Documentation et bibliothèques

Éléments méthodologiques pour un traitement factuel des documents administratifs et juridiques

Indexation of Administrative and Legal Documents: A Methodology

Elementos metodológicos para un trato factual de documentos administrativos y jurídicos

François Charoy et Lylette Lacote-Gabrysiak

Volume 42, numéro 1, janvier–mars 1996

URI : https://id.erudit.org/iderudit/1033321ar
DOI : https://doi.org/10.7202/1033321ar

Aller au sommaire du numéro

Éditeur(s)

Association pour l'avancement des sciences et des techniques de la documentation (ASTED)

ISSN

0315-2340 (imprimé)
2291-8949 (numérique)

Découvrir la revue

Citer cet article

Charoy, F. & Lacote-Gabrysiak, L. (1996). Éléments méthodologiques pour un traitement factuel des documents administratifs et juridiques. *Documentation et bibliothèques*, *42*(1), 19–25. https://doi.org/10.7202/1033321ar

Résumé de l'article

Le travail décrit ici vise à permettre une exploitation factuelle de documents de type administratif ou juridique. Par exploitation factuelle il faut entendre la création d'une base de données qui permette une prise en compte précise des faits et non une simple description du contenu des documents. Diverses expérimentations ont été entreprises. Elles ont abouti à une solution satisfaisante prenant la forme d'éléments méthodologiques pour la mise en place de telles applications et d'une définition des contraintes à respecter (fourniture du texte intégral, existence de moyens d'accès différenciés à l'information, fourniture d'une description de la structure de la base de données sous la forme, par exemple, d'une classification, apparition de la dimension analytique de ce type de base de données).

Éléments méthodologiques pour un traitement factuel des documents administratifs et juridiques

François Charoy*
Lylette Lacote-Gabrysiak*
Université de Nancy 2

Le travail décrit ici vise à permettre une exploitation factuelle de documents de type administratif ou juridique. Par exploitation factuelle il faut entendre la création d'une base de données qui permette une prise en compte précise des faits et non une simple description du contenu des documents. Diverses expérimentations ont été entreprises. Elles ont abouti à une solution satisfaisante prenant la forme d'éléments méthodologiques pour la mise en place de telles applications et d'une définition des contraintes à respecter (fourniture du texte intégral, existence de moyens d'accès différenciés à l'information, fourniture d'une description de la structure de la base de données sous la forme, par exemple, d'une classification, apparition de la dimension analytique de ce type de base de données).

Indexation of Administrative and Legal Documents: A Methodology

The following paper describes the factual exploitation of administrative and legal documents. Factual exploitation involves the creation of a data base that considers the facts of the document instead of a brief description of the contents. Several experiments were undertaken. They helped identify a methodology and several criteria to respect, such as the availability of a full text, the existence of different ways to access the information, the availability of the description of the data base including the classification, and the emergence of the analytic dimension of this type of data base.

Elementos metodológicos para un trato factual de documentos administrativos y jurídicos

El trabajo descrito aquí intenta permitir una explotación factual de documentos de tipo administrativo o jurídico. Par explotación factual, debemos entender la creación de una base de datos que permita tomar en cuenta precisa los hechos y no una simple descripción del contenido de los documentos. Diversas experimentaciones han sido intentadas. Estas han llegado a una solución satisfactoria que toman la forma de elementos metodológicos par la colocación de tales aplicaciones y de una definición de las dificultades a respetar (provisión del texto integral, existencia de medios de acceso diferenciados a la información, provisión de una descripción de la estructura de la base de datos sobre la forma, por ejemplo, de una clasificación, aparición de la dimensión analítica de este tipo de base de datos).

Combien de temps et d'argent sont perdus chaque année à défaut de pouvoir retrouver de manière pertinente des documents et études sommairement archivés? Comment recueillir de manière simple des données précises issues de textes de jurisprudence, de conventions ou d'accords collectifs? Le besoin d'un traitement documentaire efficace et pertinent des documents administratifs ou légaux devient de plus en plus prégnant. En effet, ces documents contiennent une masse très importante d'informations qui ne sont pas ou peu exploitées. Un traitement documentaire classique semble insuffisant. Il serait en fait nécessaire de pouvoir retrouver une information précise dans un texte ou, plus exactement, de pouvoir effectuer des requêtes sur des éléments factuels. Par exemple à partir d'un corpus de jurisprudence, connaître le montant de la pension alimentaire accordée, dans un cas de divorce par consentement mutuel quand une femme au chômage obtient la garde de trois enfants mineurs, de la part d'un mari ayant un salaire supérieur à 12 000 FF ou encore d'après un corpus de PV d'accident établi par la gendarmerie, savoir le montant moyen du taux d'alcoolémie dans les accidents mortels hors agglomération, etc. Le centre ICO de Québec mène dans ce sens de nombreuses recherches, notamment afin d'éviter la perte de temps et de compétences induite par la nécessité de refaire ce qui a déjà été fait dans de nombreuses administrations (Rochon, 1992)

Si le traitement automatique des langues représente une solution possible à ce problème, l'état des recherches et, surtout, la nature des documents eux-mêmes (structure et vocabulaire non normalisés,

* François Charoy est maître de conférence en informatique à l'IUTA de l'Université de Nancy 2 et chercheur au Centre de recherche en informatique de Nancy (CRIN) et Lylette Lacote-Gabrysiak est maître de conférence à l'Université de Nancy 2 (IUTA) et chercheur du Groupe de recherche sur les enjeux de la communication (GRESEC) de l'Université Stendhal (Grenoble 3).

utilisation importante de l'implicite, etc.) ne permettent pas d'envisager d'applications immédiates[1].

Le travail qui a été réalisé consiste à évaluer les possibilités de réalisation d'une base factuelle et exhaustive d'après des documents de type administratif ou juridique. Le corpus retenu pour la réalisation des tests est constitué d'accords d'entreprise[2]. L'originalité de la conception de cette base repose sur l'approche factuelle: il s'agit de reprendre les faits contenus dans les documents. Cette prise en compte précise du contenu doit se traduire dans la structure même de la base de données (champs spécifiques adaptés à la description des faits). Par exemple, le texte «*Le salaire de base augmente de 2 % pour l'ensemble du personnel*» devra pouvoir faire l'objet d'une description sous la forme: *Thème*: salaire de base; *Modification*: augmentation; *Montant de l'augmentation*: 2 %; *Personnel concerné*: tout le personnel.

Mais bien sûr ces champs devront pouvoir avoir des contenus suffisamment normalisés pour être interrogeables. Ainsi l'utilisateur devrait avoir la possibilité d'interroger sur des faits et obtenir ces mêmes faits en résultat de recherche. La structure de la base de données contient en ce sens une dimension analytique.

Contraintes et éléments méthodologiques

Lors des enquêtes réalisées auprès des utilisateurs potentiels du produit et des diverses expérimentations entreprises, quatre contraintes générales ont pu être dégagées et des éléments méthodologiques pour le traitement factuel des documents administratifs et juridiques ont été définis.

Les contraintes dégagées sont les suivantes :

- obligation de retour au texte intégral des documents afin de limiter les difficultés inhérentes à la reprise des données implicites et les éventuelles erreurs découlant des choix, toujours contestables, du concepteur de l'application (toutes les informations ne pouvant être reprises);
- besoin de diversifier les moyens d'accès à l'information afin de pouvoir répondre à la diversité et à l'imprécision des besoins des utilisateurs;
- besoin également de présenter d'une manière simple et formelle la structure de la base de données autant pour les concepteurs que pour les utilisateurs de celle-ci;
- nécessité d'utiliser une modélisation relationnelle étendue ou objet puisque le type d'application envisagée ici oblige à l'utilisation de champs aux valeurs multiples, de relations d'héritage, de classes composites et de relations complexes entre les données (Bertino and Martino 1991).

Les éléments méthodologiques définis se divisent en cinq phrases : identification des objectifs, prémodélisation, modélisation, conception de l'interface utilisateur et réalisation de la base.

Identification des objectifs

L'identification des objectifs est effectuée grâce à une enquête sur les besoins des utilisateurs et une première étude des textes. L'enquête doit, bien sûr, permettre d'identifier les besoins, les désirs et les attentes des utilisateurs mais elle doit également permettre de déterminer qui sont ces utilisateurs. Il faut à ce niveau se méfier d'une catégorisation hâtive basée, par exemple, sur leur appartenance à un groupe professionnel et lui préférer une distinction fondée sur leurs attentes et leur comportement durant l'enquête. Il faut également déterminer au cours de cette enquête :

- le degré d'implication dans l'application des personnes interrogées: sont-elles à considérer comme des co-concepteurs de l'application, comme des membres engagés dans la réalisation du produit ou simplement comme des utilisateurs potentiels éventuellement intéressés si l'application est mise en place?
- leur degré de connaissance des documents: il s'agit ici du traitement de documents administratifs ou juridiques qui ne sont pas toujours accessibles. Aussi, de nombreux utilisateurs peuvent connaître ces documents, en avoir consulté quelques-uns, avoir lu des travaux portant sur ces textes mais ne pas en avoir une connaissance précise, ce qui implique l'existence de demandes ne correspondant pas à des réalités possibles;
- leur degré de connaissance des outils documentaires: les personnes impliquées interrogées sont-elles en mesure de formuler des requêtes, de bâtir seules des stratégies de recherche? Ces données sont intéressantes à connaître pour la mise en place de l'interface utilisateur. En effet, si le public visé connaît mal les outils nécessaires à la recherche, il faudra prévoir la mise en place d'une interface adaptée surtout si la base est consultable à distance.

Au titre de l'étude des besoins des utilisateurs, il faut signaler l'importance fondamentale de la rétroaction. Le documentaliste s'appuie sur les besoins exprimés afin de structurer son application mais une fois l'application mise en place, ces besoins évoluent et les utilisateurs expriment d'autres demandes. On entre ici dans un processus circulaire. Cette réflexivité doit évidemment être prise en compte, mais on ne peut pas faire évoluer facilement une structure de base de données. Aussi, si les besoins des utilisateurs doivent effectivement être étudiés et retenus, il ne faut pas s'appuyer uniquement sur ceux-ci pour la définition des objectifs de l'application.

La première étude des textes consiste, d'entrée de jeu, à recenser les sources d'information disponibles sur les documents. Afin de savoir si, bien sûr, ces produits peuvent être utilisés mais surtout pour éviter les redondances. Ensuite, elle est constituée par la lecture d'un corpus

1. Sur les applications du traitement automatique des langues en documentation voir Jacques Rouault, *Linguistique automatique: applications documentaires* (Berne: Peter Lang, 1987), 309 p.

2. Ce travail a été effectué dans le cadre d'une thèse en Sciences de l'information et de la communication: Lylette Lacote-Gabrysiak, *Modélisations factuelles d'informations textuelles: élaboration de bases de données d'après des accords d'entreprise* (Grenoble: Université de Grenoble III-GRESEC, 1994).

logique de documents: il est nécessaire de s'imprégner des textes. Des discussions avec des spécialistes du ou des domaines considérés sont également indispensables. Ces discussions doivent permettre de situer le document dans le domaine de connaissance, de cerner son ou ses contextes d'écriture, de juger de son importance et de ses conséquences. À ce sujet, la demande, souvent formulée, d'une double compétence chez les documentalistes peut être contestée. En effet, il apparaît peu probable qu'on trouve rassemblées, chez la même personne et pour chaque type d'application, des connaissances réelles et précises d'un domaine particulier et des connaissances documentaires suffisantes afin de pouvoir mettre en place une application complexe et originale. De plus, de nombreux types de documents sont multidisciplinaires. Le fait que leur traitement soit réalisé par un spécialiste dans un seul des domaines couverts peut conduire à privilégier un point de vue par rapport aux autres présents dans le texte. Enfin, si des personnes traitent habituellement les textes, il est indispensable de bénéficier de leur expérience des documents.

Prémodélisation

La phase suivante est une phase de prémodélisation : il s'agit de dépouiller les documents dans l'optique de leur transformation sous forme de classes reliées par des liens et des champs. Cette phase doit permettre de répondre à des questions d'ordre général, d'établir une classification et d'évaluer les problèmes inhérents aux documents.

Dans un cadre général, la phase de prémodélisation doit permettre de définir:

- *l'existence possible de critères de pertinence de l'information dans les documents*. Certains passages peuvent-ils systématiquement être considérés comme dénués d'intérêt ? On peut toujours définir des critères de pertinence généraux (originalité du texte par rapport au reste du corpus, statut de l'émetteur, etc.) mais ces critères restent informels et leur utilisation est, de ce fait, toujours problématique;

- *l'existence et l'importance des éléments contextuels*. Si les documents citent explicitement des faits ou des textes, cela oblige à prendre en compte ces informations. Certains éléments contextuels sont-ils indispensables à la compréhension des documents ?

- *l'unité documentaire à retenir*. Les textes constituent une unité documentaire physique incontournable mais il importe de déterminer s'ils sont inclus dans une unité plus vaste (par exemple un texte de jurisprudence n'est qu'une étape dans une affaire juridique) ou s'ils contiennent plusieurs unités thématiques (par exemple, un accord d'entreprise contient plusieurs thèmes plus ou moins indépendants);

- *le parallélisme des structures logique et sémantique du texte*. Si c'est effectivement le cas, cela permet d'envisager l'utilisation lourde d'un langage de balisage pour le traitement des documents. Si les deux structures coïncident seulement en partie, il importe de voir dans quelle mesure une telle structuration du texte peut être utile (repérage des différents thèmes, économie d'attributs)[3].

L'ensemble des expérimentations réalisées ont également montré qu'il est nécessaire d'offrir une présentation claire, simple et exhaustive du contenu des documents. Pour cela une solution satisfaisante est d'utiliser une classification. L'usage d'un tel outil pour représenter des informations mouvantes et complexes peut sembler étonnant. Mais, justement, il s'agit de coller sur le contenu des documents particulièrement difficile à cerner une structure rigide mais également assez riche sémantiquement pour permettre aux concepteurs comme aux utilisateurs de la base de pouvoir se repérer dans ce contenu[4]. Cette classification peut être réalisée lors de la phase de prémodélisation. Dans un premier temps, une préindexation du contenu des documents est réalisée. À chaque segment textuel qui présente un caractère thématique autonome est attribué un mot clé. Les mots clés sont ensuite rassemblés. Un premier traitement doit permettre l'élimination des synonymes. Puis les mots restant sont organisés hiérarchiquement (dans les aménagements du temps de travail, il y a les aménagements en creux et bosses, les aménagements permettant l'attribution de congés supplémentaires, etc.). Enfin, des classes virtuelles sont ajoutées afin de permettre une représentation logique et complète des thèmes pris en compte. Par exemple: le temps de travail ne sera jamais utilisé comme mot clé lors de l'indexation du texte des accords d'entreprise mais cette notion recouvre effectivement un grand nombre de mesures (horaires, congés, aménagements du temps de travail, etc.). Aussi, afin de permettre une meilleure lisibilité de la classification, la classe *Temps de travail* sera-t-elle ajoutée.

La phase de prémodélisation, notamment au cours de la préindexation des documents, permet également d'évaluer l'importance des problèmes inhérents aux documents eux-mêmes. Ces problèmes peuvent être rassemblés dans deux catégories génériques : les données implicites et l'hétérogénéité du corpus.

Tout acte de communication implique l'utilisation d'informations implicites[5]. Les documents administratifs ne font pas exception à cette règle. En fait, l'utilisation de données implicites prend trois formes principales :

- *importance du contexte*. Des références sont faites au contexte ou plutôt aux contextes existants, contexte particulier de la signature, contexte local, régional, national ou international. Or, ces références sont plus ou moins claires. Surtout si le document est exploité dans un autre espace-temps que celui correspondant à sa rédaction, c e qui peut rendre impossible l'identification du contexte. Dans certains cas enfin, les références au contexte peuvent avoir un caractère totalement implicite.

3. À propos des langages de balisage, voir Martin Bryan, *SGML: An author's guide to the standard generalized markup language* (Reading: Addison-Wesley, 1998).

4. Concernant l'utilisation d'une classification notamment dans le cadre d'une base de données hypertextuelle, voir Aboud, M. et al., «Querying a hypertext information retrieval system by the use of classification.» *Information Processing & Management* 29, no. 3 (1993): 387-396.

5. À propos de l'implicite et de ses utilisations, voir Catherine Kerbrat-Orecchioni, *L'implicite* (Paris: Armand Colin, 1986), 404 p. (Linguistique).

- *utilisation d'un langage de spécialité*. On remarque dans les documents administratifs ou légaux l'utilisation d'un ou de plusieurs langages de spécialité. Cette utilisation peut gêner la compréhension même du texte. De plus, cela pose le problème de la reprise des éléments textuels dans une base de données : faut-il reprendre les termes employés ? Dès lors, les données risquent de ne pas être compréhensibles ou faut-il, au contraire, remplacer ces mots par des termes plus courants ? Il s'agit alors d'une véritable traduction avec les éventuels glissements de sens que cela implique.

- *existence d'une stratégie rédactionnelle*. L'implicite peut également être utilisé volontairement par le rédacteur du document afin de dissimuler certains faits ou d'empêcher une compréhension complète de l'information décrite. L'existence d'une telle stratégie peut généralement être perceptible à la lecture de l'ensemble du document. En revanche, la consultation d'extraits ou la reprise factuelle des informations risque de fausser la compréhension en empêchant l'utilisateur de prendre conscience de cette stratégie. Il n'est pas non plus possible de signaler explicitement celle-ci en raison du caractère ambigu de toute interprétation.

L'hétérogénéité ou plutôt le manque d'homogénéité du corpus peut également gêner l'exploitation des documents sans que des réponses réelles puissent être apportées puisque ce problème découle des documents eux-mêmes. En fait, cette hétérogénéité peut concerner :

- *la forme des documents*. Ainsi, dans ce que l'on désigne sous le terme générique d'accord d'entreprise, coexistent des accords, des avenants, des désaccords, des quasi-accords et des refus de négociation. Or ces types de document présentent des dehors formels très différents;

- *le fond* : un document peut être plus ou moins atypique ou présenter un intérêt variable. Au sein du même document certains extraits dénués d'intérêt peuvent voisiner avec des parties particulièrement riches;

- *l'évolution des documents dans le temps*. Les documents sont évolutifs. Cette évolution peut concerner aussi bien leur forme que leur fond. De plus, un document considéré comme atypique à un moment donné peut devenir courant l'année suivante. Enfin, l'existence même de la base de données peut influer sur l'évolution des documents. On retrouve ici le délicat problème de la rétroaction.

Modélisation

Une fois la prémodélisation effectuée, il est possible de réaliser la modélisation elle-même. Cette modélisation peut se décomposer en deux parties : la modélisation des références des documents et du contexte et la modélisation du contenu.

La modélisation des références du document pose relativement peu de problème. Il importe à ce niveau d'être le plus exhaustif possible.

La modélisation des éléments contextuels peut être beaucoup plus difficile au sens surtout où il est difficile, peut-être impossible, de déterminer l'ensemble des contextes qui ont agi sur le document. Aussi, dans une optique pragmatique, il paraît plus raisonnable, dans un premier temps surtout, de reprendre les éléments contextuels cités dans le document sans en ajouter d'autres.

La modélisation du contenu est la partie la plus ardue. La détermination des différentes classes thématiques est simplifiée voire dictée par la classification définie précédemment. À chaque entrée de la classification devra correspondre une classe thématique distincte. En effet, comme cela a été dit, cette classification doit permettre une représentation de la structure de la base de données et elle doit lui correspondre. De plus, certains des modèles réalisés lors des expérimentations avaient été conçus dans l'optique d'une économie maximale de classes mais trop d'implicite était alors présent dans le modèle pour rendre celui-ci véritablement utilisable.

La détermination des liens découle en partie de la classification au sens où il faut faire figurer les liens hiérarchiques existant entre chaque classe. À ces liens il faut ajouter des liens de référence (plus particulièrement entre les classes permettant la représentation des références des documents et des éléments contextuels et entre la hiérarchie d'héritage et les références du document). Il est plus délicat de rendre compte des nombreux liens qui sont présents dans les documents de manière atypique (par exemple, lien entre un point thématique et un autre). Le fait de faire figurer ces liens de manière formalisée dans la structure même de la base de données conduit à alourdir considérablement celle-ci. Aussi, une solution satisfaisante est-elle de représenter ces liens sous la forme de liens hypertextuels. Cela implique l'utilisation d'un outil spécifique mais cela permet l'existence de liens souples et facilement réalisables sans encombrer la structure de la base de données. Ces liens faciliteront, en outre, la consultation.

Il importe également de déterminer les champs. En fait, aucune méthode générale n'a pu être dégagée à cet égard. Cette démarche demeure essentiellement empirique. Il s'agit de déterminer les champs «nécessaires» au niveau de chaque classe thématique. C'est-à-dire que pour une classe thématique donnée, il faut consulter un nombre représentatif de documents traitant ce point et déterminer quels sont les champs intéressants et nécessaires pour la description de ce point. Des tentatives de modélisation exhaustive réduites à un thème ont été affectées. Il paraît effectivement possible d'aboutir à une certaine exhaustivité dans la description d'un thème. Le problème est alors que, d'une part une description aussi fine et aussi précise ne présente pas forcément un intérêt et que, d'autre part, la description obtenue correspond précisément au corpus dépouillé et demanderait probablement des modifications chaque fois qu'un nouveau texte serait ajouté. De plus, une telle modélisation demande un lourd travail et aboutirait, si elle était conduite sur l'ensemble des thèmes, à une base difficilement gérable (la seule modélisation exhaustive du travail du week-end requiert 22 classes).

Il faut noter l'importance à ce niveau de la présence du texte intégral des documents. En effet, si une information est négligée par le concepteur du modèle, celle-ci sera néanmoins disponible pour les utilisateurs. Par ailleurs, il importe, lors de la détermination des champs, de penser le plus possible à utiliser des champs

dont le contenu sera extrait de listes de valeurs prédéfinies. En effet, l'objectif d'une base de données documentaire est de permettre de retrouver le plus rapidement et le plus facilement possible des textes ou des informations par le biais de requêtes. Or, les requêtes les plus efficaces sont celles effectuées sur des champs dont le contenu est normalisé. Il s'agit donc de reprendre les faits contenus dans le texte mais, si possible, de remplacer les segments textuels extraits par des termes normalisés en privilégiant, par exemple, les termes les plus souvent utilisés.

Interface utilisateur

Une fois, le modèle mis au point, il faut également s'intéresser à la conception de l'interface utilisateur. Cette phase est fondamentale non seulement en raison du caractère documentaire du produit mis en place mais également parce qu'il s'agit d'une démarche originale qui doit permettre à des utilisateurs d'interroger factuellement une base de données. En fait une nécessité s'impose: celle de disposer de différents moyens d'accès à l'information. Ces accès passent principalement par:

- des requêtes de type prédéfini, notamment pour l'interrogation des éléments référentiels, mais aussi des requêtes booléennes qui doivent être accessibles à différents niveaux plus ou moins spécifiques;

- la classification: cela offre d'une part à l'utilisateur une voie d'entrée simple dans l'application et cela lui permet, d'autre part, de visualiser la structure de la base de données. La visualisation lui montre quels thèmes ont été retenus. De ce fait, si l'utilisateur ne connaît pas ou peu les documents recensés, il a une présentation des points abordés dans ces documents; s'il les connaît bien, il peut juger des choix réalisés par les concepteurs de l'application;

- un ou des accès hypertextuels: beaucoup d'utilisateurs ont des besoins flous et mal définis. L'hypertexte permet de naviguer entre les textes. De plus, ces liens offrent un moyen de consultation bien adapté aux documents en texte intégral[6].

Dans le cadre de l'application réalisée sur les accords d'entreprise, l'utilisateur, après quelques pages de présentation, peut : interroger sur des données référentielles ou contextuelles par le biais de requêtes prédéfinies ou booléenes et il peut entrer dans la base par l'intermédiaire de la classification. Il choisit un thème et, depuis ce thème, il peut ou consulter des documents qui en traitent (il se trouve alors placé à l'endroit du texte où le sujet est décrit tout en ayant à sa disposition la version intégrale du document), ou il peut effectuer une requête sur les champs servant à le décrire factuellement. Il lui sera également possible d'accéder directement aux textes puis utiliser l'hypertexte (figure1).

Les résultats de recherche prennent principalement la forme des textes intégraux. L'utilisateur est alors placé à l'endroit du texte qui correspond à sa requête tout en ayant l'ensemble du texte à sa disposition. De plus, depuis le texte, l'utilisateur peut utiliser le réseau hypertextuel afin d'aller de la mention des références de l'accord, de l'entreprise, des textes de couverture de l'accord, de la mention des signataires ou des événements cités jusqu'à la fiche descriptive de ces éléments; d'un point de vue thématique vers un autre point classé au même niveau de la classification dans un autre texte; d'un segment textuel vers un autre texte ou un autre endroit du même texte si un lien atypique a été représenté.

Enfin, des fiches d'annotations ont été ajoutées, à titre expérimental, et sont accessibles par l'hypertexte. Par exemple, des fiches d'annotation précisent le statut des désaccords, des quasi-accords, des refus de négociation et des accords dérogatoires. Les liens sont ancrés à l'endroit du texte où les données sont mentionnées.

6. À propos de l'utilisation de la navigation hypertextuelle, voir : Mark Dunlop, *Multimedia information retrieval* (Glasgow: University of Glasgow, 1992) (Computing Science Research).

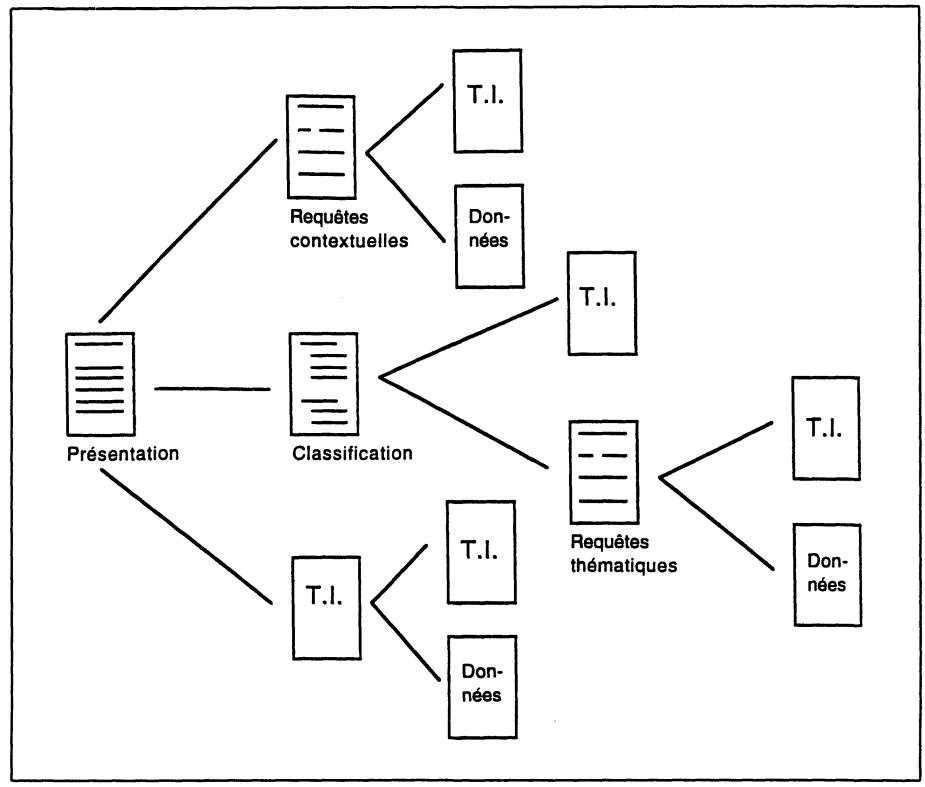

Figure 1 : Différents débuts de parcours possibles dans l'application

Il faut parler de l'importance, au titre de l'interface utilisateur, de la prise en compte d'une dimension analytique. En effet, lors des différentes enquêtes réalisées, cette demande était très importante: les utilisateurs souhaitaient, dans la plupart des cas, obtenir des réponses résultant d'analyses (dérive d'une mesure, évolution d'une autre, etc.) Or, une base de données ne peut que fournir les documents permettant aux utilisateurs de trouver eux-mêmes ces réponses. Bien sûr, il ne saurait être question de mettre au point un produit d'analyse mais il est peut-être possible d'ajouter des données analytiques. De plus, lors du dépouillement, actuellement manuel, de nombreuses constatations peuvent être effectuées par les personnes qui traitent les documents. Plutôt que de faire figurer de manière plus ou moins implicite ces informations dans la structure même de la base ou dans la manière dont sont dépouillés les textes, autant les mentionner dans des fiches d'annotation. En suivant le même principe, on peut envisager d'ajouter des liens hypertextuels qui permettraient, pour chaque point thématique de la classification, d'accéder à des textes présentant un caractère représentatif ou au contraire original voire atypique ou extrême. Ceci permettrait aux utilisateurs de pouvoir se faire une idée juste et plus représentative du contenu des textes.

Le fait d'afficher, plus clairement encore que cela n'est fait actuellement, une dimension analytique, et donc de parti pris, dans une telle application, ouvre également la possibilité d'une vision critique de l'application. Une base de données n'est pas objective. En montrant et en expliquant les choix subjectifs réalisés, on amorce la mise en place d'une critique et d'une discussion entre concepteurs et utilisateurs de l'application. Comme cela a souvent été dit, la métacommunication est sans doute le seul moyen de lever les ambiguïtés et les dysfonctionnements possibles de la communication[7].

Réalisation de la base

La réalisation de la base elle-même revêt un aspect essentiellement technique. Il semble indispensable pour une application de ce genre d'utiliser conjointement plusieurs types d'outils informatiques : un système de gestion de base de données relationnelles étendues ou objet puisqu'il est nécessaire de gérer des champs aux valeurs multiples, des hiérarchies d'héritage et des classes composites; un langage de marquage afin, au minimum de marquer les différents points thématiques dans le texte, voire de signaler de cette manière les contenus d'attributs qui, n'étant pas ou peu interrogeables, ne nécessitent pas une reprise dans la structure de la base elle-même, enfin, un système hypertexte.

La base de données ACCORD a été construite grâce à l'utilisation du logiciel Postgres[8] qui gère du relationnel étendu et de HTML (HyperText Markup Language) qui remplit à la fois les fonctions de langage de balisage et de gestionnaire de l'hypertexte, avec une interface ad hoc.

Conclusion

L'originalité principale de l'approche décrite ici est de découper les documents en classes thématiques puis de déterminer des champs spécifiques à chaque classe afin de permettre une description factuelle du contenu des documents.

La base de données ACCORD est, en outre, disponible dans Internet. En effet, il paraît pertinent d'utiliser cette possibilité pour le type d'applications décrites. L'utilisation d'Internet peut engendrer trois mutations importantes dans ce domaine à plus ou moins long terme.

D'abord, on peut imaginer, si le nombre de connectés continue à augmenter, que de plus en plus de documents pourront être communiqués directement sous forme électronique aux centres de documentation qui assurent leur traitement. Ensuite, cela devrait permettre le développement d'applications coûteuses par les centres de documentation. En effet, si les centres peuvent récupérer facilement toutes les données correspondant aux traitements courants auprès d'un centre serveur, cela dégagera des plages horaires libres pour les documentalistes. De plus, si ceux-ci mettent au point des applications lourdes et coûteuses mais intéressantes, il sera possible de permettre (contre paiement) l'accès à ces applications d'une manière simple. Enfin, la mise sur réseau d'applications de plus en plus nombreuses permettrait des connexions entre ces dernières. Par exemple, pour les accords d'entreprise, pourquoi ne pas imaginer de connecter la base ACCORD avec des bases traitant les textes juridiques, les conventions collectives ou les accords de branche? Les liens pourraient prendre la forme de liens HTML et un hypertexte permettrait facilement d'aller de l'accord vers le texte intégral de la convention collective qu'il cite, d'un accord de branche vers des exemples d'accords d'entreprise qui en constituent l'application, etc. On peut, dans la même optique, penser à relier la base ou plus spécifiquement les références des entreprises à des articles de presse qui parlent de la situation de ces entreprises. Dans le même esprit, on peut rêver à une mise en connexion des travaux analytiques sur le sujet et des accords qui leur correspondent. Ainsi, on pourrait pallier les inconvénients liés à la réalisation d'une application isolée et donc incomplète, génératrice de bruits et d'ambiguïtés. Dans la réalité, les accords ne sont qu'un maillon dans la chaîne de négociation et sont fortement dépendants de leurs contextes de signature.

En somme, on peut dire qu'afin de compenser l'aspect réducteur de la modélisation, il semble nécessaire de fournir le texte intégral des documents en résultats de recherche, de fournir une description simple de la structure de la base et de permettre à l'utilisateur de connaître les choix effectués par le concepteur afin de pouvoir les comprendre et donc, éventuellement, de les critiquer.

Les expérimentations entreprises ne constituent que la première étape d'un travail de plus longue haleine. Les éléments méthodologiques définis méritent d'être testés sur d'autres corpus de documents administratifs ou juridiques (par exemple sur des textes jurisprudentiels)

7. Communication ici n'est pas à prendre au sens de communication des informations par la base de données mais au sens où le processus documentaire est, dans son ensemble, un acte de communication voire de double communication entre l'auteur et les documentalistes d'une part, entre le documentaliste et les utilisateurs de l'application d'autre part.

8. Pour plus de détails sur le fonctionnement de Postgres, voir *The Postgres user manuel* (Berkeley: Postgres Group, Computer Science Div., Dept of EECS, University of California, 1994).

afin de mesurer leur validité, de les affiner tout en les généralisant. Enfin, il serait utile d'envisager l'utilisation d'autres outils informatiques (TAL, systèmes à base de connaissances, informations floues ou incomplètes) afin de juger de leur usages possibles.

D'un point de vue plus général, une contrainte semble se dégager: on ne peut pas interpréter les documents et on ne peut pas ne pas les interpréter. En effet, il paraît impossible de faire figurer explicitement dans la base de données l'interprétation des faits décrits en raison de l'ambiguïté résultant de toute interprétation. Mais, à l'inverse, on ne peut pas non plus faire figurer les faits, extraits de leurs contextes sans les pervertir (le texte a été écrit afin d'être interprété; il faut donc laisser aux utilisateurs les moyens de forger leur propre interprétation).

Enfin, il semble important de ne pas mésestimer les conséquences de la mise en place de telles applications. Comme cela a été dit, les phénomènes de rétroaction risquent d'être très importants à l'occasion des demandes des utilisateurs bien sûr, mais également en rapport avec les documents eux-mêmes. De nombreux documents légaux ou administratifs restent confidentiels. Cela influe sur leurs rédacteurs. Si ces documents pouvaient faire l'objet d'une exploitation pointue et exhaustive et donc d'une diffusion élargie, il est probable que cela provoquerait des réactions de la part des émetteurs. Réactions qui pourraient se traduire par exemple pour les accords d'entreprise par un déplacement des instances de décisions: les mesures conflictuelles ou marginales ne seraient plus prises dans le cadre des négociations annuelles obligatoires mais au sein d'instances plus discrètes comme les réunions du comité d'entreprise. Bien sûr les accords continueraient d'exister mais ils ne seraient plus alors que de pâles et sibyllins comptes rendus. Des conséquences plus graves pourraient également survenir allant notamment dans le sens d'une plus grande homogénéisation des décisions (si on peut facilement savoir ce qui se fait ailleurs...).

Sources consultées

Bertino, Elisa and Lorenzo Martino. 1991. Object-oriented database management systems: concepts and issues. *IEEE Computer* (April):33-47.

Rochon, Yves. 1992. SAGÉE: un développement informatique adapté aux besoins en gestion de l'information de la Direction des évaluations environnementales. *Documentation et bibliothèques* 38 (2):117-126.