L'intelligence de l'archiviste soutenue par l'intelligence artificielle : les systèmes experts et la linguistique computationnelle au service de la gestion des Archives

Autor(en): **Kellerhals-Maeder, Andreas**

Objekttyp: **Article**

Zeitschrift: **Geschichte und Informatik = Histoire et informatique**

Band (Jahr): **2 (1991)**

PDF erstellt am: **28.05.2024**

Persistenter Link: https://doi.org/10.5169/seals-4263

Nutzungsbedingungen

Die ETH-Bibliothek ist Anbieterin der digitalisierten Zeitschriften. Sie besitzt keine Urheberrechte an den Inhalten der Zeitschriften. Die Rechte liegen in der Regel bei den Herausgebern.
Die auf der Plattform e-periodica veröffentlichten Dokumente stehen für nicht-kommerzielle Zwecke in Lehre und Forschung sowie für die private Nutzung frei zur Verfügung. Einzelne Dateien oder Ausdrucke aus diesem Angebot können zusammen mit diesen Nutzungsbedingungen und den korrekten Herkunftsbezeichnungen weitergegeben werden.
Das Veröffentlichen von Bildern in Print- und Online-Publikationen ist nur mit vorheriger Genehmigung der Rechteinhaber erlaubt. Die systematische Speicherung von Teilen des elektronischen Angebots auf anderen Servern bedarf ebenfalls des schriftlichen Einverständnisses der Rechteinhaber.

Haftungsausschluss

Alle Angaben erfolgen ohne Gewähr für Vollständigkeit oder Richtigkeit. Es wird keine Haftung übernommen für Schäden durch die Verwendung von Informationen aus diesem Online-Angebot oder durch das Fehlen von Informationen. Dies gilt auch für Inhalte Dritter, die über dieses Angebot zugänglich sind.

Ein Dienst der *ETH-Bibliothek*
ETH Zürich, Rämistrasse 101, 8092 Zürich, Schweiz, www.library.ethz.ch

http://www.e-periodica.ch

L'intelligence de l'archiviste soutenue par l'intelligence artificielle: Les systèmes experts et la linguistique computationnelle au service de la gestion des Archives

Andreas Kellerhals-Maeder, Archives fédérales, Berne

1. Introduction

Les Archives fédérales, prévoyant les problèmes liés à l'introduction de l'informatique dans les offices de l'administration fédérale, participent depuis 18 mois, en collaboration avec l'Institut pour les études sémantiques et cognitives (ISSCO) à Genève et l'Istituto degli studi d'ingelligenzia artificiale (IDSIA) à Lugano, au programme national de recherche 23 «Artificial Intelligence and Robotics» avec un projet nommé «Specification and Prototyping of a System for the Intelligent Management of Information». Nous voulons développer des instruments qui nous aident à garder le contrôle logique sur les masses de dossiers électroniques qu'on peut aujourd'hui mémoriser facilement pour garantir à long terme l'utilisation des archives modernes. Le but principal est de construire de nouveaux instruments de recherche, adaptés à la structure des fonds d'Archives, utilisant des techniques d'intelligence artificielle qui, plus tard, en combinaison avec les instruments déjà existants ou d'autres instruments à développer encore, permettront de faciliter et d'améliorer la recherche archivistique.

Les Archives, faisant partie du large domaine de la documentation, se distinguent des bibliothèques et des centres de documentation par quelques éléments caractéristiques. En simplifiant un peu les choses on peut dire qu'elles comprennent toute sorte d'actes d'une provenance clairement définie, réunis automatiquement et organiquement. Les documents conservés aux Archives ont été produits dans l'administration au courant de l'exercice de ses fonctions. Ils reflètent l'accomplissement des tâches du pouvoir exécutif et - dans le contexte de leur production - ne sont pas destinés à un usage par un large public.

Dans l'administration les documents ne sont pas classés individuellement car le courant de chaque action administrative s'exprime normalement en plusieurs documents. Ce collectif de documents, produit pendant une action administrative, défini par une cote, est nommé dossier et constitue l'unité d'information de base des archives. Il contient plus d'informations que la somme des documents puisqu'il contient aussi des informations sur le contexte du déroulement de l'action administrative nécessaires à toute analyse administrative ou historique.

Les dossiers des archives courantes sont - dès leur production - classées

suivant des plans renonçant à un ordre par matière; au contraire, les archives courantes sont classées par actions administratives. Ces plans de classement servent tout d'abord à la gestion des archives courantes, mais chaque versement d'actes aux Archives est accompagné d'un bordereau de versement qui reprend - en partie - cette structure du plan de classement original.

Différence entre un classement par matières et un classement par actions - première limitation des fonds d'Archives à consulter:

Classement par matières	**Classement par actions**	**Office en question**
hélicoptère	Certifications de type	Office fédéral de l'aviation civile
	Autorisation exceptionnelle pour le héliskiing	Office fédéral de l'environnement, des forêts et du paysage

Le caractère organique des Archives a comme conséquence qu'on garde les archives courantes dans leur ordre original. Les actes d'un office restent réunis physiquement après leur versement aux Archives.

La provenance des actes structure les fonds d'Archives et les plans de classement informent l'archiviste de la structure de l'ordre original. La répartition des compétences sert de médiateur entre la structure d'organisation et la structure des plans de classements. La constitution et la législation fédérales fixent les compétences du pouvoir exécutif et définissent la répartition des compétences entre les divers offices. Chaque compétence est donc liée à un office. Ces compétences peuvent changer, disparaître ou réapparaître indépendamment de la structure d'organisation, de même que la structure d'organisation peut changer indépendamment de l'ordre des compétences. Le plan de classement reflète les compétences attribuées à un office et il change suivant les changements d'organisation ou de compétence. Un fichier systématique des compétences donne une vue d'ensemble et aide à organiser les archives, au moins sur le plan logique, selon une sorte d'ordre par matière.

La structure organique des fonds d'Archives détermine les possibilités et les limites de toute recherche. C'est donc suivant les trois axes de coordonnée archivistique - organisation, compétences, plans de classement - qu'on arrive à cerner le nombre des dossiers à consulter pour répondre à une question. La recherche concrète se fait en plusieurs étapes, utilisant les informations des différents instruments de recherche, informatisés depuis plus d'une dizaine d'années. Les questions suivantes se posent:

1. A quelle compétence ma question touche?
2. Quel office a été chargé de cette compétence?
3. Comment a-t-on classé les archives courantes dans les offices en question?
4. Quand ces dossiers ont-ils été versés aux Archives?

Le fichier systématique des compétences répond aux deux premières questions. Pour faciliter l'accès aux informations il est possible de commencer la recherche avec un mot-clé qui renvoie le chercheur à une ou plusieurs compétences à suivre et aussi aux offices, chargés de cette ou de ces compétences.

Les plans de classement en vigueur pendant une certaine période et les bordereaux de versement répondent aux questions trois et quatre.

Le système EDIBAR, en fonction aujourd'hui permet de localiser d'une manière très générale les fonds à consulter. La recherche plus détaillée se fait toujours sans l'aide de l'informatique. Un développement du système EDIBAR, permettant de continuer la recherche sur le niveau des plans de classement et des titres des dossiers, est prévu.

Dans le meilleur des cas, le chercheur trouve à la fin d'une recherche un ou un petit nombre de dossiers contenant les sources, qui lui permettent de répondre à sa question. Mais si - partant d'un mot-clé - les instruments de recherche nous indiquent plusieurs compétences, plusieurs offices chargés de chacune de ces compétences etc. on est vite forcé de préciser sa question pour limiter la recherche.

EXEMPLE D'UNE RECHERCHE: COMMENTAIRE

1. Le mot clé nous renvoie directement à deux compétences. Le choix est facile à faire.
2. Dix offices ont été chargés de cette compétence «Protection des oiseaux».
3. Nous limitons notre intérêt sur l'élaboration de la loi fédérale de 1925. Cette limitation temporelle nous permet de concentrer la recherche sur les dossiers provenant d'un seule office.
4. Dans un autre sens cette précision de la question nous permet de limiter notre recherche sur un petit nombre de positions du plan de classement de cet office
5. A la fin de la recherche le problème est évident: Les sources cherchées ne se trouvent pas entre les dossiers versés aux Archives par l'Inspectorat fédéral des forêts, chasse et pêche, mais entre les dossiers versés aux Archives par son successeur, l'Inspection fédérale des forêts.

EXEMPLE D'UNE RECHERCHE:

Je m'intéresse à l'histoire de la protection des oiseaux. Avez-vous des documents concernant ce sujet?

QUESTIONS	REPONSES DU SYSTEME DE RECHERCHE
1 A quelle compétence la protection des oiseaux touche?	
11 Quel mot-clé puis-je utiliser?	- **PROTECTION DES OISEAUX**
12 Ce mot-clé renvoie à quelle compétence?	- COMMISSION FEDERALE DE LA CHASSE (1946 -) - **CHASSE ET PROTECTION DES OISEAUX** (1874-)
2 Choix de la compétence «chasse et protection des oiseaux» et recherche des offices fédéraux chargés de cette compétence: Limitation de la recherche sur les dossiers concernant l'élaboration de la Loi fédérale sur la chasse et la protection des oiseaux (1925) -> Limitation temporelle. -> Limitation du nombre des offices en question	- INSPECTORAT FORESTIER (1875-79) - INSPECTORAT FEDERAL DES FORETS (1880-82) - [*DEPT DU COMMERCE ET DE L'AGRICULTURE*:] DIVISION: FORETS, CHASSE ET PECHE (1883-87) - [*DEPT DE L'INTERIEUR*:] DIVISION: FORETS, CHASSE ET PECHE (1888-96) - INSPECTORAT FEDERAL DES FORETS (1897-1904) - DIVISION: FORETS, CHASSE ET PECHE (1905-08) - **INSPECTORAT FEDERAL DES FORETS, CHASSE ET PECHE (1909-66) ==> E 3270** - INSPECTION FEDERALE DES FORETS (1967-79) - OFFICE FEDERAL DES FORETS (1979-85)
3 La limitation de la recherche renvoie le chercheur à une période de classement:	plans de classement valable: - **Plan A: jusqu'à 1966** - Plan B: de 1966 à 1985 **==> E 3270 (A)**
4 Quelle versement d'archives contient les dossiers concernant l'élaboration de la Loi fédérale?	bordereaux de versement suivant la structure du plan de classement: - **E 3270 (A) -/1**: 1 versement, ne contenant rien sur la protection des oiseaux. **E 3270 (B) -/1**: 19 versements, 5 versements contenant des dossiers sur la protections des oiseaux. Positions dans le plan de classement/Bulletin de versement: 2. Chasse et protection des oiseaux 1. Délibérations concernant la loi fédérale sur la chasse et la protection des oiseaux - Documents préliminaires concernant le message et avec un projet de loi - Préparation de la révision de la loi fédérale sur la chasse et la protection des oiseaux - Conseil des Etats - Conseil national - Commission de rédaction

2. Le Soutien de l'archiviste par l'Intelligence artificielle

Depuis longtemps on utilise dans l'administration le traitement électronique de l'information et depuis 1984 les Archives ont réglé l'archivage et l'utilisation de certains types de données informatisées, avant tout des données statistiques.

L'introduction des systèmes bureautiques ou des systèmes de traitement de texte de plus en plus répandue dans les différents offices force les Archives à s'occuper des problèmes concernant l'archivage d'un nouveau type de documents: les documents de texte lisibles par machine. Ce développement à prévoir, l'accroissement d'archives électroniques, forme l'arrière-plan de notre projet de recherche qui finira avec la construction d'un prototype d'un nouvel instrument de recherche, profitant de cette lisibilité par machine des documents versés aux Archives.

L'archivage de ces documents électroniques pose différents problèmes, dont deux doivent être mentionnés ici:

- Beaucoup d'informations, liées au dossier et au document physique, disparaîtront. Le dossier, aujourd'hui entité physique - une enveloppe comprenant plusieurs documents - et entité logique - position dans un plan de classement - ne sera plus qu'une entité logique. Sans une cote dans les documents, l'ensemble d'un dossier ne peut être que difficilement reconstruit ainsi que le déroulement de l'action administrative.
- D'autres informations disparaîtront parce qu'elles sont liées plutôt au logiciel qu'au document; je pense avant tout à une série d'informations concernant la présentation d'un document (caractères spéciaux: italiques etc.).

A ces pertes d'informations dans les archives électroniques s'oppose donc un avantage: la possibilité d'utiliser de nouvelles techniques de la gestion documentaire - l'analyse linguistique automatique.

Pour l'analyse automatique de la langue nous nous limitons au documents français car la langue française est plus facile à analyser (l'ordre des mots est plus fixe, ils n'existent pas autant de termes composés qu'en allemand). Les documents réels qui nous servent de base de travail ont été exportés d'un système bureautique en fonction dans un office fédéral. Ils ont un format ASCII et ils portent tous une cote, les ensembles de dossiers sont donc conservés. Les mots-clés attribués aux documents au moment de leur production ont été saisis séparément et se trouvent maintenant dans une banque de données. Nous ne les utilisons pas encore mais ils peuvent peut-être nous aider à comparer notre système de recherche aux possibilités

offertes par l'indexation libre, appliquée directement pendant le travail administratif.

STRUCTURE DU PROTOTYPE

DOCUMENT	-->	PARSING LE DOCUMENT	-->	RECONNAITRE LA STRUCTURE D'UN DOCUMENT	-->	
	-->	PARSING LE TEXT	-->	INDEXATION DU TEXTE	-->	
	-->	SYSTEME EXPERT	-->	ANALYSE DU DOSSIER	-->	
						RÉPONDRE à UNE QUESTION
	-->	SYSTEME EXPERT	-->	DEMANDES DE PRECISION	-->	
	-->	PARSING LA QUESTION	-->	INDEXATION DE LA QUESTION	-->	
RECHERCHE	-->	INTERFACE CHERCHEUR - ORDINATEUR	-->	FORMULATION DE LA DEMANDE	-->	

Les méthodes d'intelligence artificielle choisies sont d'une part des méthodes de systèmes experts, d'autre part des techniques de l'analyse automatique de la langue. La partie système expert se trouve encore dans un stade peu développé. Les éléments d'analyse de la langue sont plus élaborés.

2.1 *La partie système expert*

Un système expert contient normalement dans une base de connaissances des faits et des règles qui permettent de résoudre des problèmes d'une manière logique ou heuristique. Dans notre domaine c'est le savoir des archivistes sur les fonds d'Archives qui nous fournit les faits. Une grande partie de ces connaissances existe déjà dans une forme informatisée: ce sont les données qui ont été entrées dans les tables de la banque de données EDIBAR.

En analysant des recherches réelles et en discutant avec les archivistes nous sommes maintenant en train d'extraire le savoir, les règles des archivistes-experts sur le procédé de la recherche. A l'aide de ces règles nous construirons ensuite un modèle de recherche.

L'exemple de l'histoire de la protection des oiseaux nous a montré que - partant d'une question, traduit en mots-clés - nous avons la possibilité de chercher les compétences, les offices chargés de cette compétence et les versements d'actes de ces offices, contenant des dossiers qui peuvent nous aider à répondre à notre question. La première partie d'une telle recherche peut se faire à l'aide du système EDIBAR; la recherche plus détaillée doit toujours se faire à l'aide d'instruments de recherche traditionnels. L'archiviste-expert saute souvent les premiers pas d'une recherche - ayant ce savoir en tête - et consulte directement les bordereaux de versement.

Le modèle d'une telle recherche peut être illustré avec l'image de l'arbre: A un nombre restreint de mots-clés correspond un certain nombre de compétences - existant successivement ou simultanément - ainsi qu'à chaque compétence correspond un certain nombre d'offices - eux aussi existant succéssivement ou simultanément. On ne peut pas toujours faire un choix net entre plusieurs possibilités; parfois on peut seulement évaluer des priorités: analyser p.ex. en première ligne des aspects de politique intérieure ou de politique extérieure, analyser d'abord le processus de législation ou plutôt l'exécution d'une loi, ou encore on peut se limiter à l'analyse d'une certaine période de l'histoire.

Le but de la recherche ne peut donc pas être de trouver les dossiers justes ou faux, mais de proposer des dossier à consulter, même de limiter le nombre de dossiers proposés. Pour être sûr d'avoir consulté tous les dossiers nécessaires il faut, dans un second pas, de nouveau élargir le nombre de dossiers trouvés. Le système EDIBAR ne nous conduit normalement qu'à un fonds d'Archives concret. Mais: il ne faut jamais oublier que des documents peuvent être mal-classés; il faut donc savoir qu'on doit souvent aussi consulter des dossiers avec des titres plus généraux comme «Généralités» ou «Législation» etc. En plus EDIBAR ne nous indique jamais des fonds d'un caractère spécial, soit les dossiers des Secrétariats généraux, les dossiers personnels des Conseillers fédéraux ou de hauts fonctionnaires, les dossiers du Parlement. Le savoir, les expériences des archivistes nous permettent d'attribuer une certaine valeur à ces différents fonds spéciaux, une valeur globale ainsi qu'une valeur qui dépend du sujet de la recherche, une valeur qui détermine s'il faut consulter ce fonds spécial au début d'une recherche ou s'il a peu d'importance. De même, l'analyse linguistique nous fournit une table de matières des dossiers de caractère très général et nous permet aussi d'évaluer leur importance.

En somme, le système expert nous aidera à

- réunir et automatiser les différents pas d'une recherche,
- évaluer et proposer au chercheur des priorités de recherche,
- hiérarchiser les fonds d'Archives trouvés selon leur importance supposée,
- limiter le nombre de dossiers à consulter et, en même temps, en élargissant le champ de recherche, compléter la recherche.

2.2 *L'analyse automatique de la langue*

Si la partie système expert touche à la stratégie de la recherche, l'analyse automatique s'occupe directement des dossiers et des documents dans les Archives. Notre système commencera avec l'analyse du type et de la structure d'un document. Cela implique des questions comme:

- Comment peut-on segmenter un document? Quelles parties d'un document doivent être groupées ensemble pour l'analyse linguistique? (Direction fédérale des forêts, Aux inspections cantonales des forêts ...)
- Quelle fonction ont ces parties de document groupées ensemble? (p.ex. dans une lettre: l'entête, la référence, l'objet, le texte, la signature)
- Ces éléments fonctionels déterminent de quelle sorte de texte il s'agit? (p.ex. lettre, procès-verbal: une grammaire des sortes de texte nous décrit la forme possible d'une lettre, les éléments absolument nécessaires, les éléments facultatifs - toujours avec des indications sur un arrangement possible).
- Quels sont les parties contenant le plus d'informations dans un certain type de document? (p.ex. l'objet dans la lettre).

Les variations formelles que nous avons trouvées dans les documents électroniques nous ont forcés à limiter la reconnaissance du type de document à deux sortes de textes: la lettre et le procès-verbal.

Ayant reconnu les éléments formels et fonctionnels d'un document, on peut réduire l'analyse linguistique, soit l'analyse morphologique et l'analyse syntaxique, à des parties essentielles (p.ex. les indications concernant l'objet). Comme plusieurs documents forment un dossier, on peut gagner

plus d'informations ou des informations plus précises en comparant p.ex. les indications concernant l'objet de tous les documents formant un dossier. Cette analyse automatique des documents ou de parties comparables de documents fournira à la fin un profil de document ou un profil de dossier.

De l'autre côté la question d'un chercheur sera traitée d'une manière comparable afin d'obtenir pour elle aussi un profil, soit un descripteur très complexe.

Répondre à une question c'est comparer les profils de documents, mémorisés dans une banque de données, avec le profil de la question. Si cette comparaison ne livrait pas de résultat ou des résultats non satisfaisants, il nous reste toujours la possibilité de continuer la recherche avec des méthodes traditionelles en remontant aux textes originaux.

Le profil d'un document, le profil d'un dossier ainsi que le profil d'une question constituent tous une sorte d'index très complexe, établi automatiquement et incorporant des informations syntaxiques.

En incorporant des informations syntaxiques dans ce profil de document notre méthode se distingue des méthodes d'indexation traditionnelles et de leurs restrictions.

- Une indexation «à la main» ne peut s'effectuer aux Archives, en premier lieu parce que cette méthode n'est pas assez consistante, mais aussi parce qu'il nous manque des capacités de travail.

  Si les descripteurs peuvent être choisis librement, l'indexation est inconsistante car chaque personne chargée de l'indexation peut attribuer des descripteurs de son goût. Si les descripteurs sont choisis d'avance et fixés à un nombre limité et à une orthographe précise, la recherche elle-même est limité. Les sujets ne figurant pas dans la liste de descripteurs possibles, ne peuvent plus être recherchés. De plus, l'indexation à la main reste inconsistante soit arbitraire même avec un vocabulaire prédéfini. Les Archives devant prendre en compte les besoins des chercheurs futurs, difficiles à pronostiquer, l'établissement d'un instrument de recherche aussi limité paraît bien inutile.

- Les méthodes traditionnelles d'indexation automatique utilisées jusqu'aujourd'hui traitent les textes dans la plupart des cas selon des méthodes statistiques. Dans un premier pas, tous les mots vides comme prépositions, conjonctions etc. sont éliminés. Dans un deuxième pas, les mots restants sont réduits à leur forme de base, p.ex. les verbes à l'infinitif. Le troisième pas consiste à contracter les mots de différentes catégories mais avec une même racine, p.ex.

subventionner, subventionnement à subvention. Dans un quatrième pas, les descripteurs restants sont pondérés. P.ex. on juge importants les termes que l'on trouve fréquemment dans un seul document. En revanche, on estime sans importance les termes que l'on trouve fréquemment dans tous les documents.

Cette indexation automatique peut diminuer le travail manuel mais n'aide pas forcément à améliorer les résultats d'une recherche. Un index, se composant de termes employés dans le texte à indexer, force le chercheur de formuler sa question dans les mêmes termes; autrement il ne trouvera pas les documents cherchés.

- Au lieu de passer par un index, le chercheur peut chercher directement dans le texte les termes qu'il juge important. Le problème reste: Il faut chercher des termes qui ont été utilisés dans les documents. A long terme cela complique la recherche parce que l'usage de la langue change au cour du temps (p.ex. usine atomique -> usine nucléaire, coopération technique -> aide au développement).

- Un système basé sur un thésaurus semblerait être une solution possible. La composition d'un thésaurus est compliquée par le fait qu'il n'y a pas de langage administratif spécialisé auquel on pourrait se borner. Pour le moment nous y renonçons.

En conséquence, nous nous rallions à la création des profils de document servant de base pour les recherches. Grâce aux informations syntaxiques intégrées, nous améliorerons le résultat des recherches en comparaison à l'indexation traditionnelle. Je veux illustrer cela brièvement:

Par l'indexage traditionel une liste de descripteurs est attribuée à un document. Ces descripteurs peuvent être extraits automatiquement p.ex. d'une indication concernant l'objet. Ensuite la recherche se formule en combinaison booléenne de descripteurs. P.ex. la question

Existe-t-il des vues aériennes et un inventaire
des dégâts causés par les tempêtes?

se formulerait comme combinaison des descripteurs

dégât &
inventaire &
tempête &
vue_aérienne

Si un document portait l'indication concernant l'objet:

Vues aériennes et inventaire complet des dégats
causés par une tempête

l'indexation automatique nous fournirait la liste de descripteurs suivante:

dégât &
inventaire &
tempête &
vue_aérienne

La recherche se terminerait donc avec la présentation du document, portant l'indication concernant l'objet cité avant. Malheureusement cette analyse répondrait à la même question aussi en présentant un document portant l'indication concernant l'objet suivant:

Inventaire complet des vues aériennes de dégâts
causés par une tempête

Il est peu satisfaisant de trouver en réponse non seulement les documents concernant les inventaires des dégâts mais aussi les documents concernant les inventaires des vues aériennes. Avancer de l'index au profil de document, prendre en considération la structure syntaxique permettra d'éviter de telles réponses imprécises.

Les représentations des deux indications concernant l'objet utilisées dans l'exemple ne se distinguent pas en-haut au niveau de l'index mais elles se distinguent au niveau d'un profil de texte:

aérienne(v1),	aérienne(v1),
cause(d1,t1).	cause(d1,t1),
dégat(d1,x1),	dégat(d1,x1),
inventaire(i1,**d1**),	inventaire(i1,**v1**),
tempête(t1),	tempête(t1),
vue(v1,d1),	vue(v1,d1).

On peut faire encore d'autres exemples: L'objet

Droit de douane sur des fruits et légumes
produits dans le Marché commun et vendus en
Suisse.

est représenté dans un index traditionnel par la liste de descripteurs

droit_de_douane &
fruit &
légume &
Marché_commun &
production &
Suisse &
vente

Même si dans notre prototype l'analyse syntaxique se limite à l'analyse automatique des phrases nominales, elle permettra de distinguer par le profil de document la direction de la vente des fruits et des légumes (importation ou exportation de la Suisse).

3. Conclusion

Avec un instrument de recherche intelligent une recherche doit pouvoir se faire sans l'intermédiaire obligatoire d'un archiviste, sans supposition d'un savoir trop spécialisé sur les techniques de recherche du côté du chercheur. Un nouvel instrument de recherche devrait être capable de répondre en langue naturelle à des questions posées en langue naturelle.

Avant de pouvoir atteindre tous nos buts ambitieux il nous reste encore beaucoup de travail de recherche. Sortir du monde du laboratoire et appliquer les méthodes y développées sur des documents réels nous cause toujours certains problèmes car la structure de ces documents et leur forme ainsi que le langage écrit avec ses fautes d'orthographe et des constructions syntaxiques irrégulières sont d'une diversité gênante.

Malgré tout ces problèmes nous croyons avancer - lentement - dans la bonne direction.

Annexe

ANALYSE D'UN DOCUMENT: SEGMENTATION D'UN TEXTE

TEXTE ORIGINAL:

DIRECTION FEDERALE DES FORETS

3003 Berne
Laupenstrasse 20
Case postale 5662
Tél.: 031 / 61 80 91
n.réf.: 214 FN/FC

Aux
Inspections cantonales
des forêts

Berne, mai 1989

Lettre circulaire concernant la planification des investis
sements pour le subentionnement de câbles-grues mobiles

RESULTAT DE LA SEGMENTATION:

bloc: orig(1, 1), 29 x 1,
DIRECTION FEDERALE DES FORETS

bloc: orig(40, 1), 23 x 4,
Aux
Inspections cantonales
des forêts
---

bloc: orig(1, 3), 20 x 5,
3001 Berne
Laupenstrasse 20
Case postale 5662
Tél.: 031 / 61 80 91
n.réf.: 214 FN/FC

bloc: orig(40, 10), 15 x 1,
Berne, mai 1989

bloc: orig(1, 14), 58 x 3
Lettre circulaire concernant la planification des investis
sements pour le subentionnement de câbles-grues mobiles

EXTRAIT D'UNE GRAMMAIRE DE LA SORTE DE TEXTE «LETTRE»

textsorte	->	... | lettre | ...
lettre	->	lettre_normale | lettre_circulaire
lettre_normale	->	(entête_de_lettre) (adresse) datation + référence (sujet) (formule_de_salutation +) texte + transmission_de_compliments ...
entête_de_lettre	->	(box) | office | (expéditeur) | (adresse) | (date) | (référence)

..