore parvenu à établir une méthodologie qui arrive à mieux saisir le style d'un texte que celle qui s'appuie sur des éléments lexicaux (1994). Holmes explique qu'il n'y a pas meilleurs paramètres pour établir une comparaison objective entre des auteurs:
The lexical level is the obvious place to initiate stylistic investigations, since questions about style are essentially comparative and more data exist at the lexical level than at any other in the form of computed concordances (1994, p.91).

Les caractéristiques stylistiques d'un texte doivent, pour être étudiées par ordinateur, avoir ces propriétés décrites par Bailey: "they should be salient, structural, frequent and easily quantifiable, and relatively immune from conscious control" (Holmes, 1994, p.88, citant Bailey, 1979). On espère, en mesurant de telles caractéristiques, découvrir l'unicité de l'écriture d'un auteur et arriver à distinguer son style de celui d'un autre. On veut faire la distinction entre les véritables différences stylistiques et les variations dues au hasard (Holmes, 1994).

Le meilleur outil pour faire l'analyse stylométrique d'un texte, pour en tirer des statistiques, est probablement l'informatique. Les logiciels d'analyse de texte[1] permettent aux chercheurs de repérer des mots et expressions, de produire des profils statistiques, des graphiques, des tableaux et ce, rapidement et efficacement.

2. Les procédés

L'ordinateur ne peut évidemment pas analyser un texte avec la même profondeur que le ferait un chercheur humain. Les objets de l'analyse par ordinateur sont de deux ordres: les données "brutes" (les chaînes de caractères, les syllabes, la ponctuation, etc.) et celles qui sont plus de l'ordre du contenu (le vocabulaire et les thèmes).

2.1 Analyse de données "brutes"

Comme cela a déjà été mentionné plus haut, les mots sont les données brutes de la stylométrie. L'ordinateur peut les identifier grâce aux espaces et aux marques de ponctuation (Fortier, 1995).

L'ordinateur permet de compter les mots contenus dans un texte, de repérer ceux qui sont les plus utilisés, de les localiser pour mieux voir le contexte de leur utilisation ou encore dans le but de faire un index qui facilitera leur repérage ultérieur, de déterminer les intervalles entre les différentes occurences d'un mot, etc. (Johnson, 1996b). En 1887, dans son article "The Characteristic Curves of Composition" paru dans la revue *Science*, T.C. Mendenhall affirmait que la longueur des mots était une caractéristique pouvant permettre de distinguer les auteurs (Holmes, 1994). Depuis, plusieurs études ont été faites à partir de cette théorie. L'avènement de l'informatique a de beaucoup facilité ce type d'analyse.

Une autre façon de découvrir des traits stylistiques distinctifs, selon Holmes, est de calculer les pourcentages de noms, verbes, adjectifs, adverbes, etc., à condition, bien sûr, qu'ils puissent être reconnus fidèlement (1994). En effet, certains mots peuvent laisser planer un doute quant à leur nature (les mots *brise*, *marche* et *porte*, par exemple, peuvent aussi bien être des noms communs que des verbes accordés au présent de l'indicatif).

L'usage des mots offre plusieurs possibilités de discrimination. Selon Holmes, "some words vary considerably in their rate of use from one work to another by the same author. For discrimination purposes we need context-free or "function" words to be able to conduct reliable comparisons between literary works" (1994, p.90).

Outre les mots, d'autres aspects d'un texte peuvent être pris en considération par l'ordinateur pour analyser le style. Les syllabes, notamment, peuvent apporter de bons indices aux études sur la paternité d'un texte, selon Holmes. Certains auteurs ont un style plus homogène en ce qui concerne la distribution du nombre de syllabes par mot (Holmes, 1994).

Les signes de ponctuation sont également facilement repérables et analysables par ordinateur. Une étude menée par Étienne Brunet montre, par exemple, l'usage des divers signes de ponctuation chez six auteurs français: Marcel Proust, Émile Zola, René de Chateaubriand, Jean-Jacques Rousseau, Jean Giraudoux, et Victor Hugo. Brunet a notamment découvert que Chateaubriand et Rousseau sont les seuls à continuer à cultiver les signes de ponctuations "intermédiaires" (deux points, point-virgule), que Hugo et Giraudoux utilisent beaucoup le point et les signes exepressifs (point d'interrogation et point d'exclamation) et que Proust utilise beaucoup la virgule, étant donné que ses phrases sont très longues (Brunet, 1991)[2].

L'ordinateur peut aussi permettre d'étudier l'évolution de la longueur des phrases dans un texte ou encore la place des dialogues par rapport aux passages narratifs (Johnson, 1996b).

2.2 Analyse du contenu

Les analyses basées sur les données "brutes" peuvent paraître assez simples, mais on peut s'en servir pour aller plus loin en analysant le vocabulaire et les thèmes (Johnson, 1996b).

2.2.1 Le vocabulaire

L'une des notions fondamentales de la stylométrie, c'est la mesure de ce que l'on peut appeler la "richesse" et la "diversité" du vocabulaire de l'auteur. La prémisse de base est que l'auteur a à sa disposition une certaine quantité de mots, une certaine "banque" de mots et que parmi ceux-là, il en privilégiera certains aux dépens des autres (Holmes, 1994). Si on prend un échantillon de l'oeuvre d'un auteur, on peut s'attendre à y retrouver le reflet de son vocabulaire. Si on peut trouver une mesure qui puisse représenter statistiquement le vocabulaire, on pourrait l'utiliser pour fins de comparaisons.

Une telle mesure existe. En anglais, on l'appelle le "type-token ratio". Il s'agit du nombre d'unités lexicales (i.e. le nombre de formes différentes) formant le vocabulaire de l'échantillon divisé par le nombre d'unités (i.e. le nombre total de mots) formant l'échantillon (Holmes, 1994)[3]. Une autre façon de mesurer le vocabulaire est de calculer la chance que les deux membres d'une paire de mots (choisie au hasard) appartiennent au même "type" (i.e. la même forme d'un mot). On peut aussi se servir de la fréquence des mots en comptant le nombre de mots utilisés une fois, deux fois, trois, quatre, etc. Plus il y a de mots qui ne reviennent pas souvent, plus le vocabulaire est riche. On pourrait aussi tenter d'établir la quantité de mots rares ou techniques. Mais il faudrait alors une bonne part de travail humain pour déterminer quels mots pourraient appartenir à ces catégories (Holmes, 1994).

Le vocabulaire peut être utilisé comme moyen de comparaison entre plusieurs textes, comme l'a fait notamment Étienne Brunet. Afin de comparer le vocabulaire des différents textes de Victor Hugo, il a fait une analyse qui tient compte de ce qui pourrait être traduit comme la "jonction lexicale" ("the lexical connection"), c'est-à-dire la distance entre le vocabulaire de deux textes. Plusieurs textes de Hugo ont été étudiés par paires. Pour chacune de ces paires, le chercheur a considéré tous les mots des deux textes étudiés et la fréquence de chacun dans chaque texte. Les calculs effectués par l'ordinateur ont permis de tracer une carte que Brunet décrit ainsi: "at the bottom are grouped all the poetic collections, while the novels and plays are placed in the upper half, without merging too much into each other" (Brunet, 1991, p.76). Il a donc constaté que le vocabulaire des poèmes se distingue de celui des romans, des pièces de théâtre et des lettres.

2.2.2 Les thèmes et les champs lexicaux

L'étude par ordinateur du vocabulaire dans un texte peut aussi permettre de saisir les thèmes dont il est principalement question dans ce texte. Il s'agit d'utiliser l'ordinateur pour tracer la distribution du vocabulaire, des mots qui évoquent les différents thèmes (Fortier, 1995). La présence de concepts donnés et leur importance relative n'est pas toujours évidente à l'oeil nu (Laffal, 1995). Leur repérage par ordinateur peut donc faciliter le travail du chercheur.

Pour Julius Laffal, l'analyse des concepts d'un texte a deux utilités: "One is to gain insight into the similarities and differences between the texts for comparative studies. The second is to garner cues to the author's conceptual orientation in the texts under study" (1995, p.343).

L'analyse de concepts - telle que décrite par Laffal (1995) - est une forme d'analyse de contenu basée sur des catégories d'idées (ou *concepts*) représentées par les mots d'un texte. Chaque mot du texte est cherché par l'ordinateur dans un dictionnaire contenant des mots auxquels les concepts qu'ils évoquent sont associés. Puis, un profil de la fréquence et de la distribution des concepts est généré. Paul A. Fortier (1995) précise que le texte doit être encodé avant d'être comparé au dictionnaire. L'encodage dont il parle concerne les unités linguistiques et les parties du discours.

Pour construire le dictionnaire automatisé, les thèmes sont identifiés d'après des catégories sémantiques. La catégorisation se construit beaucoup par oppositions et similitudes à partir desquelles on établit une structure hiérarchique. Le résultat est donc très semblable au célèbre thésaurus de Roget[4] (Fortier, 1995; Laffal, 1995). Chaque catégorie doit être exclusive et pas trop large pour être significative. Lorsque les thèmes ont été établis, il ne reste plus qu'à y associer les différents mots qui les évoquent. Selon le logiciel utilisé, l'ordinateur cherchera le mot exact parmi les entrées du dictionnaire ou bien il pratiquera une certaine lemmatisation et trouvera l'entrée la plus semblable (Laffal, 1995).

Étienne Brunet (1991) a fait l'analyse de l'évolution des thèmes de la nature et du temps dans *À la recherche du temps perdu* de Marcel Proust. Il a donc étudié ces champs lexicaux, c'est-à-dire l'ensemble des mots se rapportant à ces thèmes, dans les sept romans formant ce récit. Les graphiques tracés par le logiciel utilisé par Brunet montraient que le thème de la nature devient moins important au fil de la progression du récit tandis que celui du temps l'est de plus en plus.

Paul A. Fortier (1995) a fait l'analyse de certains champs lexicaux du roman *L'Immoraliste* d'André Gide. Il a étudié plusieurs thèmes (et leurs champs lexicaux) regroupés sous les grands thèmes de la santé et de la maladie du point de vue de la fréquence de leurs occurences pour vérifier leur importance relative les uns par rapport aux autres.

3. Les usages

La stylométrie, l'analyse statistique des mots, du vocabulaire et des thèmes composant une oeuvre littéraire, peut mener à des études un peu plus poussées. On utilise souvent les résultats de ce type d'analyse pour comparer entre eux des oeuvres et des auteurs, pour déterminer la paternité d'un texte, ou pour trouver ce qui distingue une oeuvre "authentique" d'un pastiche ou d'une imitation. L'analyse des motifs rythmiques des vers en poésie ou dans les pièces de théâtre par l'étude des mots et des syllabes est une autre façon d'observer le style d'un auteur.

3.1 Comparaisons

L'un des types d'études que l'on retrouve le plus fréquemment dans le domaine de l'analyse de textes littéraires assistée par ordinateur, est la comparaison de deux ou plusieurs textes entre eux, qu'ils soient d'un même auteur ou d'auteurs différents. La comparaison peut être basée sur différents aspects, notamment le vocabulaire et les concepts abordés.

3.1.1 Par le vocabulaire

Le chercheur Lee Sigelman (1995), s'est servi de l'analyse du vocabulaire par ordinateur pour juger de la qualité de travail de Marion Mainwaring qui a tenté de compléter *The Buccaneers*, un roman de l'américaine Edith Wharton. Cette oeuvre, demeurée inachevée en raison du décès de l'auteure en 1937, avait tout de même été publiée en 1938. Aux vingt-neuf chapitres qui avaient été écrits, Mainwaring en a ajouté douze. L'oeuvre "complétée" a été publiée dans la controverse en 1993. Les critiques étaient effectivement très sceptiques et divisés quant au succès avec lequel Mainwaring avait réussi son entreprise. C'est pourquoi Sigelman a voulu analyser les deux parties de l'oeuvre (celle produite par Wharton et celle ajoutée par Mainwaring) afin d'évaluer la fidélité avec laquelle Mainwaring avait complété le récit de Wharton.

Sigelman a fait une analyse, chapitre par chapitre, du ratio entre les mots nouveaux (i.e. n'apparaissant pas dans les chapitres précédents) et le nombre total de mots ("the ratio of new types to tokens", 1995, p.273). L'avantage de cette mesure statistique, c'est qu'elle est relativement simple et que la richesse du vocabulaire - qu'elle mesure - est reconnue comme étant une caractéristique généralement stable chez un auteur.

Sigelman a d'abord appliqué cette méthode à trois romans d'Edith Wharton: *The House of Mirth* (1905), *Ethan Frome* (1911) et *The Age of Innocence* (1920). Il a ainsi pu constater que le ratio de nouveaux mots suivait une progression semblable dans chacun des textes de Wharton. L'analyse des vingt-neuf chapitres de *The Buccaneers* écrits par Wharton a ensuite montré que cette progression suivait, là aussi, sensiblement le même modèle. Par contre, Sigelman a pu remarquer une brisure dans la progression de l'apparition de nouveaux mots dans les chapitres ajoutés par Mainwaring. "The ratio of types to tokens turns out to be 8840/89494, or 0.099, for Wharton's twenty-nine chapters of *The Buccaneers*, as compared to 5791/33023, or 0.175, for Mainwaring's twelve chapters" (1995, p.273). Il apparait donc que Mainwaring emploie un vocabulaire plus riche que Wharton. Cependant, pour que l'analyse soit plus représentative, Sigelman a repris l'analyse avec des échantillons de même taille, soit les douze chapitres de Mainwaring et les dix premiers de Wharton (un peu plus de 30,000 mots dans chaque cas). La différence entre le ratio de nouveaux mots chez les deux auteurs était alors beaucoup moins importante: 0,159 pour Wharton et 0.175 pour Mainwaring. Il y a donc peu de différence dans la richesse du vocabulaire si on étudie globalement chacune des deux parties.

Mais ce qu'il est plus intéressant de vérifier, c'est si la narration du roman est "rompue" à cause du changement d'auteur, en observant s'il y a rupture dans l'évolution du ratio de nouveaux mots d'un chapitre à l'autre. Sigelman pose la question: "Is the junction between Wharton's and Mainwaring's chapters seamless, or does it betray clear evidence of disruption of the narrative flow?" (1995, p.274). Il a comparé - du point de vue du ratio de nouveaux mots - les chapitres de Mainwaring avec les autres chapitres de *The Buccaneers*, mais aussi les autres romans de Wharton. Il a pu voir que dans les romans de Wharton et ses chapitres de *The Buccaneers*, le ratio fait une parabole vers le bas, avec quelques rares et faibles remontées. Cela s'explique simplement par le fait que, au début de la narration, l'auteur a

besoin de plusieurs mots pour décrire les lieux, introduire les personnages, mettre l'action en contexte. Dans *The Buccaneers*, à l'endroit où Mainwaring fait son entrée (au chapitre 30), il y a une grosse remontée du ratio, donc beaucoup de nouveaux mots d'un coup.

Deux autres chercheurs, J.F. Burrows et D.H. Craig (1994), ont aussi utilisé l'analyse du vocabulaire pour fins de comparaison. Des critiques ont qualifié les tragédies anglaises romantiques de pauvres imitations des tragédies de la Renaissance. Burrows et Craig ont voulu voir à quel point ces deux groupes d'oeuvres étaient semblables ou dissemblables. Leur objectif était de déterminer les différences systématiques entre ces deux groupes de textes du même genre mais produits à des périodes très éloignées dans le temps. Ils espéraient ainsi apporter de nouveaux éléments au débat.

Burrows et Craig ont choisi dix pièces de chaque groupe et ont fait des comparaisons statistiques sur la base des mots les plus utilisés. L'analyse leur a fait voir des différences évidentes entre les deux groupes, allant au-delà des simples changements historiques du langage. "The Romantic tragedies are more expository; the Renaissance ones include more commonplace interactions between characters. The later plays do not show the marked variations in function-word frequencies of their predecessors" (Burrows et Craig, 1994, p.63). Ils ont pu constater que, parmi les pièces de la Renaissance, celles de William Shakespeare présentent à la fois les plus grandes affinités et les plus grandes différences par rapport aux tragédies romantiques.

3.1.2 Par les concepts abordés

L'analyse de concepts est un autre bon moyen de comparaison. On peut comparer, par exemple, les profils de concepts de deux oeuvres entre eux ou encore les fréquences relatives de différents concepts pour une même oeuvre.

C'est un peu ce qu'a fait Julius Laffal (1995). Voici comment il décrit son procédé:
To determine if profiles A and B are significantly different both are correlated with a third profile, C, thus providing the two correlation values, r(AC) and r(BC). A z' transformation is applied to the correlations. The difference between z'(AC) and z'(BC), divided by the sampling error of the difference, is evaluated for significance against the normal curve (Laffal, 1995, p.342).

Laffal a voulu vérifier si la pensée de Jonathan Swift avait changé entre 1697 et 1725 en étudiant les concepts traités dans deux de ses oeuvres: *A Tale of a Tub* et *Gulliver's Travels*, le premier publié en 1704, mais écrit vers 1697 alors que l'auteur avait trente ans et le second écrit entre 1721 et 1725. Il a donc produit, pour chaque livre, deux profils, l'un étant une liste alphabétique des concepts avec leur fréquence et leur importance (en pourcentage) par rapport au total, et l'autre étant une liste de ces mêmes concepts classés selon leur fréquence. Il a aussi fait l'analyse (et donc établi les profils) d'autres écrits de Swift (des lettres et des poèmes) contemporains à ces deux textes afin de pouvoir évaluer ses résultats dans un contexte plus large. Il a par exemple vérifié, à l'aide des profils de concepts, si *Gulliver* ressemblait davantage aux textes écrits par Swift durant la même année qu'à Tub et d'autres textes de 1697. Les corrélations ont été converties en cote Z et on a observé les différences.

Ces résultats permettent, selon Laffal, de faire trois constats: en ce qui concerne les concepts traités, (1) *Gulliver* est plus semblable aux textes qui lui sont contemporains (1725) qu'à *Tub* et aux autres textes de 1697; (2) *Tub* est légèrement plus semblable aux autres textes de 1697 (qui lui sont contemporains) qu'à *Gulliver*, mais la différence est peu significative; (3) il n'y a pas de différence entre la corrélation de *Tub* avec les autres textes de 1697 et celle de *Gulliver* avec les autres textes de 1725. Ces résultats suggèrent donc, toujours selon Laffal, que les concepts utilisés par Swift n'ont pas changé entre 1697 (lorsqu'il a écrit *Tub*) et 1725 (alors qu'il a écrit *Gulliver*), mais que les concepts spécifiques à *Gulliver* diffèrent de ceux qui sont spécifiques à Tub et aux autres textes de la même époque. De plus, en comparant les profils de *Tub* et *Gulliver* à ceux de textes contemporains produits par d'autres auteurs, Laffal a trouvé que *Gulliver* présente une plus grande corrélation que *Tub* avec d'autres textes du début du 18e siècle. Il affirme:
This finding affirms that Gulliver represents a unique departure in Swift's use of concepts rather than overall shift in his concepts between the 1690s and the 1720s. (...) (It(also suggests that Gulliver is atypical with respect to contemporary 18th century writings (Laffal, 1995, p.346).

Pour mettre tous ces résultats en contexte, il faut cependant en faire une évaluation qualitative, en observant la fréquence des concepts trouvés pour chaque texte. Laffal a donc regardé quels concepts sont plus traités dans *Gulliver* que dans *Tub* et les autres textes du 18e et vice-versa. Il a aussi comparé les concepts utilisés par Swift par rapport à ceux utilisés par les autres auteurs de l'époque pour voir en quoi Swift se distingue de ses contemporains. Il a fait des regroupements de concepts (par exemple, un incluant les concepts reliés à la culture, la religion, le langage et l'éducation) et des oppositions (concepts à consonnance négative vs positive). Il a trouvé notamment que (1) Swift a moins abordé les valeurs négatives (le mal, le crime, la destruction, la mort, la maladie, etc.) que ses contemporains tandis qu'il faisait plus référence aux valeurs positives; (2) il était plus attentif que ses contemporains à ce qui est matériel et au commerce et moins à l'éducation et à la culture, ce qui reflèterait, selon Laffal, son intérêt pragmatique pour la vie quotidienne.

3.2 Déterminer la paternité d'un texte

Les méthodes informatiques de comparaison entre différents textes ou différents auteurs peuvent permettre de déterminer la paternité d'une oeuvre (i.e. son auteur) (Holmes, 1994; Lowe et Matthews, 1995; Elliot et Valenza, 1996). On peut aussi, quand on connaît l'auteur, déterminer la période de sa vie durant laquelle il a écrit un texte (Holmes, 1994; Johnson, 1996b). Une étude statistique voulant déterminer la paternité d'un texte implique des comparaisons du texte en question avec des oeuvres des auteurs possibles en utilisant les tests statistiques appropriés qui analyseront les caractéristiques quantifiables des textes, caractéristiques réflétant le "style" de l'écriture, comme cela a été expliqué précédemment. Il s'agit de déterminer si le texte évoque plus le style de l'auteur A ou de l'auteur B.

Bailey a proposé, en 1979, trois règles pour définir les circonstances nécessaires à la détermination de la paternité d'un texte:
(i) the number of putative authors should constitute a well-defined set; (ii) the lengths of the writings

should be sufficient to reflect the linguistic habits of the author of the disputed text and also those of each of the candidates; (iii) the texts used for comparison should be commensurate with the disputed writing (Holmes, 1994, p.87, citant Bailey, 1979).

S'il y a un auteur pour lequel la question de la paternité se pose souvent, c'est bien William Shakespeare. Plusieurs des oeuvres qui lui ont longtemps été attribuées soulèvent maintenant des débats: Shakespeare en est-il bien l'auteur? Des milliers de livres et d'articles ont été consacrés à ce sujet. Deux chercheurs, Ward E.Y. Elliott et Robert J. Valenza (1996), ont voulu tenter d'apporter des éléments de réponses. Ils ont fait passer une batterie de tests logiciels à des pièces et poèmes dont on est sûr qu'ils sont de Shakespeare. Suite à cette étape de validation des tests, ils en ont retenu 51 qu'ils ont fait passer aux textes dont la paternité est contestée. Les analyses statistiques portaient sur des aspects comme les mots, les contractions, certains modèles de phrases, les préfixes et suffixes, etc. Ces tests ont cependant été critiqués par Donald W. Foster qui affirme que plusieurs sont imparfaits (1996).

Parmi les oeuvre attribuées à Shakespeare et dont la paternité fait l'objet d'un débat, il y a les trois pièces de théâtre *The Two Noble Kinsmen, The Double Falsehood* et *The London Prodigal*. Certains croient qu'elles ont été coécrites par Shakespeare et John Fletcher, alors que d'autres contestent cette hypothèse. Deux spécialistes, David Lowe et Robert Matthews (1995), ont utilisé le "Radial Basis Function Network" (RBF), une méthode du domaine de ce qu'on appelle en anglais le "neural network", pour accomplir la tâche stylométrique de déterminer l'auteur ou les auteurs de ce texte. Le RBF est une méthode assez complexe que Lowe et Matthews décrivent ainsi:
although the original motivation of this particular network structure was in terms of functional approximation techniques, the network may be derived on the basis of statistical pattern processing theory, regression and regularisation, biological pattern formation, mapping in the presence of noisy data etc. (1995, p.450). Cette méthode leur a permis de comparer le vocabulaire des pièces mentionnées plus haut à celui d'un corpus de textes dont la paternité est indubitable (étant attribuée soit à Fletcher soit à Shakespeare). Les analyses qu'ils ont menées les ont amenés à dire que *The Double Falsehood* et *The London Prodigal* devraient être prioritairement attribuées à Fletcher. Le cas de *The Two Noble Kinsmen* est différent. Cette pièce a longtemps été considérée comme une véritable collaboration entre les deux auteurs. Selon l'étude de Lowe et Matthews, chaque acte de la pièce porte les marques des deux auteurs - et a donc pu avoir été écrit en collaboration - mais avec une prédominance tantôt de l'un tantôt de l'autre.

3.3 Distinguer les imitations des oeuvres authentiques

Comme on l'utilise pour déterminer la paternité d'un texte, l'analyse de textes littéraires assistée par ordinateur pourrait être utilisée pour distinguer une imitation d'une oeuvre authentique, pour identifier le plagiat (Johnson, 1996b). Sans ordinateur, cette tâche peut être assez difficile à accomplir, puisqu'une imitation se veut d'un style identique aux oeuvres authentiques d'un auteur donné.

Le pastiche est une oeuvre qui se veut une imitation du style d'un auteur donné. Il ne faut pas le confondre avec la parodie qui est plus caricaturale. Le pastiche est donc une forme d'imitation, mais non dissimulée; il ne s'agit pas de plagiat. Le roman policier, fantastique ou de science-fiction sont des

genres qui font assez fréquemment l'objet de pastiches. L'oeuvre de l'écrivain américain Raymond Chandler n'y a pas échappé.

Lee Sigelman et William Jacoby (1996) ont utilisé les outils de l'analyse statistique pour évaluer la distinction entre les pastiches de son style et ses oeuvres originales. Ils se sont donc basés sur des éléments stylistiques pour faire ressortir les faiblesses des imitateurs. Ils ont d'abord "mesuré" le style de Chandler pour ensuite le comparer aux styles des différents pastiches. Toutefois, ils n'ont pas voulu s'attarder aux petits détails, mais plutôt à ce qu'ils considèrent comme les principales caractéristiques du style de l'auteur. Ils ont utilisé l'ordinateur pour analyser quatre caractéristiques:

1. la simplicité du vocabulaire (mesures utilisées: le degré de lisibilité, i.e. le nombre de syllabes par mot et de mots par phrase; l'usage d'un vocabulaire "de base", i.e. selon une liste de 850 mots permettant d'exprimer toute pensée);
2. l'action (mesures: le ratio entre les adjectifs et les verbes; le ratio entre les mots reliés à la violence et la criminalité et ceux reliés à la contemplation et la réflexion);
3. les dialogues (mesures: la densité, i.e. le nombre de mots qui font partie de dialogues divisé par le nombre total de mots; la fréquence; la longueur);
4. le langage des personnages (mesures: la fréquence relative des mots d'argot (selon une liste prédéterminée); la fréquence des comparaisons; la fréquence relative des mots considérés comme vulgaires ou obscènes; la fréquence relative des conjonctions de coordination.

Ces analyses ont permis à Sigelman et Jacoby de remarquer une certaine constance dans le style de Chandler. Ils ont comparé leurs résultats avec les pastiches et ont ainsi pu voir les similitudes et différences (dont certaines sont presque systématiques).

3.4 L'étude des motifs rythmiques dans les vers

L'analyse par ordinateur des mots et des syllabes peut contribuer à l'étude du rythme dans les vers de poèmes ou de pièces de théâtre. C'est ce que Sharon Diane Nell (1993) a fait.

La plupart des textes de pièces de théâtre du XVIIe siècle français sont composés d'alexandrins - des vers de douze syllabes. Les alexandrins classiques comportent toujours une césure, c'est-à-dire un repos entre la sixième et la septième syllabe. Elle marque la cadence du vers en le séparant en deux hémistiches composés de six syllabes chacun.

Pour faire son analyse, Nell part de *Théorie du vers de Benoît* de Cornulier. Cornulier a remarqué que certains types de mots ou de syllabes ne paraissent pas de part et d'autre de la césure, c'est-à-dire en sixième ou septième position, dans certaines circonstances. Il les a regroupés selon cinq "critères" nommés *masculin*, *proclitique*, *enclitique*, *prépositionnel* et *féminin*[5]. Il ne peut pas y avoir d'accent sur ces mots ou syllabes. Or, comme en français l'accent est mis sur la dernière syllabe d'une phrase, ces mots ne peuvent pas être placés à la fin de l'une des hémistiches d'un alexandrin. Autrement dit, ils ne peuvent pas se trouver en sixième ni en douzième position. Dans la langue française, les mots ou

syllabes susceptibles d'être accentués sont les conjonctions, les adjectifs, verbes ou adverbes monosyllabes, et la dernière syllabe d'un mot qui en compte plus d'une. Ces observations amènent Nell à tirer cette conclusion: "in addition to the metrical division of the alexandrine line into two *hémistiches* of six syllables each, these two halves of the same line may be subject to further subdivision" (1993, p.187). Ce sont les divisions non métriques, c'est-à-dire les syllabes pouvant être accentuées et se trouvant ailleurs qu'en sixième ou douzième position, qui créent le rythme dans un alexandrin. Par opposition, les divisions métriques sont celles qui se trouvent en sixième ou douzième position.

Nell a utilisé la théorie des critères de Cornulier et la technologie informatique pour faire une analyse comparative des pièces *Polyeucte* de Pierre Corneille, *Phèdre* de Jean Racine et *Le Tartuffe* de Molière, toutes trois tirées du répertoire théâtral français du 17e siècle. L'objectif de son étude était formulé ainsi: *to determine if there were any constant qualities between the plays or if there seemed to be stylistic differences, indicated by a wide variation in the use of the internal rythmic patterns, for example, between the three playwrights, or if the occurrence of these patterns displayed similarities in all three works* (Nell, 1993, p.190).

Nell a utilisé le tableur *Excel* de Microsoft pour effectuer son analyse. Elle a créé quatre documents dans *Excel*: (1) un contenant les sections d'alexandrins (elle a choisi d'étudier les hémistiches séparément plutôt qu'en alexandrins) codifiés selon les critères de Cornulier; (2) un autre contenant les 42 combinaisons de critères possibles; (3) un modèle pour la compilation statistique; (4) une feuille de macros qui assurait l'automatisation et la communication entre les autres documents.

Nell a collecté quatre types d'informations: (1) la fréquence globale des IRP (pour *Internal rhythmic pattern*) dans les trois pièces; (2) les différents types d'IRP présents dans chaque scène; (3) les pourcentages globaux des types d'IRP dans les trois pièces; (4) le comportement des IRP hexasyllabes dans les trois pièces. Elle a pu voir, par exemple, que les motifs (IRP) les plus fréquents sont les mêmes dans les trois pièces. Elle a tracé des portraits graphiques de chaque pièce en calculant la fréquence de chaque longueur de motifs (monosyllabes, bisyllabes, etc.) dans chaque scène. La longueur d'un motif correspond à la distance en syllabes entre deux syllabes accentuées.

3.5 Marques d'un auteur dans l'évolution du langage

Selon Dennis Taylor (1993), l'ordinateur permet de répondre à la question: comment un auteur donné contribue-t-il à l'évolution du langage? Il est possible de comprendre l'influence d'un auteur en faisant la corrélation entre son vocabulaire et ses expressions, d'une part, et des dictionnaires informatisés, d'autre part, pour ensuite produire le profil et l'historique de l'entrée de certains mots dans la langue. De grands auteurs ont inventé de nouvelles façons de dire les choses, ont modifié le langage et certaines de ces inventions et modifications font maintenant partie intégrante du langage courant. Il peut s'agir de nouveaux mots ou encore de nouvelles façons d'employer d'anciens mots. Auparavant, il n'y avait pas d'outils pour faire des liens entre les oeuvres littéraires et l'état du langage. Cela était donc un problème pour les littéraires de trouver comment mesurer l'originalité et la créativité des grands auteurs et, par la même occasion, leur participation à l'évolution du langage.

Pour ce faire, il faut trouver une façon de mesurer "les moments-clés du langage littéraire, ces moments où le langage viole une norme et constitue une déviation ou, mieux, une variation et un développement" (traduction de l'auteure)[6] (Taylor, 1993, p.342). Il faut comparer l'oeuvre avec des dictionnaires informatisés qui lui sont contemporains (pour voir comment l'oeuvre se distingue des normes de l'époque à laquelle elle a été produite) et des dictionnaires plus "tardifs" (pour voir quels aspects du langage de l'oeuvre ont été incorporés au langage). Il faut aussi se servir de grammaires et autres sources également informatisées afin de faire ressortir les changements dans la façon d'employer les mots. "Our ultimate task is to computerize all dictionaries and all texts, and then conduct a study of what changes in the language correspond to what sources" (Taylor, 1993, p.343). À l'aide des dictionnaires et d'autres oeuvres antérieures au texte étudié, il est aussi possible de repérer les changements de collocations, c'est-à-dire qui concernent la position d'un mot par rapport à d'autres, la proximité des éléments entre eux. Selon Taylor, les nouveautés de collocation marquent les points de transition du langage, là où les changements s'opèrent. L'informatique peut aider à repérer ces lieux.

4. Complémentarité homme-machine

Les exemples d'applications de l'analyse de textes littéraires assistée par ordinateur exposés précédemment montrent que l'informatique permet au chercheur d'aller plus loin. Les logiciels font des tâches qui seraient longues et laborieuses, voire impossibles, pour l'humain seul. L'ordinateur est objectif. Il examine le texte entier sans que son attention ne soit davantage attirée sur un passage en particulier. Il n'est pas sujet aux distractions, pas plus qu'aux idées préconçues (Fortier, 1995). Une analyse statistique par ordinateur bien faite peut redonner des bases plus fermes à un débat qui, jusque-là, se perdait en conjectures (Burrows et Craig, 1994). Elle peut faire ressortir des aspects d'une oeuvre qui sont difficilement visibles à l'oeil nu. L'ordinateur est donc avant tout un accélérateur et un facilitateur (Olsen, 1993).

Toutefois, le plus évolué des logiciels ne peut évidemment pas, seul, produire une analyse significative d'un texte littéraire. Toute statistique a besoin d'une interprétation humaine pour prendre son sens. L'ordinateur fournit les données brutes qui seront ensuite soumises à la capacité d'analyse de l'expert (Fortier, 1995). L'intervention humaine est aussi nécessaire avant celle de la machine, ne serait-ce que pour numériser les oeuvres, dictionnaires et autres documents nécessaires à l'analyse. Après la numérisation, il faut aussi corriger les erreurs de transcriptions. Le chercheur doit aussi bien souvent préparer les données à être analysées pas l'ordinateur. Dans le cas des analyses de concepts décrites plus haut, par exemple, c'est au chercheur de déterminer les concepts, leur structure et leurs liens avec les différents mots.

Pour l'analyse des oeuvres de Jonathan Swift, Laffal a aussi traduit les mots des oeuvres qui étaient dans d'autres langues que l'anglais et il a remplacé les noms propres par *name* ou *place*, selon ce qu'ils désignaient. Les mots dont l'ortographe a changé depuis la rédaction de ces textes ont été réécrits selon l'ortographe moderne. De plus, le chercheur a dû intervenir durant l'analyse pour contrer les problèmes de polysémie. Pour ce faire, il a employé deux logiciels: *one reads the text to be analysed and marks all*

words wich have more than one meaning in the dictionary. A second program advances through the marked text, stopping at each marked word with a display of numbered dictionary choices. The human editor selects the proper meaning by keying the pertinent number (Laffal, 1995, p.342).

L'informatique peut être employée pour analyser des données tirées d'oeuvres littéraires plutôt que sur les textes eux-mêmes. Mais alors, il faut une implication humaine plus grande, il faut collecter les données, les organiser, les traiter, etc. avant l'intervention de la machine. Le travail du chercheur précédant l'analyse par ordinateur devient plus important encore. Beverley Ormerod, Jean-Marie Volet et Hélène Jaccomard (1995) se sont servis de logiciels informatiques pour étudier les personnages féminins dans la littérature africaine francophone. Il s'agissait en fait d'une comparaison entre les personnages féminins chez les auteurs masculins et chez les auteurs féminins. Les chercheurs s'attendaient à trouver les résultats suivants: (1) que les personnages masculins soient beaucoup plus nombreux, autant dans les oeuvres des auteurs féminins que dans celles d'auteurs masculins; (2) que, chez les auteurs féminins, il y aurait une discrimination positive en faveur des personnages féminins "in terms of female characters' mere presence in a novel and in terms of their power, attitude and importance" (1995, p.355).

Ils ont établi un corpus de dix romans écrits par des hommes et dix par des femmes. Les données soumises à l'examen de l'ordinateur étaient constituées d'une liste exhaustive des personnages de ces vingt romans auxquels on a accordé trois notes de 1 à 5 (selon des critères préétablis), l'une correspondant à leur importance, une autre à leur pouvoir et la dernière à leur attitude dans le roman. C'est à ce niveau que se situait la plus grande part de l'intervention humaine avant le traitement des données par ordinateur. Les chercheurs ont ensuite fait ressortir la différence entre les textes d'auteurs masculins et ceux d'auteurs féminins.

Quant aux profils, graphiques, tableaux, etc. résultants des diverses analyses, ils ne sont pas eux-mêmes des interprétations des oeuvres littéraires. Ils servent plutôt de base aux études menées par les chercheurs. Dans tous les exemples mentionnés jusqu'ici, les chercheurs ont eu à interpréter les résultats fournis par les différents logiciels utilisés.

5. Critiques et limites de l'analyse de textes par ordinateur

L'analyse de textes littéraires assistée par ordinateur a bien sûr ses détracteurs. Les gens qui pratiquent ce type d'analyse ne sont généralement pas ceux que l'on considère comme des spécialistes de la littérature. Il y a bien, parmi les experts du domaine, quelques professeurs de littérature à l'université. Mais on trouve surtout des "non littéraires": des spécialistes en mathématiques, en physique, en informatique, en psychologie, voire en science politique, etc. Les littéraires qui n'utilisent pas l'informatique dans leurs travaux, eux, mettent souvent en doute cette façon d'étudier la littérature et ne semblent pas s'y intéresser outre mesure. *The conclusions of most individual CARL (computer-assisted research on literature (projects have simply been too trivial or too obvious to attract attention. A second reason put forward for the marginalization of CARL is the rebarbative presentation of its research* (Finch, 1995, p.511).

Alison M. Finch parle aussi d'une "mythologisation" des méthodes d'analyse statistiques par les experts eux-mêmes. *Some surprising figures of speech infiltrate the critical diction of many CARL analysts - figures of speech that tend to mythologize their own entreprises. (...) It may have stopped CARL experts evaluating properly the results of their own research, and it cannot but be off-putting the non-CARL critics they are trying to win over* (1995, p.512).

Pour certains experts de l'analyse de textes littéraires par ordinateur, Olsen en tête (1993), l'informatique a bien des choses à offrir à la littérature, mais elle est souvent mal utilisée et n'a pas l'impact qu'elle devrait avoir sur le champ des études littéraires. Olsen croit qu'il est nécessaire de réévaluer le rôle de l'informatique dans l'analyse de la littérature et d'aller dans de nouvelles directions. Il cite Rosanne Potter qui affirme que les spécialistes utilisant l'informatique en littérature ont trop souvent tendance à rendre leurs rapports très "techniques", ce qui n'aide pas à s'adjoindre un lectorat de littéraires. Potter remarque également que ce type d'études se limitent la plupart du temps à un petit nombre d'oeuvres (Olsen, 1993, citant Potter, 1989).

Selon Mark Olsen, les erreurs commises par les experts de l'analyse de textes par ordinateur ne sont pas d'ordre technique, mais plutôt théoriques et méthodologiques (1993). Il soulève aussi que les analyses de textes littéraires par ordinateur sont généralement faites sur la base d'éléments simples comme la longueur des mots et les ratios "type-token", alors que ces mesures donnent des résultats peu satisfaisants en eux-mêmes, selon lui. C'est aussi l'avis de David D. Miall: "The frequencies of words, collocations, or particular stylistic features, tell us rather little about the literary qualities of a text, since these aspects of a text find their meaning only within the larger and constantly shifting context constituted by the reading process" (1995, p.202).

Le problème, c'est qu'il n'est pas évident de transformer des qualités textuelles en statistiques. Paul A. Fortier soulève que, bien que les textes soient composés de mots, leurs effets sont produits par des phénomènes d'un ordre supérieur et plus complexe (1995). De plus, aucun algorithme informatique connu ne peut saisir si un mot donné est employé au sens figuré ou littéraire (Miall, 1995). Pour ce faire, il faudrait d'abord que le chercheur fasse un travail d'encodage, tâche énorme et fastidieuse. Selon Miall, prédire une nouvelle ère où un ordinateur serait capable de comprendre une oeuvre littéraire, c'est sous-estimer la complexité du processus de lecture d'un texte. Lire un poème ou un roman est un processus de transformation probablement encore plus complexe que dans le cas d'autres types de textes.

Par contre, il est convenu que les données issues d'analyses par ordinateur puissent être utiles pour des analyses plus poussées. Mark Oslen ajoute: *It would seem that the approach of using computers to analyze the linguistic and symbolic environment - the collective and social elements of language - in order to understand individual texts and rhetorical stances, suggests that computer analysis of text should play a central and well defined role in our understanding of text* (1993, p.313).

De plus, il est évident que, si certains aspects des textes littéraires sont quantifiables, d'autres ne pourront jamais l'être (Burrows et Craig, 1994).

Étienne Brunet, de son côté, soulève les dangers de l'obstination statistique. Lorsqu'un chercheur veut, par exemple, déterminer la paternité d'un texte, il avance d'abord une hypothèse puis soumet le texte à des tests. S'il n'arrive pas aux résultats escomptés et qu'il ne veut pas que ses efforts soient vains, il risque de s'acharner et d'interpréter les résultats de façon à leur faire dire ce qui lui convient le mieux. C'est que, pour Brunet, différentes personnes peuvent faire dire différentes choses aux statistiques. On a tendance à donner aux chiffres une supériorité presque divine sur les mots parce qu'ils semblent absolus; "but this apparent incontrovertibility, however impressive, often conceals relative and contingent procedures that have nothing essential about them", écrit-il (Brunet, 1991, p.70).

Conclusion

L'ordinateur permet aux chercheurs qui s'intéressent à la littérature d'ajouter un aspect quantitatif à leurs analyses. À l'aide de différents logiciels, les spécialistes peuvent obtenir des tableaux, des graphiques, des statistiques sur les mots qui composent les oeuvres qu'ils étudient, de même que sur les syllabes, les signes de ponctuation, le vocabulaire, les thèmes et champs lexicaux, etc. Ces résultats peuvent ensuite être utilisés pour comparer des auteurs ou des textes entre eux. Ils peuvent notamment aider le chercheur à déterminer la paternité d'un texte, à distinguer les imitations des oeuvres authentiques, à comprendre les motifs rythmiques dans des vers ou encore à saisir comment un auteur donné a contribué à l'évolution du langage. Mais l'informatique n'est qu'un outil, il fournit une assistance au chercheur qui doit intervenir *avant* et *après* l'analyse automatisée. L'analyse de textes littéraires assistée par ordinateur ne fait pas l'unanimité et les experts doivent composer avec les limites de cette méthode.

L'analyse de textes littéraires par ordinateur est marginalisée par les littéraires. La plupart d'entre eux ne croient pas que l'informatique puisse leur apporter une aide réelle dans leurs travaux et ne semblent pas avoir la curiosité de découvrir les possibilités de cet outil. Il faut dire qu'une bonne partie des écrits dans le domaine de l'analyse de textes par ordinateur sont assez techniques et quelquefois rébarbatifs pour qui n'est pas très familier avec les statistiques et l'informatique. D'un autre côté, les experts en analyse de textes littéraires assistée par ordinateur ne font pas toujours un usage très pertinent des outils informatiques. Bien des études se limitent à l'analyse d'aspects très simples comme la longueur des mots et des phrases, la fréquence de certains mots, etc. En eux-mêmes, les résultats de telles analyses ne sont pas très intéressants d'un point de vue strictement littéraire. Par contre, ils peuvent être pratiques lorsqu'utilisés pour fins de comparaison; à la condition, bien sûr, que la comparaison soit pertinente, que son auteur ait un objectif précis.

Il est certainement souhaitable que littéraires et experts en analyse de textes par ordinateur - qu'ils soient informaticiens, mathématiciens ou autres - s'associent, qu'ils joignent leurs connaissances et leurs spécialités pour arriver à faire un usage pertinent des outils informatiques en études littéraires. Il faut faire en sorte d'attirer les spécialistes de la littérature vers ce domaine et mettre à leur disposition des outils de qualité qui leur apporteront un soutien concret dans leurs travaux.

Toutefois, tant que le processus cognitif de la lecture ne sera pas mieux compris, les chercheurs ne

pourront faire qu'un usage assez limité de l'informatique. Et ce n'est certainement pas demain le jour où l'ordinateur pourra saisir le sens de vers comme ceux-ci, tirés du poème "C'était un bon copain" de Robert Desnos:

Il avait le coeur sur la main
Et la cervelle dans la lune
(...)
Il avait l'estomac dans les talons
(...)
Il avait la tête à l'envers
Et le feu là où vous pensez
(...)
Quand il prenait ses jambes à son cou
Il mettait son nez partout
(...)
Il avait une dent contre Étienne
(...)
Il n'avait pas sa langue dans la poche
(...) (1930, p.86).

Le cerveau humain est en mesure de saisir le sens caché derrière des expressions, de faire des associations d'idées, de créer et d'interpréter des métaphores, et même de trouver des significations nouvelles aux mots qui composent le vocabulaire courant. Jamais un ordinateur, même armé du logiciel le plus évolué qui soit, ne pourra rivaliser d'intelligence et de perspicacité avec un auteur ou un lecteur humain.

Bibliographie

Bailey, R. W. "Authorship Attribution in a Forenstic Setting". *Advances in Computer-aided Literary and Linguistic Research*, D.E Ager et al (éd.), Birmingham: AMLC, 1979.

Brunet, Étienne. "What Do Statistics Tell Us?". *Research in Humanities Computing*, n°1 (1991): 70-92.

Burrows, J. F. ; Craig, D. H. "Lyrical Drama and the "Turbid Mountebanks": Styles of Dialogue in Romantic and Renaissance Tragedy". *Computers and the Humanities* 28 (1994): 63-86.

Desnos, Robert. *Corps et biens*, Paris: Gallimard, 1930.

Elliott, Ward E. Y. ; Valenza, Robert J. "And Then There Were None: Winnowing the Shakespeare Claimants". *Computers and the Humanities* 30, n°3 (1996): 191-245.

Finch, Alison M. "The Imagery of a Myth: Computer-Assisted Research on Literature". *Style* 29, n°4

(1995): 511-521.

Fortier, Paul A. "Categories, Theory, and Words in Literary Texts". *Research in Humanities Computing*, n°5 (1995): 91-109.

Foster, Donald W. "Response to Elliot and Valenza, "And Then Were None"". *Computers and the Humanities* 30, n°3 (1996): 247-255.

Holmes, David, I. "Autorship Attribution". *Computers and the Humanities* 28 (1994): 87-106.

Johnson, Eric. "The Kinds of Words used in the Novels of Jane Austen, Charles Dickens, and James Janke". *Text Technology* 6, n°2 (été 1996a): 91-96.

Johnson, Eric. "The World Wide Web, Computers, and Teaching Literature". 1996b. http://www.triton.dsu.edu/tlwc/articles/webprof.html

Laffal, Julius. "A Concept Analysis of Jonathan Swift's A Tale of a Tub and Gulliver's Travels". *Computers and the Humanities* 29 (1995): 339-361.

Lowe, David ; Matthews, Robert. "Shakespeare Vs. Fletcher: A Stylometric Analysis by Radial Basis Functions". *Computers and the Humanities* 29 (1995): 449-461.

Miall, David D. "Representing and Interpreting Literature by Computer". *The Yearbook of English Studies: Non-Standard Englishes and the New Media* Special Number 25 (1995): 99-212.

Nell, Sharon Diane. "Toward a Theory of Rythm in French Poetry: Computer Assisted Recognition of Rythmic Groups in Traditional Isometrical Alexandrines". *Computers and the Humanities* 27 (1993): 185-223.

Olsen, Mark. "Signs, Symbols and Discourses: A New Direction for Computer-aided Literature Studies". *Computers and the Humanities* 27, n°5-6 (1993): 309-314. http://tuna.uchicago.edu/homes/mark/Signs.html

Ormerod, Beverly ; Volet, Jean-Marc ; Jaccomard, Hélène. "The Female Voice and Traditional Discourse Biases: The Case of Francophone African Literature". *Computers and the Humanities* 28 (1995): 353-367.

Potter, Rosane. *Literary Computing and Literary Criticism: Theoretical and Practical essays on Theme and Rhetoric*. Philadelphie, 1989.

Sigelman, Lee. "By Their (New) Words Shall Ye Know Them: Edith Wharton, Marion Mainwaring, and

The Buccaneers". *Computers and the Humanities* 29 (1995) 271-283.

Sigelman, Lee ; Jacoby, William. "The Not-So-Simple Art of Imitation: Pastiche, Literary Style, and Raymond Chandler". *Computers and the Humanities* 30, n°1 (1996): 11-28.

Taylor, Dennis. "Literary Texts and the State of the Language: The Role of the Computer". *Computers and the Humanities* 27 (1993): 341-347.

Notes

[1] Ces logiciels sont souvent faits sur mesure pour l'analyse particulière que le chercheur désire faire. Il y a donc presque autant de logiciels que de types d'analyses.

[2] Toutes les analyses de Brunet ont été faites grâce à la base de données FRANTEXT contenant 3,000 textes littéraires intégraux et qui permet d'exécuter simplement divers types d'analyses statistiques.

[3] David I. Holmes le décrit ainsi: *"If N = the number of units (word occurences) wich form the sample text (tokens), and V = the number of lexical units wich form the vocabulary in the sample (types), then the type-token ratio is defined by R = V/N"* (Holmes, 1994, p.92).

[4] *"I have taken as my guide the more obvious characters of the ideas for wich expressions were to be tabulated, arranging them under such classes and categories as reflection and experience has taught me would conduct the inquirer most readily and quickly to the object of his search"* (Roget cité par Laffal, 1995, p.342).

[5]

Masculin:
: *"Se dit d'un mot qui, s'appuyant sur le mot suivant avec lequel il forme une unité phonétique, est dépourvu d'accent tonique"* (Le Nouveau petit Robert 1), ie les articles, les pronoms personnels relatifs, les prépositions monosyllabiques;

Proclitique:
: Toutes les syllabes sauf la dernière d'un mot multisyllabe, à moins qu'une ou plusieurs de ces syllabes soient des "e" muets (Nell, 1993);

Enclitique:
: *"Mot qui prend appui sur le mot précédent et forme avec lui une seule unité accentuelle"* (Le Nouveau petit Robert 1), ex: "ce" dans "qu'est-ce?";

Prépositionnel:
: Une préposition est un *"mot grammatical, invariable, introduisant un complément (d'un substantif, d'un verbe, d'un adjectif, d'un adverbe) en marquant le rapport qui unit ce complément au mot complété"* (Le Nouveau petit Robert 1), ex: à, après, avec, jusque, outre, par,

sauf, etc.;

Féminin:

"e" muet se trouvant entre deux consonnes (Nell, 1993).

[6] *"The key moments in literary language, those moments where the language violates a standard norm and constitues a deviation or, better, variation and development. Some of these deviations eventually become very influential and indeed become part of the standard language wich is then again subject to variation"* (Taylor, 1993, p.342).

Survol du monde de l'indexation des images

par

Kumiko Vézina

Cursus vol. 4 no 1 (Automne 1998)

Cursus est le périodique électronique étudiant de l'École de bibliothéconomie et des sciences de l'information (EBSI) de l'Université de Montréal. Ce périodique diffuse des textes produits dans le cadre des cours de l'EBSI.

ISSN 1201-7302

C. élec. : cursus@ere.umontreal.ca
URL : http://www.fas.umontreal.ca/ebsi/cursus/

Droits d'auteur

L'auteure

Ayant terminé sa maîtrise à l'EBSI en juillet 1995 avec un stage à New York, elle fait partie de la première vague d'étudiants au Doctorat en sciences de l'information à l'EBSI dès septembre de la même année. Ses intérêts portent principalement sur les systèmes d'indexation d'images fixes ainsi que les besoins d'usagers de collections visuelles.

Pour joindre l'auteur : vezinak@ere.umontreal.ca

Table des matières

Résumé

Le domaine de l'indexation des images est encore relativement jeune si on le compare à celui de l'indexation textuelle mais il est en pleine effervescence. À partir de la fin des années 70, de nombreuses institutions voulurent mettre en valeur leurs collections visuelles. Ils mirent donc sur pied des projets de thésaurus et de classification de grande envergure destinés à mieux organiser leurs images. Toutefois, ces projets soulevèrent de nombreux problèmes vis-à-vis l'indexation du matériel visuel et moussèrent l'intérêt académique dans le domaine de la bibliothéconomie et des sciences de l'information. Ce survol présente donc les premiers grands projets d'organisation d'images, les systèmes actuels d'indexation des images utilisant de nouvelles techniques comme thésaurus visuels et langages de description pictorielle, les problèmes reliés à ce domaine et les recherches à faire.

Introduction

Au temps de la préhistoire une des premières formes de communication étaient les manifestations soit gravées soit peinturées sur les parois de grottes. À prime abord abstraites, i.e. non-figuratives, ces manifestations ont pris la forme de représentations animales ou végétales. De nombreux exemples ont été détectés en Europe dont les plus connus sont sûrement la grotte de Lascaux en France et celle d'Altamira en Espagne. Ailleurs dans le monde, les aborigènes d'Australie sculptaient des formes animales et végétales afin, entre autres, de délimiter leur territoire, d'indiquer les endroits propices à la chasse ou encore les utilisaient comme points de repère pour se diriger dans la brousse.

À mesure que l'Homme évoluait, les nombreuses civilisations de part le monde développèrent des systèmes de communication plus sophistiqués grâce à des symboles qu'ils inventaient. Dans certaines cultures, l'écriture avait des racines dans le visuel (pictogrammes, idéogrammes, etc.) alors que d'autres développèrent des symboles arbitraires comme l'alphabet hébreu. Jusqu'à nos jours, le texte était à la base de l'éducation et de l'échange d'information.

Maintenant, en grande partie due aux avancements technologiques, nous entrons dans une nouvelle ère où les images deviennent prépondérantes. Les images animées ou en mouvement sont générées en très grande quantité et à une vitesse fulgurante à la fois par la télévision mais aussi, cependant en moins grand nombre, par la création de films. De plus, la production de processus commerciaux, industriels et scientifiques qui génèrent des témoignages visuels d'opérations continuelles comme la banque de données d'images de satellites de la NASA dépasse déjà les téra-octets d'espace-disque sans qu'il ne semble y avoir de relâche. Les images fixes, quant à elles, créées par le monde de l'édition, sont produites à une échelle bien plus réduite et les images de représentations artistiques encore beaucoup moins. Il en découle qu'au niveau de l'organisation, le stockage et le repérage d'images fixes, des progrès ont pu être effectués alors que dans le domaine des images animées, la situation est tout autre. En effet, l'expérience encore toute jeune des spécialistes de l'information dans le domaine de l'image animée jumelée au fait que ce type d'information augmente de façon exponentielle explique le manque de systèmes adéquats (Enser, 1995).

Un survol du monde de l'indexation des images permettra de voir les grands projets d'indexation d'images, les problèmes reliés à ce domaine ainsi que les recherches à faire.

1. Premiers grands projets d'organisation d'images

En Amérique du Nord, vers la fin des années 70 et le début des années 80, un effort se faisait sentir au niveau de la construction de thésaurus et de liste d'autorité pour classifier les images visuelles (Sunderland, 1982). Thomas Ohlgren (1982), spécialiste d'iconographie aux États-Unis, nous fait part de quatre projets inter-institutionnels qui s'adressent aux besoins thématiques du grand public en donnant accès à leurs collections visuelles :

1.1 Image Access Society of North America

Société fondée en août 1979 par des archivistes, conservateurs, historiens d'art, muséologues, bibliothécaires de diapositives et de photographies et des spécialistes de l'information lors du premier International Conference on Computers and the Humanities, au Collège de Dartmouth à Hanover au New Hampshire. L'association a profité de l'événement pour définir ses buts :

a) Recueillir les schémas de classifications en utilisation et faire une compilation de ce matériel ;
b) Établir une nomenclature pour décrire et évaluer les nombreux schèmes iconographiques ;
c) Recueillir de l'information portant sur les différentes approches à l'accès par sujet en Amérique du Nord auprès des membres de la Société en question ;
d) Commanditer et encourager des projets pilotes dans les domaines de construction de Thésaurus et d'indexation par sujet ;
e) Organiser une conférence pour discuter de la réalisation des buts mentionnés.

Une deuxième rencontre en août 1980 permet à l'association de se poser une question cruciale puis de tenter d'y trouver une réponse : Quels seraient les standards minimaux pour un système pratique d'accès par sujet ? Voici les conclusions qui ont été tirées de cette réunion.

- a) Le système devra être universel et interdisciplinaire;
- b) Il devrait être appliqué de façon consistante d'une institution à une autre de même que d'un département à un autre à l'intérieur d'une même institution. Il devrait être intelligible pour tout indexeur et utile à tous les indexeurs d'un même département ;
- c) Il doit être applicable à divers niveau de complexité
- d) Il devrait être portable ou possible de le distribuer ;
- e) Le vocabulaire d'indexation utilisé devrait susciter les mêmes images visuelles peu importe le bagage de connaissances ("background ") des personnes ;
- f) Il devrait être possible pour quiconque d'avoir accès au système et d'obtenir des réponses comparables à des questions comparables ;
- g) Il devrait être laissé ouvert (" open-ended "), c'est-à-dire permettre la modification de termes existants ou l'ajout de nouveaux termes.

1.2. List of Subject Headings pour la Library of Congress Prints and Photographs Division

Elisabeth W. Betz, spécialiste en catalogage, a complété au début des années 80 une liste alphabétique de 3500 vedettes-matières avec références dont le but est d'améliorer l'accès-sujet aux 10 millions de gravures et de photographies de cette division. Les 65 000 notices de l'index-sujet de cette division furent comparées avec les vedettes-matières au Library of Congress Subject Headings et ont été

modifiées pour s'appliquer à des images visuelles.

1.3. Thésaurus de la Division de l'iconographie (Archives publiques du Canada)

Il s'agit de la création d'un thésaurus informatisé de termes iconographiques vers la fin des années 70. Cette entreprise était menée par Raymond Vézina et Douglas Schoenherr avec la collaboration de Denis Castonguay. L'approche scientifique de cette opération fait en sorte qu'il s'agit d'un des projets d'indexation-sujet les plus importants au monde (Ohlgren, 1982). Le thésaurus a fourni un vocabulaire contrôlé d'indexation pour les indexeurs de quelque 150 musées canadiens fédéraux, provinciaux et municipaux (Vézina, 1998). En plus, l'inventaire informatisé a permis une consultation en-ligne à environ 100 000 œuvres d'art de la Division de l'iconographie, un des plus grands dépositaires d'iconographie récente canadienne au Canada regroupant notamment des aquarelles, des gravures, des eaux-fortes ainsi que des toiles. Voici les grandes catégories de leur thésaurus : architecture, costume, activités, artefacts, flore, faune, insignes, paysages, gens et transport. Ce système de classification est constitué d'un vocabulaire structuré utilisant trois niveaux de spécificités :

1. Liste contrôlée de catégories générales de classification (celles déjà mentionnées) ;
2. Liste contrôlée de termes de classification consistant en des subdivisions bien définies des catégories générales. Par exemple, pour la catégorie générale " architecture ", une des subdivisions est " architecture domestique " ;
3. Une liste ouverte où l'on peut ajouter des termes d'indexation spécifiques (ces termes ont des références vers des termes génériques, spécifiques ou associés). Encore une fois, si l'on reprend l'exemple de la catégorie générale" architecture " et de sa subdivision " architecture domestique ", la liste ouverte de termes spécifiques comporterait les termes suivants, entre autres, " fermes ", " maisons historiques ", " igloos ", " wigwams", etc.

1.4. Art and Architecture Thesaurus (AAT)

En 1979, Pat Molholt, Toni Petersen et Dora Crouch annoncent le projet d'un thésaurus dans le domaine de l'architecture qui sera à la fois une liste d'autorité pour la sélection des descripteurs par les indexeurs ainsi qu'un guide pointant vers l'information ou les objets les plus importants pour les besoins des usagers. Ce thésaurus est en fait une liste de termes organisés de façon hiérarchique : attributs physiques, styles et périodes, agents, activités, matériaux et objets. Nous savons maintenant que le résultat final de ce projet a été publié en 1990 et qu'il y a déjà un bon nombre d'articles qui ont été écrit sur ce thésaurus. En effet, le Art and Architecture Thesaurus a été décrit, critiqué et même comparé à d'autres systèmes d'indexation comme ICONCLASS.

On peut également mentionner d'autres projets d'organisation de collections visuelles comme celui de Eleanor Fink qui a développé des vedettes-matières pour le Musée d'Art américain du Smithsonian et celui de Elizabeth Glass qui a développé son système de classification par sujet pour le Musée Victoria and Albert à Londres.

2. Systèmes d'indexation des images

En ce moment, il existe plusieurs systèmes dont le but est d'indexer des images afin de les stocker pour un jour être en mesure de les repérer. Alors que chaque système adopte une approche différente, ils tendent tous vers un même objectif, i.e. faire en sorte que l'indexation soit la meilleure possible afin que le repérage soit des plus facile et efficace. Voici la description de plusieurs projets en cours de nos jours mentionnés dans un article de Graeme Baxter et Douglas Anderson (1995).

2.1. Méthodes d'indexation basées sur thésaurus ou système de classification

Les méthodes d'indexation basées sur thésaurus ou système de classification utilisent du texte pour décrire les images. Les thésaurus sont des listes hiérarchiques de termes, en général par ordre alphabétique, avec pour chacun des termes des renvois vers des termes plus génériques, plus spécifiques ou des termes associés. Quant aux systèmes de classification, ce sont de grandes catégories qui permettent de faire des regroupements thématiques.

Déjà mentionné, le projet Art and Architecture Thesaurus (AAT), publié en 1990 et se basant sur la structure des vedettes-matières médicales de la National Library of Medicine, est un thésaurus de plus de 50 000 termes (Petersen, 1990). Toutefois, il semblerait qu'un indexeur, même à la suite d'une formation, prend en moyenne quarante minutes pour indexer une diapositive à l'aide de ce système (Keefe, 1990).

Un autre mode de classification, ICONCLASS, représente 17 volumes de codes arrangés de façon hiérarchique associés à des descriptions textuelles. Il semblerait que ce système est très satisfaisant pour décrire de l'iconographie traditionnelle en histoire de l'art (peintures, sculptures, etc.) mais peu satisfaisant pour les objets communs comme meubles, bijoux, et autres. De plus, le processus d'indexation doit être effectué par des spécialistes en histoire de l'art et demeure une opération demandant en moyenne trente minutes par item (Sherman, 1987).

Mis de côté de nos jours semble-t-il, le système de classification TELCLASS a été créé pour le BBC-TV en 1979 pour indexer leur matériel d'images en mouvement. Les codes alphanumériques ainsi que les termes associés utilisés sont regroupés sous six grandes catégories : verbal, schématique, actualité, simulation, technique et formel (Evans, 1987). TELCLASS fut aussi adopté par d'autres institutions notamment par la maison d'édition McGraw-Hill pour leur encyclopédie multimédia portant sur la biologie des mammifères (Baxter et Anderson, 1995). Mais depuis, aucune littérature ne porte sur ce système.

2.2. Thésaurus visuels et systèmes hybrides associés

Une nouvelle vague de recherche explore en ce moment la création de systèmes pour lesquels le texte ne servira plus à indexer ou à repérer les images mais sera remplacé par des éléments visuels. Toutefois, les expériences jusqu'ici utilisent toujours des mots. Ces systèmes se basent sur le concept du " browsing " qui utilise le processus de sélection le plus performant qui soit, i.e. la combinaison de l'œil et du cerveau. Ces nouveaux systèmes sont efficaces pour de petites collections où les images sont relativement simples. Pour une plus grande collection, il faudrait que l'indexation soit assez exhaustive pour que le taux de rappel ne soit pas trop élevé.

Il existe un projet à la NASA au Johnson Space Centre où des images concernant l'espace sont regroupées dans une banque de données qui ne fonctionne pas par logique booléenne mais plutôt à l'aide d'algorithmes se basant sur des statistiques. Le processus de repérage affiche ainsi les données par ordre d'importance par rapport à la requête de l'usager. De plus, le repérage affiche à la fois le terme du thésaurus choisi par l'usager avec les images qui l'accompagne avec en plus les termes génériques, spécifiques et associés avec leurs images liées respectives, tout ceci par ordre d'importance (Seloff, 1990).

À l'Université de Syracuse, un autre projet se dédie à l'identification de feuilles de plantes mais pourrait éventuellement être utilisé pour identifier des fossiles, des os, de la monnaie, etc. Il s'agit de choisir à l'écran, parmi les exemples présentés, une forme similaire à celle de la feuille que le chercheur veut identifier. Par la suite, le chercheur peut manipuler la forme grâce à des " points de contrôle " pour qu'elle corresponde le plus exactement à ce qu'il recherche. Le système retrouvera ensuite, dans la banque d'images, celles qui ressemblent le plus à la feuille à identifier. Ici, l'utilisation de mots pour effectuer une recherche a été complètement mis de côté (Hogan et al., 1991).

Au Danemark, le Riso National Laboratory a créé une base de données, The Book House, qui permet aux usagers de repérer des livres grâce à des icônes (Pejterson, 1993). Par exemple, l'icône du crime représente un homme non-rasé qui se trouve derrière des barreaux. Chacun des icônes est relié à des termes provenant d'un index permettant ainsi à l'usager de construire une requête complexe à mesure qu'il choisi des icônes. Par la suite, la base de données sera en mesure d'identifier une série de livres correspondant aux besoins de l'usager. Il est intéressant de noter que pour déterminer l'image des icônes, il a fallu effectuer une analyse cognitive des besoins et des requêtes des usagers potentiels. De plus, ils se sont rendus compte que la perception et la compréhension d'images étaient bien différentes d'un adulte à un enfant ce qui les a amenés à créer deux séries d'icônes, une pour adultes et l'autre pour enfants, pour des sujets identiques. Il serait intéressant de voir l'efficacité d'un tel système pour le repérage d'images.

À l'Université de Californie à Berkeley, le système qui utilise le logiciel IMAGEQUERY permet aux usagers d'effectuer en premier lieu une recherche textuelle. Le système affiche par la suite une mosaïque d'imagettes que l'usager peut regarder pour ensuite choisir celles qui concorderaient le plus avec ses besoins. Il s'agit d'un modèle hybride qui tente de s'éloigner de l'indexation textuelle des images en utilisant à la fois du texte et du visuel pour effectuer le repérage du matériel visuel (Besser, 1990).

2.3. Langages de description pictorielle

Ces systèmes utilisent un langage spécial (algorithmes en général) où la description de l'image peut être codée et lue par une machine. De telles techniques demandent un indexeur très spécialisé et ne sont efficaces que dans le cas de petites collections regroupant des images simples (Baxter et Anderson, 1995). Voici trois exemples de systèmes de description pictorielle.

Le chercheur Leung et ses associés ont établi une méthode qui se base en fait sur la narratologie de l'image pour la décrire. Le projet Entité-Attribut-Relation utilise des " phrases " avec nom, adjectif et verbe pour décrire ce qui se passe sur l'image. Par exemple, une requête pourrait ressembler à : assis, (chat noir, chaise) (Leung et al., 1992).

Un groupe de chercheurs-physiciens à Milan provenant d'institutions telles que le CNR-SIAM et l'Université de Milan, se sont penchés sur la description de la silhouette d'une image grâce aux coordonnées de points spécifiques. Le code ainsi créé ressemblerait à ceci : " ARM FROM (9,19) TO (9,23) WITH ENDPOINT IN (3,25) ". Jusqu'ici, le prototype a servi à décrire des images d'ellipses en astronomie (Bordogna et al., 1990).

À Taiwan, des chercheurs ont développé des chaînes de caractères qui décrivent la relation spatiale entre les différents objets dans une image (Chang et Wu, 1992). En guise d'exemple : T = *(A,B,7), (B,C,8), (C,D,1), (D,E,4)} où 1 = nord de l'objet de référence, 7 = est de l'objet de référence, 8 = nord-est de l'objet de référence et où A, B, C et D sont des objets.

2.4. L'indexation et repérage d'images par contenu

Ces méthodes ne se basent par sur de la description textuelle mais analysent plutôt le contenu d'une image, en se basant sur des analyses mathématiques de déploiements de pixels, pour ensuite les jumeler à la requête de l'usager. La requête en question peut avoir deux formes, soit l'usager choisi une image avec une structure semblable à ce qu'il recherche ou une question décrivant la structure de l'image voulue (" pattern-matching "). Ces systèmes ne sont valides que pour des collections où les images sont simples. Pour le moment, le contenu devra encore être indexé avec des mots (Cawkell, 1992).

En 1987 à l'Université de Victoria en Colombie-Britannique, deux chercheurs ont développé une méthode probabiliste de question-réponse entre ordinateur et usager pour créer un appariement de traits caractéristiques des images voulues. Donc à mesure que l'usager répond aux questions posées par l'ordinateur, ce dernier accumule les paramètres qui lui permettront de restreindre la quantité d'images qu'il va retourner à l'usager. Même si ce système très intéressant et convivial a beaucoup plu aux usagers, il ne semble pas y avoir eu de suite à ce projet (Lee et MacGregor, 1987).

Au Royaume-Uni, deux projets se sont démarqués. Celui de Rickman et Stonham à l'Université de Brunel est une base de données faciales avec 500 visages qui utilise un processus de " pattern-matching ", i.e. reconnaissance de formes (Cawkell, 1992) et le projet Shape Analysis For Automatic Retrieval of Images (SAFARI), développé par Eakins à l'Université de Northumbria au milieu des années 80, qui est un système de commandes qui indiquent les coordonnées, la direction, la longueur, l'arc d'un angle et bien d'autres aspects de formes se trouvant dans une image afin de construire une requête de la forme recherchée (Eakins, 1993).

Finalement, le système semi-automatique Query By Image Content (QBIC) conçu par Niblack et associés, pour le groupe de recherche IBM, prend en considération les couleurs, textures et formes d'une image qu'il transforme en formules algébriques. Lors d'une recherche, ce système permet de réduire le taux de rappel dans une grande base de données pour que l'usager puisse ensuite effectuer du " browsing " (Niblack et al., 1993).

3. Problèmes liés à l'indexation des images

À prime abord, il serait important de faire la différence entre les notions de catalogage, classification et indexation. Le problème fut posé à Michèle Hudon, professeur à l'École de bibliothéconomie et des sciences de l'information de l'Université de Montréal, qui à l'aide d'exemples réussit à résumer l'essentiel de ces opérations (Hudon, 1998).

Le catalogage consiste à décrire physiquement un document, quel que soit son format, permettant d'une part de l'identifier de façon unique et d'autre part de le repérer par le biais d'une caractéristique qui n'a pas rapport à son contenu (numéro ISBN, nom de l'auteur, etc.).

La classification permet de placer un document, après en avoir analysé le contenu de façon générale, dans l'ensemble des documents qui traitent du même sujet. Le document est ici considéré comme une entité. C'est un peu comme placer le document dans une boîte étiquetée: Animaux, Meubles, Romans français du 19ème siècle, etc.

Quant à l'indexation, on ne considère plus le document comme une entité distincte mais on considère plutôt les éléments d'information qui s'y trouvent. Le but de l'indexation est toujours de créer des regroupements de documents sur un même sujet, mais la description est plus précise. L'indexation nous permettra donc d'accéder par exemple à tous les documents qui fournissent de l'information sur la meilleure alimentation pour les chiens de telle ou telle race, peu importe que le document qui contient cette information ait été classé dans la grande classe Animaux, sous-classe Chiens, ou dans une classe qui serait plutôt du côté Alimentation, Médecine vétérinaire, etc.

3.1. Indexation trop spécialisée

Karen Markey Drabenstott (1986, 1988) s'est penchée sur l'indexation faite au sein des collections

d'images en art. La compréhension des images en histoire de l'art se base, en partie, sur les trois distinctions proposées par l'historien d'art Erwin Panofsky qui sont : description pré-iconographique (sujet primaire : dénotation), l'analyse iconographique (sujet secondaire : connotation) et l'interprétation iconographique ou l'iconologie (sujet tertiaire : symbolisme). Selon Markey, les collections d'images en art sont presque toujours organisées pour des gens qui possèdent des connaissances en Histoire de l'art où l'emphase est placé sur l'aspect iconographique de l'image, i.e. le sujet secondaire sans aucune mention de l'aspect pré-iconographique, i.e. le sujet primaire.

John Sunderland (1982) mentionne aussi que l'indexeur de collections d'images, dans le domaine de l'analyse du sujet et de l'iconographie, fait face à des problèmes historiques, pratiques et théoriques. Les collections d'images n'ont pas de systèmes d'organisation universellement acceptés. Donc, à cause de leur organisation, la recherche dans de telles collections est presque exclusivement réservée à des personnes versées dans la lecture d'iconographie. Mais de plus en plus, des gens de divers domaines consultent maintenant ces collections (sociologues, musiciens, publicistes, designers graphique, etc.) pour trouver ce dont ils ont besoin soit pour une annonce publicitaire ou une jaquette de livre par exemple. Le problème majeur réside dans le fait que ces nouveaux usagers, qui n'ont pas nécessairement de connaissances dans le domaine de l'iconographie, ne peuvent facilement consulter ces collections avec les systèmes actuels de repérage des images en art. L'accès de ce genre de collections est donc limité. Bref, Markey (1986, 1988) et Sunderland (1982) s'entendent sur le fait qu'il faut tenir compte des besoins de l'usager. La question demeure comment peut-on organiser la collection pour satisfaire tout le monde?

3.2. Polysémie de l'image

Un autre problème réside dans la polysémie de l'image. Ginette Bléry (1981) donne un exemple (voir figure 1) d'une photographie où l'on voit un petit garçon souriant et une petite fille se donnant un baiser alors qu'une deuxième petite fille se trouvant assise de l'autre côté du même garçon arbore une expression furieuse. Sur la même photographie, on retrouve donc, entre autres, les concepts d'amour, de joie, de jalousie, de fureur et même de trahison! Ceci rappelle une phrase de Sara Shatford Layne (1986) : " the delight and frustration of pictorial ressources is that a picture can mean different things to different people ". En effet, selon Sunderland (1982), un des grands problèmes de l'image est qu'elle peut avoir une infinité de significations. Il a donc mené une petite enquête où il a demandé à un enfant de 12 ans, à un profane et à un historien d'art de décrire une image spécifique de J.-E. Millais " Le Christ dans la maison de ses parents " (voir figure 2). La preuve quant à la polysémie de l'image selon le récepteur a été faite; chacune de ses personnes a donné, pour la même image, une interprétation bien différente l'un de l'autre :

- L'historien d'art a identifié l'artiste, le titre de l'œuvre, sa date d'exécution, l'endroit où se trouvait le tableau, le style de l'œuvre (pré-Raphaélite) ainsi que le sujet (moment prophétique pour Jésus lorsqu'il se blesse avec un clou) et il continu son interprétation en expliquant la symbolique de l'œuvre ;
- Le profane, quant à lui, indique que c'est une image religieuse où est représentée la Famille

Sainte dans l'atelier de Joseph. Il mentionne aussi le fait qu'il est évident que la famille est heureuse et que l'on voit la relation d'amour et de bonheur qui existe entre le père, la mère et l'enfant ;

- L'enfant de 12 ans a plutôt décrit les éléments se trouvant dans l'image : une femme à genou tenant un enfant. L'enfant à un trou dans sa main. Un homme tient la main de l'enfant et un clou ou quelque chose. Il y a des copeaux de bois sur le plancher, un jeune garçon avec un bol dans ses mains, etc.

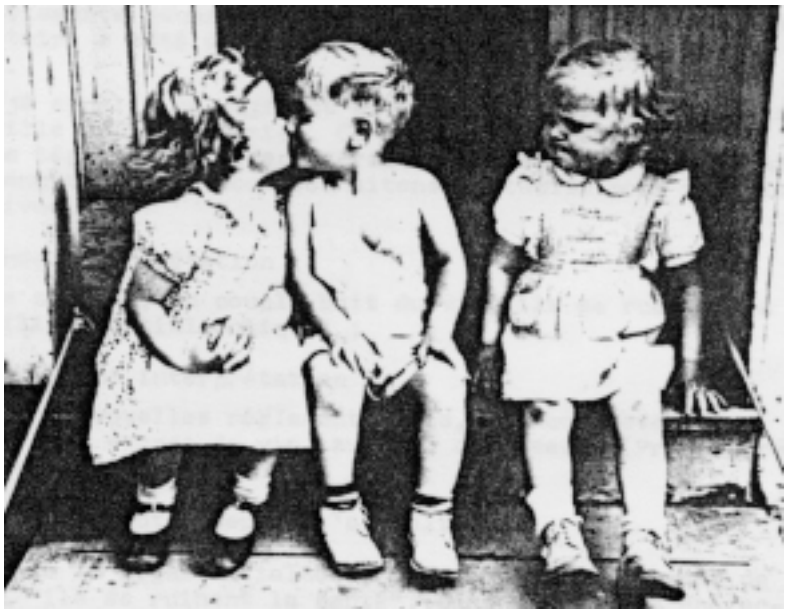

Figure 1

Bléry, Ginette. 1981. La mémoire photographique, étude de la classification des images et analyse de leur contenu grâce à l'aide de l'informatique. Interphotothèque, no 41. p. 17.

Figure 2
Sir John Everett Millais (1829-1896). " Christ in the House of His Parents ", 1850

.

3.3. Choix des termes

Selon Michael Krause (1988), un des grands problèmes porte sur le choix des termes pour l'indexation des images. En effet, comment trouver sous quelles vedettes sera placée l'image? La difficulté ne consiste pas à choisir le système d'indexation pour organiser les images mais plutôt de son application, i. e. comment définir les sujets et quels aspects de l'item ou du sujet devraient être mentionnés et les termes même qu'il faudrait choisir. Il explique qu'il existe deux types d'indexation, le " hard indexing " ou le " ofness " d'une image (l'image est de …), i.e. ce que l'indexeur voit dans l'image, soit un chat ou une femme par exemple. Et le " soft indexing " ou le " aboutness " d'une image (l'image est à propos de …) qui porte sur la signification; donc la femme sur l'image est en fait la Muse. C'est à ce niveau que les réactions personnelles et les émotions stimulées par l'image peuvent être relevées. Justement, toujours selon Krause (1988), les bibliothèques devraient donner un accès aux usagers à des idées abstraites comme la faim. Le fait d'indexer la signification des images améliore l'accès et la valeur donnés à la collection. Les usagers n'auraient pas à chercher à plusieurs endroits, à l'aveuglette, pour trouver ce dont ils ont besoin. Les indexeurs devraient se pencher plus sur la signification des images et sur les différentes utilisations qu'on pourrait faire d'une image pour mieux l'indexer. Ceci permettra à l'usager de trouver beaucoup plus rapidement ce dont il a besoin avec une plus grande efficacité puisque ces images auront été indexées avec des termes qui correspondent mieux à leur besoin. Mais Krause (1988) note que ce n'est pas le cas, les indexeurs ne veulent pas aborder la signification de l'image de peur de

devenir subjectif au moment de l'indexation.

Ginette Bléry (1981) pense aussi qu'il faudrait aborder cette question de la subjectivité de l'image. Elle a effectué une expérience où elle a demandé à des personnes (elle ne mentionne pas le nombre de participants) d'indiquer à l'aide de mots abstraits leurs impressions pour chacune des images qu'on leur montre (ni la quantité ni le genre d'image n'étaient indiqués). Des 438 réponses obtenues, elle a pu tirer quatorze couples d'opposition. Selon Bléry (1981), il n'y a pas une infinité de possibilités de significations pour les images et il suffit d'identifier les émotions que peut susciter une image parce qu'en général elle est la même pour tout le monde. Toutefois, il faudrait répéter l'expérience de façon plus scientifique avec un plus grand nombre de participants et d'images. Voici donc les quatorze couples d'opposition en question qui pourraient servir à décrire l'aspect subjectif d'une image :

- Abstrait et Sensuel
- Actif et Passif
- Ancien et Moderne
- Apaisant et Stimulant
- Artificiel et Naturel
- Beau et Laid
- Sérieux et Frivole
- Chaud et Frais
- Coloré et Terne
- Gai et Triste
- Comique et Tragique
- Décontracté et Angoissé
- Érotique et Froid
- Ordonné et Discordant

3.4. Constance de l'indexation

Karen Markey Drabenstott (1984) aurait découvert qu'il y avait une faible constance entre indexeurs lors d'une de ses expériences. Et James Turner (1994) s'est rendu compte que le terme le plus populaire était le même pour environ 60% des participants. Il en découle que la constance inter-indexeur sera plus forte pour le sujet principal et les aspects objectifs que pour les sujets secondaires ou les aspects subjectifs d'une image. Par ailleurs, Sara Shatford (1994) mentionne dans un article plusieurs auteurs ayant étudié la constance de l'indexation.

3.5. Les besoins des usagers

Kevin Roddy (1991), quant à lui, relève que même si les images sont emmagasinées de façon intelligente, elles restent inaccessibles étant donné qu'elles possèdent trop peu de descripteurs. Et le problème des descripteurs est qu'ils sont ambigus, arbitraires et portent à confusion ou à des désaccords. Il indique qu'il serait important que le système indique des valeurs aux items repérés par degré de

rapprochement avec la requête. De plus, le plus grand problème de l'accès aux images est l'impossibilité de fournir de l'information sur ce qui serait une session typique de recherche d'information. Il faudrait pouvoir anticiper les besoins du chercheur et il est rare, par exemple, que quelqu'un veuille tout genre de maisons ou d'églises. L'usager habituellement veut un certain genre de maison en particulier.

Dans son mémoire, Murray (1984) mentionne un article de Schroeder qui indique que le problème majeur pour l'indexation d'une collection est l'utilisation future des images. Selon lui, il est impossible d'anticiper les besoins des usagers parce qu'on ne peut jamais savoir ce qui va servir dans une image. Comment faire si un usager veut savoir à partir de quand les femmes commencent à se croiser les jambes? Quand les hommes ont commencé à poser avec des sourires? Quand est-ce qu'on a commencé à montrer les dents en souriant? Ce genre d'information ne fait pas partie des sujets primaires qui seront indexés. Voici d'autres exemples d'éléments qui pourraient être considérés de l'information par des usagers mais qui n'aura pas été tenu en considération lors de la création de l'image et qui ne sont donc pas indexés :

- Contenu accidentel (objets, activités, etc.) ;
- Intangibles (relation d'espace entre les gens, entre les objets, etc.) ;
- Conventions photographiques (arrangement typique d'une équipe
- sportive, portrait de famille, etc.) ;
- Convention de perspective et de sélection ;
- Semi-intangibles (gestes, postures, expressions faciales, etc.).

À l'opposé, Wright qui a écrit un des chapitres du livre Picture Librarianship (écrit par Helen Harrison en 1981) considère qu'il y a très peu d'information dans les images et que l'utilisation n'est pas importante lors du processus d'indexation. La seule chose indexable sont les objets (Murray, 1984).

3.6. Transfert de la signification

Un autre grave problème dans le domaine de l'indexation des images est le transfert de la signification d'un médium visuel vers du verbal. On croit en effet que ce transfert cause la perte d'une partie de l'information (Murray, 1984). Surtout lorsque l'œuvre d'art ne contient pas d'information comme titre, attribution, description, etc. En effet, Sunderland (1982) explique qu'il y a des images où l'on ne sait pas de quoi il s'agit à moins que le créateur ne l'ait indiqué. À la figure 3, cette image pourrait aussi bien représenter une famille partant en pique-nique un dimanche qu'une famille de récents immigrés en Alberta. En fait, elle représente des fermiers américains de l'Arkansas qui à cause de la fameuse crise des années 30 se voient obligés d'abandonner leur terre et de suivre l'exode vers la Californie (voir le film Les raisins de la colère de John Ford tourné en 1940). Par ailleurs, certains croient qu'il est impossible de traduire de l'information visuelle à l'intérieur d'une structure verbale et même que le langage verbal est inadéquat pour exprimer le langage visuel. Parce que " les images sont des sujets alors que des livres sont à propos de sujets " (Murray, 1984). Il est peut-être possible de faire passer le message sémantique d'une image qui est d'ordre logique mais pas son aspect esthétique qui n'est pas transférable à moins de perdre une partie de la valeur du message pictoriel (Mole dans Murray, 1984). Encore de nos jours il est

question de ce problème, Elaine Svenonius (1994) se penche sur la question de transfert du visuel vers le textuel et Dennis Hogan (Baxter et Anderson, 1995) dit même qu'il faut utiliser autre chose que du verbal pour indexer le non-verbal.

Figure 3
Bléry, Ginette. 1981. La mémoire photographique, étude de la classification des images et analyse de leur contenu grâce à l'aide de l'informatique. Interphotothèque, no 41. p. 22.

3.7. Normes

Par ailleurs, Graeme Baxter et Douglas Anderson (1995) mentionnent le manque de normes pour l'indexation des images. Déjà au début des années 80, Thomas Ohlgren faisait état du besoin évident de normes pour la classification et la description d'images.

3.8. Autres problèmes

Il y a d'autres aspects à considérer aussi : certaines images comme diapositives, coupures de journaux ou illustrations de livres coûtent peu et ont une durée de vie courte, l'indexation de tels items vaut-il la peine surtout si son coût dépasse celui de l'acquisition et le maintient de telles collections? De plus, la quantité d'images augmente tellement ces jours-ci qu'un système d'indexation qui pouvait être efficace pour de petites collections peut devenir inutilisable lorsque la collection double ou triple d'importance. Aussi, si les images sont indexées ou classifiées sous des vedettes-matières trop larges, il est alors presque impossible de les repérer ce qui équivaut à les perdre. De plus, il est courant que le personnel s'occupant de telles collections ne possède ni de formation en sciences de l'information ni la connaissance des principes et procédures pour le maintien de ce genre de documents. Il en découle que les personnes responsables de ces collections se sentent débordées et que la situation est difficile à améliorer étant donné la pénurie de cours offert sur l'organisation de matériels pictoriels (Murray, 1984).

Comme le mentionne Peter Enser (1995), il est facile de générer des documents mais difficile d'y obtenir un accès physique et encore plus difficile de repérer les quelques-uns qui satisferaient un besoin spécifique d'information.

4. Recherches à faire

En effectuant la revue de littérature dans le domaine de l'indexation des images, bon nombre d'articles mentionnent des pistes et idées de recherche qui contribueraient à l'avancement du domaine. Voici les grands champs qui ont pu être déterminés grâce aux lectures.

Shatford Layne (1994) mentionne un besoin de recherche concernant la stratégie de recherche d'images, le choix des termes d'indexation, les requêtes utilisées par l'usager et les raisons qui font en sorte que les images repérées sont utiles. Toujours selon Layne (1994), il faudrait effectuer des recherches quantitatives pour déterminer quels seraient les meilleurs attributs des images examinées sous lesquels elles devraient être indexées. Or, la récente thèse de doctorat de Layne (1997) pourra servir de tremplin à d'autres recherches dans ce domaine.

Selon Karen Markey Drabenstott 1988), il y avait un manque de recherche au niveau des usagers de collections visuelles. Cependant, est-ce encore le cas de nos jours, dix ans après la publication de cet article? En se basant sur la revue de littérature que ses étudiants mettent sur pied au niveau de la maîtrise et du doctorat en sciences de l'information à l'Université de Michigan, il semblerait que la situation soit toujours la même, i.e. qu'il existe très peu de littérature et de recherche concernant les usagers de collections visuelles (Drabenstott, 1998). Que sait-on de leurs requêtes ? De leurs besoins en information ? De l'utilisation éventuelle du matériel qu'ils recherchent dans ces collections ? Sait-on s'ils réussissent à trouver ce qu'ils cherchent dans ces collections ? Y a-t-il un type d'usager qui réussi plus que d'autres ? Quelles sont les caractéristiques de ces usagers en particulier ? Est-ce que le taux de réussite lors de recherches dans les collections de ressources visuelles est différent de recherches effectuées dans des collections traditionnelles basées sur l'imprimé ?

Il faudrait aussi se pencher sur la question de perception et mémoire visuelles. Rita Murray (1984) considère que de telles recherches permettraient aux bibliothécaires de mieux choisir la façon d'organiser leur collection d'image et de la rendre accessible ce qui contribuerait peut-être éventuellement à l'élaboration d'une théorie de l'analyse du sujet et même de la représentation de l'image.

L'indexation d'une collection d'images est primordiale pour son bon fonctionnement. Odile Le Guern (1989) mentionne que le processus d'indexation devrait s'intéresser à la sémiologie de l'image qui lui indiquerait des pistes pour une meilleure description des documents tout en lui fournissant une grille d'analyse valable une fois pour toutes et pour toutes les images.

Selon Graeme Baxter et Douglas Anderson (1995), il faudrait concevoir un projet qui examinerait les normes pour l'indexation et le repérage d'images, d'un point de vue multidisciplinaire comme ce que le

Getty Art History Information Programme fait en ce moment pour les images en art et les objets culturels. Même que la mise en œuvre de recherches dans le but d'identifier et de développer des mécanismes reconnus de façon internationale pour le repérage d'images devrait être une priorité.

Le nombre impressionnant de projets en chantier de nos jours atteste de la popularité du repérage des images par des moyens informatiques. En effet, ce domaine ouvre la voie à de nombreuses recherches pleines de défis. Toutefois, le langage visuel qui est à la base des thésaurus visuels, de la reconnaissance de formes, et de bien d'autres systèmes de classification nécessite de solides connaissances dans le domaine mathématique et informatique.

5. Conclusion

Ce survol a permis de brosser un portrait général du monde de l'indexation des images en mentionnant les grands projets d'organisation d'images dans le domaine des arts lorsque l'intérêt pour les collections visuelles s'est accru à la fin des années 70 et le début des années 80 comme la Image Access Society of North America, la List of Subject Headings pour la Library of Congress Prints and Photographs Division, le Thésaurus de la Division de l'iconographie des Archives publiques du Canada et le Art and Architecture Thesaurus (AAT). Par la suite, les grands systèmes d'indexation d'images ont été regroupés et expliqués sous les rubriques suivantes : méthodes d'indexation basées sur thésaurus et systèmes de classification, thésaurus visuels et systèmes hybrides associés, langages de description pictorielle et indexation et repérage d'images par contenu. Finalement, les problèmes liés à l'indexation des images ainsi que les recherches à faire dans le domaine des images permettent de faire le point sur les directions futures que devraient considérer les chercheurs en sciences de l'information.

En effet, Cawkell (1993) mentionne qu'en plus d'un manque extraordinaire de littérature dans le domaine de l'indexation des image, il semblerait même que cette question soit évitée par les professionnels et les chercheurs concernés. Il en découle que les personnes oeuvrant dans le domaine du traitement et de la reconnaissance d'images ignorent souvent les travaux de ceux qui s'occupent de l'organisation et du repérage d'images et vice-versa. Il faudrait donc qu'une atmosphère de collaboration règne si l'on veut profiter de ces deux branches des sciences de l'information où seul l'échange d'information, de théories, d'idées et d'aide pourra faire en sorte que des progrès se réalisent.

6. Bibliographie

Baxter, Graeme et Douglas Anderson. 1995. Image indexing and retrieval : some problems and proposed solutions. New Library World 96, no 1123 : 4-13.

Bell, Lesley Ann. 1994. Gaining Access to visual Information. Art Documentation 13, no 2 (Summer) : 89-94.

Besser, Howard. 1990. Visual access to visual images : the UC Berkely Image Database Project. Library Trends 38, no 4 (Spring) : 787-798.

Bléry, Ginette. 1981. La mémoire photographique, étude de la classification des images et analyse de leur contenu à l'aide de l'informatique. Interphotothèque, no 41 : 9-34. (no spécial sur l'analyse de l'image fixe).

Bordogna et al. 1990. Pictorial indexing for an integrated pictorial and textual environment. Journal of Information Science 16, no 3 : 165-173.

Brilliant, Richard. 1988. How an art historian connects art objects and information. Library Trends 37, no 2 (Fall) : 120-129.

Brooks, Diane. 1988. System-system interaction in computerized indexing of visual materials : a selected review. Information Technology and Libraries 7, no 2 (June) : 111-123.

Cawkell, A.E. 1992. Selected aspects of image processing and management : review and future prospects. Journal of Information Science 18 : 179-192.

----------. 1993. Developments in Indexing Picture Collections. Information Services & Use 13 : 381-388.

Chang, C.-C. et Wu, T.-C. 1992. Retrieving the most similar symbolic pictures from pictorial databases. Information Processing and Management 28, no 5 : 581-588.

Couprie, L.D. 1983. Iconclass : an iconographic classification system. Art Libraries Journal 8, no 2 (Summer) : 32-49.

Drabenstott, Karen. 1998. Courrier électronique envoyé le 17 avril à l'auteur de l'article.

Eakins, J.P. 1993. Design Criteria for a Shape Retrieval System. Computers in Industry 21 : 167-184.

Enser, P.G.B. 1995. Progress in Documentation Pictorial Information Retrieval, Journal of Documentation 51, no 2 : 126-170.

Evans, A. 1987. TELCLASS : a structural approach to TV classification. Audiovisual Librarian 13, no 4 : 215-216.

Falk, Jane E. 1983. Subject Access Systems and the Visual Arts. Master diss., John f. Kennedy University.

Garnier, François. 1984. Thésaurus iconographique : système descriptif des représentations. Paris : Léopard d'or.

Greenberg, Jane. 1993. Intellectual control of visual archives : a comparaison between the Art and Architecture Thesaurus and the Library of Congress Thesaurus for Graphic Materials. Cataloging & Classification Quarterly 16, no 1 : 85-117.

Hogan, M. et al. 1991. The visual thesaurus in a hypermedia environment : a preliminary exploration of conceptual issues and applications, International Conference on Hypermedia and Interactivity in Museums, Pittsburgh, October 1991, Archives and Museum Informatics, Pittsburgh, 1991 : 202-221.

Hudon, Michèle. 1998. Courrier électronique envoyé en avril à l'auteur de l'article.

Keefe, J.M. 1990. The image as document : descriptive programs at Rensselaer. Library Trends 38, no 4 : 659-681.

Krauze, Michael G. 1988. Intelletual problems of indexing picture collections. Audiovisual Librarian 14, no 4 (November) : 73-81.

Le Guern, Odile. 1989. Images et bases de données. Bulletin des Bibliothèques de France 34, no 5 : 422-435.

Lee, Eric et James MacGregor. 1987. Computer retrieval of graphic information. Canadian Journal of Information Science 12, nos. 3-4 (1987) : 80-88.

Leung, C.H.C. et al. 1992. Picture retrieval by content description. Journal of Information Science 18, no 2 : 111-119.

Markey, Karen. 1984. Interindexer consistency tests : a literature review and report of a test of consistency in indexing visual materials. Library and Information Science Research 6 : 155-177.

----------. 1986. Subject access to visual resources collections : a model for computer construction of thematic catalogs. New York : Greenwood Press.

----------. 1988. Access to iconographical research collections. Library Trends 37, no 2 (fall) : 154-174.

Milstead, Jessica L. 1994. Needs for Research in Indexing. Journal of the American Society for Information Science 45, no 8 : 577-582.

Murry, Rita. 1984. Criteria for Subject Indexing of Pictures : an Introductory Survey. Master diss.,

University of Alberta.

Niblack, W. et al. 1993. The QBIC project : querying images by content using colour, texture, and shape in Niblack, W. (Ed.), Storage and Retrieval for Image and Video Databases : Proceedings of the SPIE, no 1908, San Jose, California, February 1993, SPIE, Bellingham, 1993, pp. 173-187.

Ohlgren, Thomas H. 1982. Image analysis and indexing in North America : a survey. Art Libraries Journal (summer) : 51-60.

Pejterson, A. M. 1993. A new approach to design of document retrieval and indexing systems for OPAC users in Raitt, D.I. and Jeapes, B. (Ed.), Online Information 93 ; Proceedings of the 17th International Online Information Meeting, London, 7-9 December 1993, Learned Information (Europe), Oxford, 1993, pp. 273-290.

Petersen, T. 1990. Developing a new thesaurus for art and architecture. Library Trends 38, no 4 : 644-658

Roddy, Kevin. 1991. Subject access to visual resources : what the 90s might portend. Library Hi Tech 9 : 1 (no 33) : 45-49.

Seloff, G.A. 1990. Automated access to the NASA-JSC image archives. Library Trends 38, no 4 : 682-696.

Shatford, Sara. 1984. Describing a picture : a thousand words are seldom cost effective. Cataloging & Classifcation Quarterly 4, no 4 (Summer) : 13-30.

----------. 1986. Analyzing the subject of a picture : a theoretical approach. Cataloging & Classifcation Quarterly 6, no 3 (Spring) : 39-62.

----------. 1994. Some Issues in the Indexing of Images. Journal of the American Society for Information Science 45, no 8 : 538-588.

Sherman, C.R. 1987. ICONCLASS : a historical perspective. Visual Resources 4, no 3 : 237-246.

Soergel, Dagobert. 1994. The Art and Architecture Thesaurus (AAT): A Critical Appraisal. Visual Resources X, no 4 : 369-400.

Svenonius, Elaine. 1994. Access to Nonbook Materials : The Limits of Subject Indexing forVisual and Aural Languages. Journal of the American Society for Information Science
45, no 8 : 600-606.

Teel, Kay. 1992. Subject access to visual materials. SIG/CR News (American Society for Information Science, Special Interest Groupe/Classification Research) 3 (August) : 1-5.

Turner, James M. 1993. Subject Access to Pictures : Considerations in the Surrogation and Indexing of visual Documents for Storage and Retrieval. Visual Resources IX : 241-271.

----------. 1994. Indexing " Ordinary " Pictures for Storage and Retrieval. Visual Resources X : 265-273.

Vézina, Raymond. 1998. Entrevue le 23 février.

Survol du monde de l'indexation des images.

Retour à la Table des matières -- Page d'accueil de Cursus -- Page d'accueil de l'EBSI
