CHAPITRE VIII

De la pertinence au parcours interprétatif : l'interface

Aperçu

L'interrogation du système se fait sous forme d'un texte, au lieu d'une équation de recherche (avec des opérateurs comme OU ou ET) ou au lieu de quelques mots-clés (mode usuel d'interrogation des moteurs de recherche Web). Le cas de la requête textuelle, véritablement innovant, voit ici ses atouts et ses points faibles identifiés.

En particulier, le texte apporte un contexte qui a le double avantage de préciser (désambiguïser) et d'élargir le sujet de la recherche. Mais l'utilisateur doit aussi avoir le moyen de spécifier les aspects du texte sur lesquels il attend plus particulièrement des réponses du système. Deux modes de surlignage sont proposés, l'un s'attachant plus à un choix terminologique précis (*surlignage horizontal*), l'autre à l'idée sous-jacente exprimée par un passage (*surlignage vertical*).

Les Sciences de l'Information montrent que la connaissance d'ensemble de la base documentaire est mobilisée lors d'une recherche : elle éclaire les résultats et éventuellement explique leur petit nombre, leur orientation particulière, etc. De même ici pour DECID, l'utilisateur peut demander un *aperçu de la prise en charge effective* du texte qu'il soumet par rapport à l'ensemble des profils : telle zone du texte a un fort potentiel de susciter des rapprochements, telle autre sort des domaines d'intérêts des chercheurs EDF. C'est en quelque sorte une relecture du texte soumis, une « première impression » générale, à l'éclairage des activités de la DER.

Ces indications sur le texte ont donné lieu à la mise au point d'un mode d'affichage adapté au texte dans toute sa longueur. Il concilie les points de vue local (le passage en train d'être lu) et global (le positionnement de ce passage dans la configuration d'ensemble du texte). Des *surlignages* traduisent des informations locales ; un *histogramme* donne, lui, un profil d'ensemble du relief que prend le texte. Associé à l'ascenseur, l'histogramme permet de se positionner directement aux points locaux, pics ou aux gouffres, détectés dès le niveau global.

L'affichage des résultats suppose d'abord une bonne description de chaque destinataire proposé : description générale (nom, rattachement) mais aussi informations par rapport à la demande : indication des unités (lexicales) qui ont le plus activement contribué au rapprochement, *projection* du texte de requête sur le texte de définition du profil (ou l'inverse, notamment si des considérations de confidentialité l'exigent). L'organisation en liste par ordre (total) de pertinence a été dénoncée, au profit d'une conception différentielle de la pertinence. Celle-ci est opérationnalisée par une arborescence sémantique, dynamiquement construite, dépliable progressivement. Sous cette forme, les résultats sont explorés en suivant des *pistes* ; puis dans un second temps c'est l'*originalité* de telle ou telle proposition qui pourra motiver de la retenir. A un ordre de pertinence figé et préétabli s'est substituée une navigation, qui accompagne la construction dynamique de l'interprétation et l'appropriation personnelle des suggestions du système. De multiples points de vue peuvent être adoptés, au gré de l'utilisateur : notamment, les résultats peuvent être reversés dans un tableur (ex. Excel), permettant, dans un environnement puissant et familier, encore d'autres tris et d'autres rapprochements.

Table des matières du Chapitre VIII

A. ACCÈS SÉMANTIQUE AUX BANQUES TEXTUELLES : LE TEXTE COMME POINT D'ACCÈS À D'AUTRES TEXTES

1. Les modes d'interrogation des bases textuelles

a) Par delà les mots-clés : la diversité des types de requêtes

Les modèles de la recherche documentaire assistée par ordinateur, et tout particulièrement le courant de l'*information retrieval*, conduisent insensiblement à se représenter les requêtes documentaires comme des ensembles de mots-clés, éventuellement pondérés. Cette notation uniforme oblitère la gamme très contrastée des modes d'interrogation.

La problématique du texte est comme évacuée. Du côté de la base, il s'agit d'informations, non plus de documents (et moins encore de textes). Quant à la requête, elle est vue soit comme un masque d'interrogation, une formule directement en prise avec les formalismes internes du traitement, soit comme une expression pleinement libre et spontanée apparentée à l'oral, et dont la description comme texte est loin d'être envisagée.

L'esquisse d'une petite typologie de requêtes, ci-après, va nous permettre de faire ressortir ensuite les caractéristiques d'une interrogation par un texte.

b) L'équation de recherche

C'est la forme courante pour l'interrogation professionnelle de bases documentaires, telles que Pascal ou INSPEC. Elle s'articule en opérandes (descripteurs contrôlés, codes de classement, chaînes de caractères avec éventuellement des jokers) et opérateurs (qualification par un champ, et/ou/mixte (CUMUL chez TOPIC), négation, proximité), ou autrement dit elle a son lexique et sa syntaxe.

Exemple de requête :

```
(   ( CC=c72? OR CC=6180n OR CC=c6130d OR linguist?/DE )
    AND
    (       ( AU=salton? OR AU=biber? )
            OR
            (       ( dissemination?/TI,AB OR selective (W) diffusion? )
                    AND
                    ( information?/DE or document?/DE or full (W) text?/DE,ID )
            )
            OR
            ( routing (2N) information? or routing (2N) document? )
    )
)
NOT ( speech?/DE OR image?/DE )
```

L'équation traduit une variété de motifs que l'on recherche dans les notices bibliographiques. Son extrapolation au cadre du texte intégral est un déplacement de signification majeur, même si les notices bibliographiques comportent généralement un résumé rédigé (de l'ordre d'un paragraphe). Les contraintes de présence et d'absence, de choix de vocabulaire, de co-présence et de proximité n'ont pas la même incidence sur des résumés d'une base donnée, ou sur des textes dans toute leur étendue et avec les spécificités de leurs genres.

L'équation s'interprète originellement comme une famille de combinaisons logiques sur des champs typés (auteur, titre, descripteurs, résumé, etc.). Or les textes ne se laissent pas penser en termes de logique booléenne. La disjonction (non exclusive) est alors utilisée pour décrire des effets paradigmatiques, la conjonction et les contraintes de proximité pour la dimension syntagmatique et l'interaction contextuelle. Les opérateurs logiques ne sortent pas indemnes de cette réinterprétation linguistique : beaucoup voient leur rôle possible s'affaiblir, certains disparaissent, des opérateurs de type nouveau sont élaborés[1].

L'équation de recherche se caractérise par son mode de fonctionnement *explicite* et *déterministe*. C'est là tout à la fois sa force et sa faiblesse. Grâce à sa clarté (au moins pour l'utilisateur expérimenté), la recherche est menée au plus fin, en connaissant et en exploitant au mieux toutes les possibilités de la machine. L'utilisateur (expérimenté...) contrôle[2] le déroulement de la recherche, s'appuie sur les mécanismes mis en œuvre pour déployer une tactique adéquate, et a les moyens de donner du sens aux résultats, car il sait à quels critères obéit leur obtention[3]. En revanche, on sait bien que les chemins de l'expressivité de la langue ne passent pas par l'explicitation littérale et systématique. Qui a pratiqué les équations de recherche sur du texte intégral connaît bien leurs limites : on pense à tel mot, et c'est en fait tel autre qui est utilisé ; on prévoit telle combinaison de termes, et ce n'est pas exactement ce qui apparaît dans les textes recherchés. La disjonction est trop lâche, et l'on se bat à détailler un inventaire fatalement incomplet. La conjonction pèche par l'excès inverse –trop de rigidité– et oblige à s'en tenir à une contextualité maigrelette : une évocation trop riche, au lieu d'ouvrir les possibilités d'investigation, les anéantit (sélection nulle ou infime par rapport au potentiel « enfoui » dans la base).

L'accès par équation de recherche peut être très efficace pour retrouver un document précis (dont on a en mémoire une caractéristique discriminante). Le succès est en revanche très loin d'être assuré quand il faut ne laisser échapper aucun document concernant un sujet donné. Ce que l'on peut résumer par : bonne précision, mais silence rémanent. De surcroît, lorsque la recherche ne se limite pas à retrouver une information (dans quelque document qu'elle soit) ou un document bien défini, le balayage des résultats est long et fastidieux, puisque les références proposées ne sont pas triées (sinon par une critère externe : ordre inversement chronologique, nom d'auteur) (Turtle 1994).

c) La requête en Langage Naturel

La vogue de la communication homme-machine a mis à l'honneur une forme d'interrogation mimant[4] le dialogue humain. Dans le contexte des systèmes documentaires, la requête prend la forme de l'expression succincte d'un besoin d'information. En théorie, c'est une ou deux phrases, du type :

[1] Le lecteur intéressé par cette question trouvera, dans le chapitre consacré à la construction d'unités, une étude détaillée et approfondie des opérateurs utilisés dans les systèmes documentaires et la discussion de leur transposition dans un contexte linguistique et textuel. Le système TOPIC, techniquement réputé et mondialement diffusé, est pris comme référence.

[2] Il s'agit du contrôle tel que le définit Bruno BACHIMONT (Bachimont 1992), c'est-à-dire la possibilité pour l'utilisateur de se construire une représentation du traitement effectué, et ainsi de comprendre ce traitement et d'être en mesure de le maîtriser (adaptation, réglage de paramètres, etc.).

[3] La tactique se conforme au fonctionnement de la requête booléenne, qui sélectionne un « paquet » de références à dépouiller : classiquement, on procède par enrichissement progressif de la requête pour atteindre, par réductions successives, un volume de résultats acceptable. Aux premières étapes de la recherche, l'ajustement se fait moins par *relevance feedback* (avis sur la pertinence des propositions du système) que par *magnitude feedback* (on se fait simplement une idée sur le nombre de références rapportées : c'est trop, ou trop peu). (Saracevic & *al*. 1991)
Les tactiques de conduite de la recherche (par ajustement progressif de la requête) font l'objet de tentatives de formalisation. Ainsi, (Denos 1997, §IV.4.2, p. 144 sq.) forge une typologie des situations problématiques qui suscitent une correction de la requête (pénurie, non-pertinence, non-discrimination, non-focalisation), et formalise des principes de correction associés, en vue d'une reformulation automatique de la requête (*ibid*., §IV.4.3, p. 149 sq.).

[4] Mimer... ou singer ? L'acharnement à dissimuler le fonctionnement (calculatoire) et à déguiser la machine en « interlocuteur » n'a pas toutes les vertus escomptées, qu'il s'agisse d'efficacité (qualité des résultats) ou d'ergonomie (maîtrise du déroulement du traitement et exploitabilité des résultats).

> Je cherche des informations sur ..., et plus précisément concernant ... ; je ne suis pas intéressé par les documents au sujet de ...

Que peut-on noter de caractéristique pour ces requêtes ? Le lexique consiste en des formules de construction (*je cherche*, *je suis intéressé par*, *avez-vous des documents sur*), et des termes de niveau « méta », qui désignent des domaines, des disciplines. Les relations entre ces termes sont décisives. En particulier, la négation doit être prise en compte.

S'insérant dans le cadre d'une pratique –l'interrogation de tel fonds documentaire par telle population d'utilisateurs–, les requêtes pourraient être linguistiquement décrites comme un genre. Une étude d'un tel corpus de requêtes est nécessaire pour fonder ensuite les procédures d'analyse, celles qui « traduisent » le motif de la recherche en termes de représentation interne pour la machine (Cousins 1992). Il faut en attendre la mise en évidence de particularités morphosyntaxiques (à savoir traiter) :

> [Une étude des comportements des utilisateurs face à un système de recherche en langage naturel a abouti aux observations suivantes :]
>
> - les énoncés tapés par les utilisateurs paraissent souvent pauvres, mal formés et non grammaticaux ;
>
> - on constate de nombreuses fautes d'orthographe, souvent dues à une insuffisante maîtrise de la frappe au clavier [...] ;
>
> - il y a très peu d'utilisation des tournures interrogatives. Les sujets formulent leur requête sous forme de mot clé ou de syntagme nominal complexe. Les entrées tapées par les sujets contiennent une proportion importante de phrases sans verbes, ce qui pose un problème certain pour une grammaire centrée sur le verbe. Le traitement des énoncés fragmentaires est un problème important qu'il faudra étudier lors de la construction de la grammaire ;
>
> - l'usage des pronoms est peu répandu, ce qui est réconfortant quand on pense à la difficulté d'établir des coréférences en traitement automatique de la langue naturelle ;
>
> - les structures syntaxiques sont plutôt simples et leurs variations limitées.
>
> (Polity 1994, p. 142)

Au plan sémantique, une exploration approfondie des formulations ferait par exemple ressortir des schémas d'expression du type *discipline + descripteur spécifiant + détermination spatio-temporelle*, comme l'observe Denise Malrieu sur des demandes de bibliographie par correspondance (Malrieu 1992).

Une analyse purement lexicale montre ici ses limites ; on est plus volontiers dans le domaine de l'analyse syntaxique au niveau de la proposition. Automatiser ces traitements syntaxiques est très délicat (Polity 1994) : à notre connaissance, il n'y a guère de systèmes opérationnels allant au-delà d'une reconnaissance des expressions[5].

Aleth-IR par exemple transforme la requête en langage naturel en équation booléenne, par des mécanismes comme ceux-ci : les groupes nominaux reconnus sont traduits par l'application d'un opérateur de proximité sur leurs composantes ; les variantes identifiées par lemmatisation sont regroupées par un opérateur de disjonction (OU) ou factorisées par une troncature.

Si le système épargne effectivement à l'utilisateur le recours au formalisme booléen, on peut douter du gain en matière d'efficacité : le calcul n'est en mesure que de produire une équation relativement élémentaire (que la plupart des utilisateurs auraient su composer directement), et potentiellement assez éloignée de la signification voulue (erreurs d'analyse et non superposition de la

[5] L'aide générale de SPIRIT-W3 (`http://www-dist.cea.fr/ext/spirit/aide/aide.html`) a le mérite d'expliquer très clairement que le traitement est plus efficace si l'on s'abstient de formuler la requête comme une question, et que l'on se contente des termes désignant le thème de recherche :
« L'utilisation d'une syntaxe correcte du français est souhaitable pour obtenir le meilleur résultat car le système se base sur la syntaxe pour déterminer les mots composés et lever les ambiguïtés. *Mais attention le système n'interprète pas la syntaxe des questions*, il est donc inutile (sinon nuisible) de poser une requête de la manière suivante : 'y a-t-il quelque chose sur le traitement des déchets nucléaires ?'
On se contentera de la formulation suivante : 'le traitement des déchets nucléaires'. »
Ce qui à notre avis ne retire rien à l'ergonomie du système, au contraire.

syntaxe de la langue et de celle du langage booléen). La tentation est alors de normaliser la formulation de la requête (grammaire canonique, vocabulaire connu du système)... et de retrouver, de façon sournoise, le genre de contraintes que l'on voulait éviter :

> si l'on peut imaginer qu'un tel système puisse gérer un minimum d'irrégularités ou d'écarts par rapport à la syntaxe de la langue (exemple de question susceptible d'être traitée : *Quels sont les documents relatifs aux mémoires informatiques possédez-vous ?*), une requête de type « structure énumérative » pourra difficilement être interprétée [pour obtenir une indexation structurée] du fait de l'impossibilité d'établir des règles adéquates (exemple : *œil physiologie anatomie pathologie chirurgie*). L'utilisateur désirant effectuer des requêtes en langage naturel devra faire un minimum d'effort dans leur rédaction. (Coret & *al*. 1994, p. 153)

d) La requête pour moteur de recherche Web

Les observations montrent qu'elle consiste dans la très grande majorité des cas en un ou deux mots, qui évoquent le thème de la recherche (Clarke, Coormack, Tudhope 1997) (Pinkerton 1994).

Les moteurs proposent subsidiairement l'expression de requêtes élaborées (par opposition à la simple juxtaposition de mots) qui articulent les termes avec des opérateurs, reconnaissent les expressions composées, modulent l'importance des composants de la requête (obligatoire / facultatif).

Requête simple :	`recipe oatmeal raisin cookies`	(AltaVista)
Requêtes élaborées :	`recipe cookie +oatmeal -raisin`	(AltaVista)
	`« auto parts » BMW`	(HotBot)
	`cat AND NOT kitten`	(Excite)

La requête courante (95 %) est pauvre en contexte, ce qui, d'une manière générale, est connu pour détériorer très sensiblement la qualité des résultats (Heine 1995). D'où les aides à la reformulation et à l'enrichissement, comme LiveTopics pour AltaVista (option *refine*, voir FIGURE 1 et FIGURE 2). Le principe de LiveTopics (Bourdoncle 1997) est de suggérer du vocabulaire à ajouter à la recherche, compte-tenu des mots indiqués dans la requête et des textes de l'espace de recherche qui pourraient être sélectionnés par la requête. Les différentes « interprétations » des mots-clés de la requête dans le contexte de la base apparaissent via la mise en évidence de familles de documents qui concernent différents domaines[6].

On s'accorde sur le fait que ces requêtes sont « plates » : tous les termes semblent s'équivaloir. Pourtant, l'ordre des termes n'est peut-être pas fortuit[7] : n'aurait-on pas tendance à commencer par ce qui est au cœur de ses préoccupations, ce qui est le plus représentatif de la thématique, puis à compléter par des termes périphériques ou plus spécialisés ? à décrire successivement différents aspects, groupant ainsi les termes sémantiquement ? Ces aspects ne sont pas étudiés à notre connaissance : mais peut-être n'obtiendrait-on pas des régularités suffisantes pour en tirer des procédures opératoires. D'ailleurs, la brièveté des requêtes ne permet pas, dans bien des cas, d'y modeler un quelconque relief !

[6] A propos de LiveTopics, on ne peut que déplorer l'intégration totalement inefficace de ce module dans AltaVista. En effet, LiveTopics fait des suggestions intéressantes, mais les sélectionner est traduit par l'introduction du mot dans la requête *avec une présence obligatoire*. C'est évidemment beaucoup trop contraignant, surtout s'agissant d'apporter des indications de contexte. La conséquence est, qu'après un enrichissement de la sorte, la requête non seulement est déportée de son ancrage initial, mais encore ne permet plus de retrouver aucune page...

[7] Dans le cadre d'une pratique d'indexation, Suzanne BERTRAND-GASTALDY observe que l'ordre des descripteurs est conventionnel, et donc fortement significatif. Les valeurs sémantiques liées à l'ordre des descripteurs varient d'un domaine à l'autre (ici : droit pénal, procédure civile, assurances-responsabilité), chaque liste de descripteur formant comme un genre à part entière. (Bertrand-Gastaldy 1996)

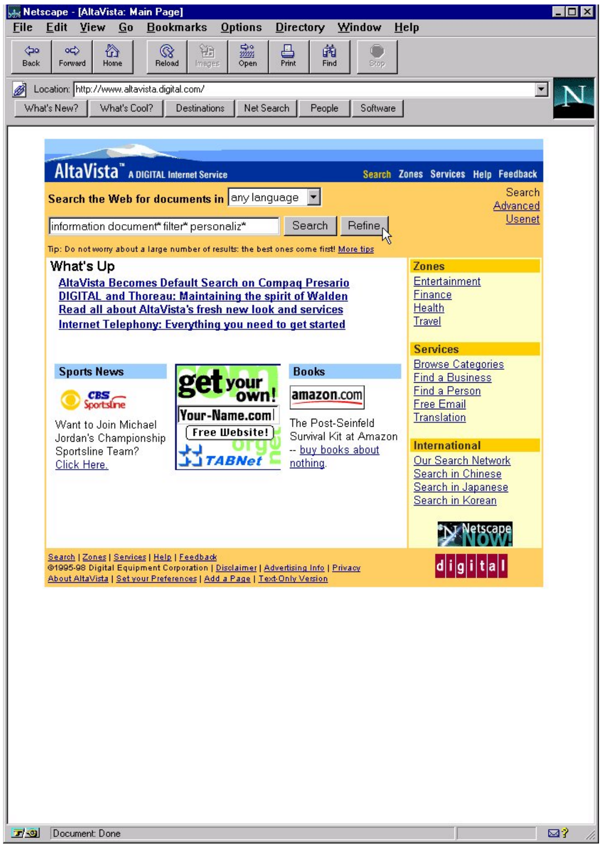

FIGURE 1 : **AltaVista, lancement de l'option *Refine* sur une requête.**

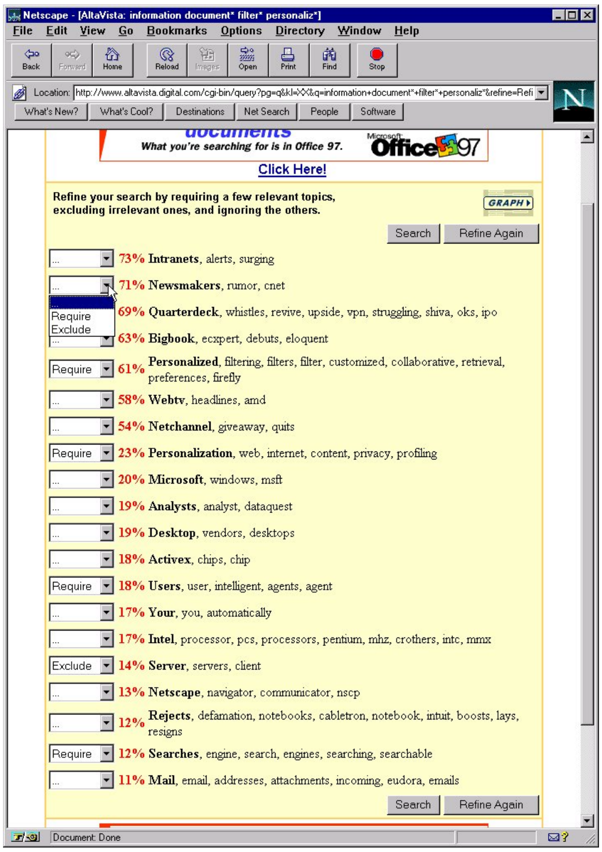

FIGURE 2 : **AltaVista, contextes des mots de la requête, calculés par *Refine*.**

Pour permettre cependant l'expression du rôle plus ou moins important que l'on veut voir jouer à chaque mot de la requête, plusieurs modes de formulation existent (Pincemin, Lemesle 1996, §3.3) :

- *pondérations numériques* (Excite) : ce mécanisme semble simple, mais en fait sa signification est totalement opaque à l'utilisateur. Le seul effet dont on puisse se douter intuitivement, c'est qu'à valeur de pondération croissante correspond l'indication d'une importance croissante. En revanche, rien ne permet de savoir comment choisir les valeurs, dans l'absolu et les unes par rapport aux autres.
- *présence obligatoire* : alors que le calcul permet d'explorer toutes les combinaisons d'occurrences des mots de la requête, indiquer la présence obligatoire de certains mots leur donne un rôle central et décisif. Mais juger qu'un élément de la requête est plus important qu'un autre n'imposerait pas toujours que cet élément doivent figurer explicitement et exactement sous la même forme dans tous les documents pertinents.
- *influence sur le classement des résultats* (AltaVista) : l'indication d'importance ne joue pas sur la recherche elle-même (les mêmes documents sont trouvés), mais sur l'ordre de présentation des documents. De fait, si les documents sont extrêmement nombreux (ce qui est souvent le cas dans les interrogations sur Internet), c'est dans la pratique à peu près la même chose : les documents qui ne comportent pas ces termes sont classés trop loin pour être vus. Les reproches que l'on pourrait faire à cette fonctionnalité est qu'elle est peu nuancée (un terme est important ou ne l'est pas, il n'y a pas de degré d'importance, de préséance d'un terme important sur un autre également important), et que son rôle effectif dans le classement des résultats n'est pas clair.

Dans tous les cas, l'indication des importances relatives des termes oblige l'utilisateur à jouer à l'apprenti sorcier, car il ne lui est pas donné les moyens de maîtriser la signification des choix faits et leur impact au niveau des résultats.

Une autre approche, assez originale, consiste à estimer la fréquence « implicite » de chaque terme (jouant ensuite sur sa pondération), en estimant sa fréquence moyenne quand il figure dans un *texte* (Kwok 1996) : c'est finalement une sorte de recontextualisation artificielle.

e) La requête par l'exemple : le texte comme point d'entrée

Le désir point, au détour de la lassitude des mots-clés et du constat de leur facticité, d'un mode d'expression de la recherche plus proche des textes :

> [In the context of an information retrieval task, conducted by a search intermediary in large operational online systems like DIALOG,] one user made a suggestion for the future system : « maybe if we could give the system a broader discussion of problem, instead of giving simple words and combinations of those words. » (Su 1993, p. 97)

Texte, ou extrait ?

Des systèmes existent qui permettent l'interrogation par *du* texte, mais non vraiment par *un* texte :

> Par souci de convivialité, plutôt que d'expliciter une requête, l'utilisateur choisit dans une page un fragment de texte correspondant à un concept. Cette sélection est transformée en requête par analyse des mots clés. Une fois cette requête formulée, le système calcule quelles sont toutes les réponses potentielles et fournit la liste ordonnée par pertinence décroissante des passages [paragraphes] similaires à sa sélection. (Betaille, Massotte, Joubert 1998, pp. 141)

Le système SPIRIT donne la possibilité de sélectionner une zone d'un document préalablement trouvé et de l'utiliser comme nouvelle requête. Aucune indication précise n'est donnée sur la taille ou la constitution de la dite zone, mais plusieurs indices convergents indiquent qu'elle est de l'ordre du syntagme (long), éventuellement du paragraphe, mais pas d'un texte d'une page ou plus. Dans l'exemple donné dans (Fluhr 1994), l'extrait sélectionné pour relancer la recherche est une partie de phrase, de une à deux lignes de long. Des essais effectués sur SPIRIT-W3[8] montrent que la limite supérieure de la taille de la zone requête est de l'ordre de quelques paragraphes, au delà la requête est tronquée.

Les moteurs de recherche Web ne donnent guère de précision sur la longueur de la requête qu'on peut mettre dans la fenêtre prévue, mais expérimentalement on observe qu'à partir d'une centaine de caractères la requête est tronquée. Techniquement, le passage par cgi[9] limite les données retransmises au serveur (mais le seuil est beaucoup plus haut : il correspond à une dizaine de pages). En revanche, certains moteurs proposent de relancer la recherche en cours en demandant des pages proches d'une des pages trouvées (« *More like this* » de Excite par exemple).

Le texte comme source de mots-clés

Le moteur de recherche commercialisé par la société canadienne *Fulcrum* présente ainsi sa fonction *Intuitive Searching*, qui s'apparente dans la forme à une requête textuelle :

> Fulcrum's innovative Intuitive Searching feature provides an automatic query generation capability. Intuitive Searching allows a complete document or text fragment to be used directly as a query. Documents with similar content are located and ranked according their relevance. [...]
>
> Intuitive Searching also supports a natural language-like capability. A user can enter any words representing the subject matter of interest and this text fragment is used directly as a query.
>
> Intuitive Searching's sophisticated algorithms are based on sound, established academic research. A set of terms are generated from the terms in the query text. Search term selection is influenced by the frequency with which terms occur in the query text and in the tables being searched, as well as by various weighting factors.
>
> (Présentation commerciale de *Fulcrum SearchServer* v2.0, octobre 1994, 18 pages).

Nous n'avons malheureusement pas eu l'occasion d'en savoir plus sur cette fonctionnalité (démonstration, expérimentation, exemple, ou détail de la technique). Le fait qu'il n'y ait

[8] SPIRIT-W3 est accessible sur le site du CEA (`http://www-dist.cea.fr`) pour la consultation des catalogues des bibliothèques.

[9] Cgi : *common gateway interface*.

apparemment pas reprise de tous les termes trouvés dans l'extrait, mais construction d'un ensemble de termes représentatif, est intéressant. Les points faibles pourraient néanmoins être : *(i)* une sélection uniquement fondée sur la fréquence (dans l'extrait) et la distribution (sur la base), sans prise en compte de l'organisation interne du texte, notamment de zones de localité, et de la manière dont les termes sélectionnés couvrent le passage ; *(ii)* des unités d'indexation limitées à des chaînes de caractères simples (mots graphiques, découpés par les blancs typographiques), sans prendre en compte les interrelations syntagmatiques entre ces unités (notamment, les expressions composées). Quoiqu'il en soit, c'est un exemple de réalisation tout à fait comparable à la fonction initialement offerte dans DECID.

Un mécanisme analogue serait offert par le système Callable Personal Librarian (CPL) de la société Personal Library Software (PLS), à en juger par les passages suivants :

> PLS statistical search methodology can accomodate natural-language queries. In natural-language mode, you can enter a query, such as « negotiations over most-favored-nation status for China » and get back more than just the documents literally containing that phrase. In this example, words like « for » and « over » are dropped from the search, as stop words, but the others are all given some weighting if found. The PLS engine can accept phrases, questions, (« When can I plant tulips in Southern California ? ») or samples of text as search requests. Using the query by example function, located documents or paragraphs and phrases from them that are particularly « on target » can be fed back to the search engine with the instruction, in effect, to « find me more like this one. »
>
> Entering into the calculation of a relevance score for each document are variables such as :
>
> - the number of times each query term is found in a document ;
> - the number of different search terms that appear in a document ;
> - how close the « hit » terms found are to the beginning of each document ;
> - how closely together different search terms appear in the text ;
> - how closely the order in which the terms appear in a document approximates the order in which they are presented in the query ;
> - how rare are the terms matching those in the query, as indicated by frequency of appearance in document collection. (Rarer terms are more useful in indicating what a document is about, and receive a higher weightinng than higher frequency ones.)
>
> (Banet 1996, p. 5)

> *Query By Example* - CPL can search for all statistically significant words within a specified record. A user who is interested in the subject of a particular document can employ this feature to find additional records with similar content.
>
> (Documentation commerciale PLS / CPL)

Les critères pour le calcul de la pondération et de la similarité prennent en compte des effets syntagmatiques (localisations à l'intérieur du document). On peut penser que ces critères entre en jeu aussi pour un document en tant que requête, puisqu'il est soumis par une fonctionnalité spéciale (« the query by example function », « find me more like this one »). Le traitement serait donc peut-être un peu plus évolué que dans l'exemple précédent. Toutefois, on en reste clairement à la notion de mots, sans réellement construire d'unités par delà la chaîne de caractères simple.

2. Caractéristiques de la requête textuelle - Discussion

a) Enthousiasme et réalisme

La mise au point d'un système reposant entièrement sur l'interrogation par texte, comme DECID, est d'abord l'occasion de percevoir les avantages techniques et scientifiques de ce mode d'interrogation. Il est clair que le texte a des atouts majeurs par rapport aux autres types d'interrogation, à la fois sur les points de comparaison et d'évaluation les plus classiques (bruit, silence) et sur d'autres aspects, moins formalisés peut-être mais non moins importants, notamment en lien avec l'ergonomie (qui est tout autant le confort d'utilisation que la juste maîtrise par l'utilisateur du déroulement du traitement).

L'expérimentation met au jour les possibles difficultés pratiques. Les côtés très innovants sont également susceptibles de malentendus. La première tâche est de prendre acte de ces retours et de les recueillir avec attention. La seconde est d'évaluer leur pertinence et leur portée : dans quelle mesure révèlent-ils des aspects encore inaperçus, et des limitations, de ce mode d'interrogation ? La troisième tâche est de redéfinir la mise en œuvre de l'application, en intégrant ces retours : revenir sur certains choix techniques, inventer de nouvelles fonctionnalités adaptées, ajuster la communication sur le système et l'aide aux utilisateurs.

Au stade où nous en sommes, l'interrogation par le texte n'a rien perdu de son attrait ; elle a gagné en souplesse et en maturité.

b) Pas d'étape de formulation

L'économie d'un langage de requête ésotérique et spécifique

Devoir passer par un langage d'interrogation formel et spécifique introduit au moins trois difficultés :

- apprentissage du langage : syntaxe et éventuellement référentiel des termes,
- habileté à l'utiliser : arriver à exprimer avec justesse ce que l'on veut, à maîtriser la signification opératoire de chaque terme, et l'impact effectif de chaque relation,
- mémorisation : risque de confusion avec d'autres langages spécifiques à d'autres bases documentaires (pour l'utilisateur occasionnel ou/et familier avec plusieurs systèmes de recherche).

> Problème des outils informatiques : pour l'instant absolument pas au point ergonomiquement pour être utilisés librement par les chercheurs. Pour les bases de données externes, il faut des spécialistes qui sachent poser des requêtes dans le langage du serveur. (Merle, Fradin 1994, §7.1, p. 39)

> Dans les bases de données externes, il y a sûrement de l'or, mais comme on ne peut y avoir accès... (connaissance préalable de la structure de la base et de ses contenus pour décider d'aller y chercher quelque chose, complexité et diversité des langages de requête) (Merle, Fradin 1994, §7.2, p. 42)

> Notre information [vient en dernier lieu] [...] de quelques serveurs externes [...] [comme] data Star, Dialog, Questel, mais compte tenu de l'extrême complexité à les manipuler, ils ne représentent qu'une information d'appoint. Ce qui prend aussi beaucoup de temps, c'est de connaître ce que l'on peut trouver sur ces serveurs. (Merle, Fradin 1994, §7.2, p. 43)

En soumettant un texte, tel quel, au système, il n'y a pas à connaître le lexique, la syntaxe, l'organisation du référentiel, et l'usage effectif d'un langage documentaire. Car chaque système peut avoir ses conventions : par exemple, recouvrement ou complémentarité des différents champs du document secondaire, rôles réservés à chacun[10]. L'utilisateur n'est plus perplexe face au choix des mots-clés, et n'a pas à procéder à tâtons pour parcourir les combinatoires de ses mots-clés qui soient à la fois productives et sélectives.

Cette minimisation des contraintes formelles ne peut être que favorable à la fraîcheur de l'information, puisqu'il n'y a pas à attendre de prendre le temps, pour faire le travail de formulation, ou à mobiliser un ou plusieurs intermédiaires (aide à l'utilisation du système, ou retransmission indirecte de l'information par la voie hiérarchique par exemple).

L'effort d'expression

L'expression d'un besoin d'information ou la caractérisation d'un document n'ont rien d'évident[11]. Avant même de savoir comment trouver ce que l'on cherche, il faudrait savoir ce que l'on

---

[10] Il est important de savoir, pour procéder à l'interrogation, que par exemple :
- le résumé utilise un vocabulaire différent de celui du titre,
- les noms de produits sont dans des mots-clés libres mais pas dans le résumé,
- l'intitulé du colloque est considéré comme faisant partie du titre, etc.

[11] C'est d'ailleurs dans cet effort d'explicitation et d'identification progressive de l'objet de la recherche que se situe une part majeure du travail du professionnel de l'information, quand il aide à réaliser une recherche documentaire (Saracevic & *al*. 1991).

cherche... or justement, si l'on manque d'informations, ou si l'on veut transmettre un document qui ne ressort pas de son domaine de compétences, il est difficile de définir avec justesse ce qui doit conduire la recherche[12]. Désigner, cerner ce que l'on recherche, est d'autant plus délicat que cela conditionne toute la suite. L'utilisation directe d'un texte s'avère un moyen de contourner la difficulté d'expression d'un thème, en s'appuyant sur un document qui, lui, reflète une certaine maîtrise du sujet. Partir d'un document ou plusieurs documents servant d'exemple économise cette maïeutique et évite de restreindre la recherche faute d'inspiration.

Mieux encore : pour quelqu'un qui lit bien une langue étrangère, mais qui a plus de mal à s'exprimer avec cette langue (cas classique : la version est un exercice plus facile que le thème), partir d'un texte dans cette langue permet d'éviter une reformulation hasardeuse et délicate en quelques mots, tout en donnant potentiellement accès à d'autres documents sur les mêmes sujets. Même une bonne connaissance générale d'une langue ne suffit pas à faire une traduction efficace d'un document technique : toute la difficulté est dans la connaissance de la terminologie spécifique au domaine, qui s'apprend dans la fréquentation du domaine, mais n'est pas enregistrée dans les dictionnaires. La requête textuelle évite une recherche infructueuse faute de connaître la terminologie usitée, et épargne de devoir déployer des moyens importants pour trouver cette terminologie (compulser des documents du domaine, déranger des experts, etc.)[13].

Faire du document lui-même la requête économise un travail de description (code de classement, mots-clefs...) et semble simplifier beaucoup la tâche de l'utilisateur du système. Néanmoins, il reste à l'utilisateur une part de responsabilité et de décision, par exemple le découpage éventuel du document en plusieurs sous-documents plus précis, ou le choix de focaliser la recherche sur une partir jugée plus significative ou représentative. Soumettre le document dans son intégralité est en fait déjà un choix. Il ne faut pas sous-estimer cette étape de formation de la requête textuelle, car elle a clairement une incidence sur la suite du traitement. le découpage éventuel du document : c'est parfois plus par un passage que par le document entier que le lecteur est intéressé.

Le traitement direct de documents bruts en quantité importante

Ce qu'évite la requête textuelle, c'est un travail d'analyse, qui mobilise une personne, pour pouvoir soumettre un document au système. La capacité du système à tirer lui-même l'information dont il a besoin du document original (ou 'document primaire') rend faisable le traitement de volumes ou de flux importants de documents. Concrètement, il devient envisageable de signaler ou / et diffuser rapidement, et avec précision, les différentes fiches techniques d'un gros dossier, les communications d'un colloque à large couverture, les pages de sites Internet stratégiques.

La disponibilité d'une forme électronique

Le critère décisif, pour la viabilité de l'interrogation par un texte, est en fait l'existence de textes électroniques correspondant à la recherche souhaitée et facilement accessible dans l'environnement de l'application. Autrement dit, on ne rédige pas un texte *ex nihilo*, *from scratch*, pour les besoins d'une recherche documentaire : on tape au clavier les quelques mots qui viennent à l'esprit.

De même, il serait utopique de compter sur une utilisation à large échelle des technologies de passage du papier à la forme numérique (acquisition de l'image du document par scanner, puis reconnaissance optique de document et de son texte). Car il faudrait multiplier les équipements matériels et logiciels correspondants, assurer leur disponibilité et leur accessibilité rapide (immédiate) et à moindre coût, miser sur le fait que tout utilisateur soit suffisamment familiarisé avec ce genre de conversions pour y penser et ne pas hésiter à s'en servir...

[12] « si l'utilisateur ne connaît ni le domaine, ni le vocabulaire employé, il peut cependant obtenir une information de qualité grâce au processus de reformulation et d'hypertexte dynamique », l'hypertexte dynamique résultant du fait que « toute partie d'un document peut être utilisée comme question ».
(extrait de la documentation commerciale de la société T.GID sur SPIRIT)

[13] Internet est ainsi utilisable comme un immense réservoir de documents dans lesquels rechercher les usages d'un mot donné, quand on n'en est pas sûr.

En revanche, le point de départ idéal, c'est une pièce jointe reçue par la messagerie, une note interne que l'on finit de mettre en forme dans son traitement de texte, un article trouvé en navigant sur le Web : un banal copier / coller économise l'effort de rechercher quelques mots-clés descriptifs et fournit une requête parfaitement adaptée à l'application. A cette occasion, il n'est pas inutile de préciser (et de montrer) que l'opération de copier / coller, très connue *dans le cadre* d'une application (par exemple à l'intérieur d'un traitement de texte d'une part, et à l'intérieur de la messagerie d'autre part), se généralise au transfert de données textuelles d'une application dans une autre (par exemple, que l'on peut copier dans la fenêtre d'un navigateur Internet, ou dans sa messagerie, ou dans un éditeur de texte, et coller dans une autre application, en particulier dans la fenêtre de requête de DECID).

On pourrait s'inquiéter de ce que ces situations favorables, certes existantes, semblent minoritaires. Le meilleur correctif est que le système apporte lui-même des textes à partir desquels relancer la recherche. Le scénario type est alors le suivant : l'utilisateur donne quelques mots comme point de départ ; les textes calculés comme proches sont présentés à l'utilisateur, qui peut à ce moment-là en tirer la matière pour construire une requête textuelle (copier / coller d'un passage, ou indication d'un document dans son entier). C'est ce que nous appelons la requête textuelle *par rebond*.

La tactique du rebond soulève deux questions alternatives, qui semblent la piéger : *(i)* si l'utilisateur est content des premiers résultats, pourquoi irait-il en chercher d'autres ? *(ii)* et si à l'inverse l'utilisateur trouve les premiers résultats insatisfaisants, va-t-il relancer la recherche à partir d'eux ? La première objection n'est pas très sérieuse, quiconque a quelque peu pratiqué une recherche d'information en conviendra. Car soit l'on cherche à *re*trouver une personne ou une référence précise, et effectivement la recherche s'arrête dès qu'elle est localisée. Soit l'on procède à une recherche plus exploratoire, et alors il est naturel de relancer la recherche par rebond, pour confirmer et compléter les premiers résultats grâce à une seconde vague de résultats obtenus par une requête *a priori* supérieure. La seconde objection (à savoir, va-t-on relancer la recherche à partir d'un ensemble de documents insatisfaisants) doit être précisée. Les premiers résultats peuvent être dans leur ensemble insatisfaisants, mais néanmoins apporter quelques paragraphes qui approchent de ce que l'on cherche : ce peut être un premier 'petit' rebond, qui amorce la progression de la recherche par requêtes textuelles. Si en revanche absolument rien ne semble convenir, c'est peut-être, tout à la fois, un signe de la faiblesse de l'interrogation par mots-clés, une invitation à revoir la formulation de la requête, un encouragement à considérer s'il n'y aurait pas malgré tout un texte assez facilement accessible et utilisable comme requête.

c) Déploiement d'un contexte

Réduction du bruit : disparition d'ambiguïtés artéfactuelles

Un mot pris isolément a en puissance une diversité de réalisations possibles. Les termes atomiques d'une requête booléenne ont eux-mêmes à être décrits : c'est par exemple leur position dans un thesaurus qui les situe. Même une phrase est insuffisante pour expliciter l'objet d'une recherche : lui échappent des « évidences » en fait non généralisables, des oublis, des précisions éclairantes.

En revanche, les mots pris dans leur ensemble situent le domaine du texte : l'essentiel des ambiguïtés est levé. Plus il y a de contexte, mieux le sens est cerné, à condition évidemment de ne pas avoir une vision éclatée du texte (comme un paquet de mots-clés isolables), mais bien une vision globale, synthétique (coopération des différents mots à la construction d'un sens).

> We conclude that word sense ambiguity is only problematic to an IR system when it is retrieving from very short queries. (Sanderson 1994)

> In principle, different texts may exist that cover substantially related subject matter in completely different terms. In that case, a vocabulary comparison method in not adequate for text identification. In practice, identical events are not easily described without using substantially overlapping vocabularies. Such an overlap is then detectable by properly designed text-matching methods. (Salton, Allan, Buckley 1994)

Le même article (Salton, Allan, Buckley 1994) précise cependant qu'il existe des cas où des domaines différents partagent un même vocabulaire. Il donne comme exemple le cas d'un rapprochement entre une phrase sur le football américain, et une sur la théorie des jeux (mathématiques probabilistes), en raison de mots comme *games*, *play* et *team(s)*. Il faut admettre que de tels recouvrements de vocabulaire existent, mais restent relativement rares. Une présentation thématique des résultats de la recherche documentaire, comme celle présentée pour DECID, permet de départager immédiatement les domaines.

Grâce au contexte apporté par le texte, on a d'emblée ce que cherchent à recueillir des procédés classiques comme le *relevance-feedback*, ou l'enrichissement contrôlé à la mode de LiveTopics. Ces procédés visent en effet à recueillir, dans les textes de la base, du vocabulaire pour contextualiser une requête de quelques mots-clés.

Le nombre des mots : une force centrifuge ou une force centripète ?

La réaction spontanée à la présentation de l'interrogation par un texte, comme nouveau mode de recherche par rapport à l'interrogation par mots-clés, est souvent l'occasion de dissiper le malentendu suivant : « plus il y a de mots, plus le système me rapportera des documents hors-sujet ». Ou à l'inverse, la crainte héritée des systèmes d'interrogation booléenne, qui très vite tarissent quand la requête excède quelques mots.

> Contrairement aux systèmes classiques à fonction booléenne de mots-clés (ET, OU, SAUF) qui donnent d'autant moins de résultats que la question est longue (donc trop restrictive), un système modulé par des fréquences comme celui que nous exposons donne des réponses d'autant meilleures que la demande de renseignements est plus détaillée. (Fluhr 1977, §III.7.1, p. 182)

Au contraire, nous avons vu qu'une des forces d'une requête textuelle bien exploitée par le système est qu'elle fournit un contexte, qui réduit les ambiguïtés. Cela appelle donc clairement un effort de communication sur ces points précis qui préoccupent les utilisateurs potentiels : explications (théoriques) et démonstrations (concrètes) sont complémentaires.

D'autre part, des modes de focalisation (sur un élément du texte) et de canalisation de la diversité des résultats sont à prévoir. En effet, on peut souhaiter que le texte contextualise un élément particulier sans disperser les recherches sur d'autres aspects.

d) Réduction du silence : Caractérisation par delà les néologismes, les variantes

Le vocabulaire du texte de requête et du document trouvé peuvent ne pas se superposer exactement, ils seront mis en relation s'il partagent le même contexte, la même toile de fond. Or une nouveauté est bien présentée en référence à des acquis sur lesquels elle se fonde, ou en comparaison avec des « équivalents » qui font référence, qui ont fait date, sont connus et reconnu, et par rapport auxquels elle se positionne et justifie son introduction. Une nouveauté s'inscrit dans un contexte, elle est appliquée à quelque chose, elle contribue à résoudre une question prégnante, elle concerne tels acteurs dans telle situation, etc.

Par ce mécanisme, ce qui traite de nouveautés ne reste pas isolé et inaccessible. En effet, s'il fallait nommer la nouveauté pour avoir connaissance du texte qui en parle, beaucoup d'information serait perdue : par ignorance de l'existence même de le nouveauté, par difficulté à donner le nom qu'elle peut prendre si sa désignation n'est pas stabilisée, par impossibilité de la désigner si elle sort du vocabulaire prévu (indexation contrôlée). L'exploitation des informations textuelles est manifestement avantageuse lorsqu'elle prend en compte une nouvelle activité, et la situe avant même que des liens se soient officiellement constitués avec d'autres activités. On donne accès à ce qui est pertinent et que l'on ne peut pas ou que l'on ne sait pas formuler.

Le cas des nouveautés est le plus remarquable et intéresse spécialement la diffusion ciblée (diffusion de documents confidentiels novateurs, sur la 'frange' des connaissances de référence ; capacité à rendre compte des recherches actuelles, sans réduire ce qui fait justement leur force et leur originalité). Mais ce rôle de l'arrière-plan est également le moyen de dépasser d'autres difficultés ponctuelles lexicales : variantes d'écriture et emplois de synonymes, termes techniques pointus, jargon local ou /et passager, habitudes de vocabulaire différentes d'une discipline à l'autre (alors que l'on peut vouloir bénéficier de points de vue complémentaires sur un même objet).

Dans les systèmes documentaires classiques, le problème des variations lexicales, à savoir qu'un document pertinent n'est pas retrouvé car la requête a exprimé le sujet avec d'autres termes, est abordé au moyen de modules de reformulation (proposition ou adjonction de synonymes, réinjection de termes caractéristiques complémentaires présents dans les documents intéressants déjà trouvés)[14]. La requête textuelle opère déjà pour elle-même une reformulation de qualité (pas d'adjonctions incongrues), car des contraintes stylistiques amènent une variation naturelle du vocabulaire, pour les termes justement qui ne sont pas fixés dans la pratique.

e) Une expression plus juste : relief, interrelations, implicite

Les termes ont des saillances différentes, en fonction par exemple de leur visibilité (reprise fréquente, originalité, caractère central), de leur portée (titre), de leur usage dans le genre correspondant au texte. Une bonne connaissance du type de document et du corpus général (contexte documentaire) permet de trouver et de mettre en valeur l'information la plus significative, les qualités propres de chaque texte vis-à-vis d'un thème de recherche. C'est aussi une voie pour faire la part entre différents types d'informations (par exemple, sur un corpus scientifique, ne pas brouiller ce qui est relatif à la méthode et aux moyens, aux résultats).

La donnée du texte lui-même permet de ne pas s'en tenir au « mot à mot », à une juxtaposition de mots-clés, mais de s'appuyer sur des unités qualifiées voire reconstruites en contexte.

Enfin, un texte s'inscrit dans une réalité fortement informative : auteur, type de document (correspondant à un certain usage), etc. Par l'intermédiaire de son contexte général, il devient possible de situer le document dans son domaine, pas toujours clairement exprimé ou décrit dans le texte même.

Tout ceci suppose que l'approche soit elle-même textuelle. Si la première chose que fait le traitement, c'est de tronçonner le texte en mots, puis ensuite de traiter ces formes chacune pour elle-même, alors le déploiement du texte risque plutôt d'empirer les choses. Un système conçu pour traiter des mots-clés, et recevant la matière d'un texte, n'est pas à même de gérer la multiplicité des unités recensées : il ne sait ni les articuler (en les combinant), ni les organiser (en les hiérarchisant).

Les premiers rudiments d'un traitement textuel sont les tactiques mises au point pour la recherche sur le texte intégral : distinction des mots grammaticaux, et d'une manière plus générale de ce qui fait office de liant dans le corpus considéré ; pondération, qui rende compte de la dominance (locale, dans le texte) et de la spécificité (globalement, dans le cadre du corpus) ; renforcement mutuel des mots en interrelation (expressions composées, cooccurrences).

La pleine prise en compte des textes pose très vite la question de la prise en compte des genres. Tant que l'on reste à l'intérieur d'un même genre, cet aspect reste latent. Il prend en revanche une importance décisive quand sont à confronter des textes de différents genres. Chaque genre a par exemple ses habitudes de vocabulaire, ses éléments d'articulation, ses informations caractéristiques attendues. Il y a donc une hétérogénéité des usages linguistiques à travers les genres : une analyse adaptée à chaque genre s'impose pour rendre les textes commensurables. Il y aurait deux niveaux de traitement envisageables : soit qu'une indication du genre du document soumis soit un paramètre à fixer, soit que le système induise lui-même, à partir de certains indices caractéristiques, à quel genre rapporter le document.

Pour DECID, les nouveaux types d'unités proposés pour la représentation interne des textes intègrent le déploiement textuel et les relations qui se tissent entre les mots en contexte. La description spécifique de genres est également prévue. Au plan de l'interface, un menu serait à ajouter pour que l'utilisateur puisse donner une indication sur le genre du texte qu'il soumet, si ce genre fait partie des genres privilégiés prévus (sinon, un traitement 'neutre' propose néanmoins des résultats, avec une assurance de qualité moins grande).

---

[14] (Radasoa 1988) étudie de manière approfondie les procédés de reformulation (dans le cadre du système documentaire SPIRIT). Il fait figurer, comme septième et dernière méthode de reformulation, l'interrogation par partie de document : à savoir, l'analyse du passage sélectionné fournit de nouveaux mots-clés, qui s'ajoutent, avec une importance moindre, aux mots-clés initiaux de la requête.

f) Une expression plus ouverte, non déterminée par les a priori

Un texte comporte plus de matière que quelques mots, il peut avoir des aspects, liés au thème dominant, que l'on n'a pas remarqués, et qui peut-être même étaient inconnus. La confrontation du texte à un corpus en renouvelle la lecture et permet de sortir des *a priori* ou des premières impressions qui peuvent regrettablement confiner la recherche de documents en relation. Le passage par un texte permet ainsi l'accès à de l'inattendu, en relation avec le sujet.

La prise en compte des textes dans leur diversité, sans les convertir à une expression dans un format prédéfini, permet de sortir des liens établis et déjà connus. Les mots-clés vont dans le sens d'une banalisation de l'information, puisque pour être efficaces ils constituent des désignations conventionnelles et suffisamment fédératrices. Le « bon mot-clé » procède d'un accord, explicite ou tacite, sur une manière commune d'évoquer un thème, en gommant les spécialités : c'est une manière de favoriser les rapprochements, mais au détriment des singularités, alors que celles-ci sont significatives et valorisantes.

A cela s'ajoute que le pouvoir expressif des textes est bien plus grand que celui de mots-clés, ceux-ci émanant de l'organisation des disciplines, structurant et étiquetant la connaissance. Un texte permet de décrire un développement à la croisée de plusieurs disciplines, de s'affranchir de souscrire à des rubriques prédéfinies dont aucune n'est vraiment appropriée.

La réécriture du document primaire (document original) en un document secondaire (mots-clés par exemple) se traduit en quelque sorte par la perte d'un « degré de liberté » sémantique. En effet, le document secondaire, aussi complet et fidèle qu'il se veuille, adopte nécessairement un point de vue et limite les possibilités interprétatives. Un même document primaire peut donner lieu à *n* documents secondaires, tous différents les uns des autres. En travaillant directement à partir du document primaire, le système a donc accès au texte dans toute son ampleur originelle.

Les résultats sont par conséquents plus riches et plus larges, tout en gardant une certaine cohérence et une logique relative au texte de départ. Le traitement doit cependant prévoir des outils pour focaliser la recherche sur un aspect particulier ou/et canaliser la dispersion des facettes trouvées.

g) Correspondance avec le texte intégral

Partant d'un texte pour aller vers d'autres textes, on n'a pas ce décalage de niveau de vocabulaire que l'on aurait avec quelques mots-clés, qui condensent souvent une désignation du thème plus synthétique mais n'apparaissent pas toujours comme tels dans les textes.

De fait, chaque mode d'interrogation est en correspondance optimale avec un mode d'enregistrement de la base de recherche, et une optique d'interrogation.

Une requête exprimée sous forme d'équation de recherche dans une langage documentaire, s'accorde avec une base indexée par des descripteurs, et convient à une recherche sur une thématique bien représentée dans le référentiel définissant les descripteurs. C'est une recherche « méta », sur les représentations que le système a des textes.

Une requête indiquant une chaîne de caractères, ou une combinatoire de chaînes de caractères (à l'aide d'opérateurs booléens ou de troncature) vise à retrouver les occurrences d'un mot ou d'une expression. Elle est adaptée à une base en texte intégral, et convient pour des recherches de localisation plutôt que de découverte. Cette technique est efficace pour retrouver un document précis, si l'on s'en rappelle un élément caractéristique ; elle est tout à fait adaptée à la recherche d'attestations lexicales (ce mot existe-t-il avec cette orthographe, quels sont les usages de ce mot, quel est le développé de ce sigle –s'il est précisé, il y a des chances que ce soit dans le voisinage immédiat d'une occurrence du sigle). La recherche par chaîne de caractères est beaucoup moins performante pour une recherche sur une thématique : en effet, elle est purement fondée sur l'expression (le signifiant), et ne prépare aucunement une recherche sur le contenu significatif (le signifié).

L'interrogation par un texte est en harmonie avec l'exploration d'une base de textes ; elle convient naturellement à la recherche de documents en relation sémantique, sans préjuger de la thématique précise qui peut établir un rapprochement. La prise en compte des interrelations contextuelles entre les unités d'indexation permet de ne pas s'en tenir aux chaînes de caractères et d'atteindre effectivement un niveau sémantique. Il s'agit d'une sémantique implicite et dynamique, non enregistrée dans un référentiel terminologique ou conceptuel. En ce sens, la recherche est plus souple, moins calibrée sur des concepts choisis.

3. Le choix du texte comme base pour la caractérisation des profils dans DECID : réactions et discussion

a) Un détournement des textes ?

Conversion brutale d'une information organique en une information matériau

Les textes d'Actions sont, pour les chercheurs, des documents servant au fonctionnement, à l'organisation interne de l'entreprise, plutôt que des documents alimentant le travail de recherche lui-même.

Or ces deux finalités de l'information (fonctionnement de l'entreprise vs support de travail) induisent des rapports antagonistes au document :

> L'acteur individuel [...] [participe], comme partie de l'organisation, à la fois à la vie de la structure et à sa mission (organisation et production). Nous parlerons de fonctions *organiques* et de fonctions *productives*. Les informations manipulées dans ces contextes seront des informations *organiques* ou des informations *matériau*.
>
> [...] L'information *matériau* est l'objet [du] métier [de l'acteur] : il va donc la considérer en *professionnel*, avec toute l'implication que cela implique sur le plan social et affectif. Il est un expert dans son traitement, et il peut en parler de façon légitime. C'est son affaire. Traiter cette information est parfois un plaisir, en tout cas toujours une activité socialement reconnue et légitime. Il est payé pour le faire, et il est normal qu'il y consacre une partie de son temps, puisque c'est son métier.
>
> [...] [L'information organique] risque [pour sa part] [...] d'être rapidement la source de conflits, puisque, si pour un acteur A elle est *organique*, elle sera sans doute, pour un acteur « fonctionnel » quelque part ailleurs dans l'entreprise, à la source ou dans sa destination finale, une information *matériau*. D'un côté, celui des usagers, elle est perçue comme peu noble, voire comme une charge ou une contrainte. De l'autre, celle du producteur, comme un matériau utile, difficile à produire et ayant de la valeur.
>
> (Lahlou 1994, §2.5, p.16, et §4.1, pp. 44-45)

Le chercheur ne s'implique que modérément dans les documents de fonctionnement :

> La partie administrative des CERD [Contrats Externes de Recherche et Développement] est rédigée par l'Antenne de Gestion. Je ne rédige que la partie technique.
>
> C'est mon chef de groupe qui rédige les ARD [textes d'Action], et ça me convient[15]. Je pourrais le faire si on me donne les « règles du jeu ».
>
> (Merle, Fradin 1994, §6.2.2, p. 24)

> Frappe initiale des AID/ARD par les chercheurs imposée par le chef de Département précédent, mais ils ne suivent pas les modèles (égarés ?). L'Antenne de Gestion arrange présentation et orthographe, mais on a besoin de téléphoner pour vérifier les explications à donner sur le contexte et les buts.
>
> L'administratif passe pour une contrainte pour les chercheurs, qui attendent la dernière minute, on ne peut pas alors soigner la qualité du travail et contrôler l'intérêt de ce que l'on produit.
>
> Les AID et ARD sont validées successivement par le chef de Département et le chef de Service.
>
> (Merle, Fradin 1994, §6.2.2, p. 25)

Inadéquation

En prenant conscience que le texte d'Action, qu'il rédige comme un document administratif à l'intention de ses supérieurs, est aussi exploité comme matériau pour la construction de son profil

[15] Réponses à un questionnaire, lors de l'enquête PUBE (Merle, Fradin, Soinard 1995, p. 23) :
Proposition : « C'est le chef de groupe qui rédige les ARD »
Réponses : Vrai : 1 personne ; Faux : 9 personnes ; Ne sait pas : 2 personnes.
Commentaires : « Cela peut arriver mais ce n'est pas souhaitable », « non, mais il les contrôle beaucoup ! »

pour recevoir de l'information pour sa recherche, le chercheur EDF pressent une certaine inadéquation de son texte d'Action dans cette nouvelle fonction.

> Chacun comprend que les ARD [textes d'Action] sont faits pour la hiérarchie et validés par elle, et qu'il est prudent de ne pas mélanger les objectifs. Amorce d'une réflexion sur les limites de ces documents en tant qu'outils de communication interne. (Merle, Fradin 1994, §2, p. 7)
>
> Le contenu des ARD/AID [textes d'Action] est ciblé pour le directeur adjoint, et n'est pas suffisant pour y chercher l'information dont on a besoin (mais l'ARD dans sa forme actuelle a une fonction qu'il ne faut pas changer). (Merle, Fradin 1994, §6.1.8, p. 18)
>
> Les Comptes Rendus types d'ARD sont conçus pour contrôler l'activité par la hiérarchie, non dans un objectif de diffusion[16]. (Merle, Fradin 1994, §6.2.2, p. 24)
>
> Une ARD n'est pas nécessairement « autoporteuse », ça dépend de l'objectif qu'on lui donne. (Merle, Fradin, Soinard 1994, p. 72)
>
> Mon profil n'a rien à voir avec mon ARD. Mon ARD est peut-être plus significative des orientations de mon groupe à un moment. (Merle, Fradin, Soinard 1994, p. 60)
>
> On ne peut pas ajuster le profil d'un individu à des textes d'ARD/AID, c'est trop restrictif. (Merle, Fradin, Soinard 1994, p. 64)
>
> Je lis le contenu formel de l'ARD. Je sais ce qu'ils font en réalité, ça n'a aucun rapport. Ou alors c'est un petit effet de bord. Manquent les enjeux essentiels.
>
> Les enjeux caractérisant un axe de recherche ne sont pas présents dans les ARD. Il faut aller les chercher dans les Contrats de Groupe et les Plans Stratégiques des Départements. MAIS les Contrats de Groupe ne sont communiqués qu'avec l'accord du Chef de Département, et ne sont pas faits pour être compréhensibles. Ne reste que le résumé du Contrat de Groupe publié lors de sa création (et pas mis à jour s'il y a réorientation).
>
> (Merle, Fradin, Soinard 1994, p. 72)
>
> Réponses à un questionnaire, lors de l'enquête PUBE (Merle, Fradin, Soinard 1995, p. 18) :
> *Proposition* : « Mes ARD sont caractéristiques de mes centres d'intérêt »
> *Réponses* : Vrai : 5 personnes ; Faux : 2 personnes ; Ne sait pas : 4 personnes.
> *Commentaires* : « pas complètement », « je n'ai pas d'ARD[17] », « Oui, mais mes centres d'intérêt professionnels et actuels seulement ».

Réactions et propositions constructives

Le chercheur a l'impression que son texte d'Action n'explicite pas ses besoins en informations... mais aussi, ajouter cette tâche de construction d'un profil serait bien perçu comme un travail exigeant et une charge rébarbative, sauf exception (utilisateur motivé, voire enthousiasme passager).

> Le « *Qui Fait Quoi ?* »[18] est intéressant, mais à consulter dans le métro. Je doute que quelqu'un soit prêt à y entrer des informations sur son activité pour créer un réseau en retour. (Merle, Fradin 1994, §6.1.8, p. 18)

Une redéfinition du genre ?

Certains modifieraient leur mode de rédaction, pour adapter le texte au double objectif : non seulement une lecture par la hiérarchie pour approbation de leur programme d'activité, mais aussi une

---

[16] De fait, pour les textes d'Action, c'est le volet Fiches Descriptives, mais pas le volet Comptes Rendus, qui est exploité pour la caractérisation des destinataires de la diffusion ciblée. Les Comptes Rendus sont une information plus sensible et plus confidentielle ; les Fiches Descriptives ont également l'avantage d'un point de vue prospectif (les Comptes Rendus sont par nature rétrospectifs).

[17] A peu près un chercheur sur deux est responsable d'une Action.

[18] Il s'agit du fascicule annuel, présentant les textes des Actions calculées comme les plus « proches » des Actions d'une équipe (Groupe ou Département), cf. chapitre *Introduction*.

exploitation par une machine à des fins de diffusion ciblée d'information. Ce serait en fait une redéfinition du genre 'texte d'Action', avec peut-être comme conséquences : une modification du vocabulaire, davantage de mentions techniques, l'évocation de thèmes secondaires mais sur lesquels on voudrait être informé, etc.

> Pour que mon ARD [texte d'Action] soit significative en termes de centres d'intérêt, il faudrait que je la retravaille avec un « candide » pour que le texte soit compréhensible, pour qu'il sache ce que j'ai besoin de savoir. Ensuite je pourrais charger quelqu'un de compétent (du Département ou un documentaliste) de me tenir informé dans ce domaine. (Merle, Fradin, Soinard 1994, p. 60)

> [La diffusion ciblée] c'est super que ce soit fait à condition de recevoir uniquement ce qui m'intéresse. Pour cela il faudrait que le texte [sur lequel est basé mon profil] soit conçu pour que la machine le comprenne. Si je structure bien mon texte pour qu'il soit compris par quelqu'un alors il doit l'être de la machine qui devient intelligente. (Merle, Fradin, Soinard 1994, p. 62)

Cette attitude, même si elle était encouragée voire demandée, n'aurait sans doute pas des répercussions générales. Si les chercheurs ont déjà tendance à négliger ces écrits « administratifs », ils ne se prêteront pas beaucoup plus volontiers à une rédaction soignée, surtout si elle est perçue comme une exigence et non comme une intéressante source de bénéfices informationnels (tout le monde n'est pas un fervent utilisateur des services documentaires).

Incontournables mots-clés

Le recours à des mots-clés a été spontanément proposé pour suppléer au texte.

> Il faudrait pouvoir associer les mots-clés du domaine technique et les listes de contacts et de collaborations en cours[19]. Par exemple, qu'est-ce qui se fait avec le (au) SEPTEN ? (Merle, Fradin, Soinard 1994, p. 60)

> Je pourrais définir mes centres d'intérêt avec des mots-clés et des exclusions :
> protection + transport + distribution SAUF production
> production autonome SAUF machines
> Partir de mes ARD ? OK mais ce n'est pas suffisant.
> Partir des documents que je produis ? Ca dépend des cas, je dois pouvoir en valider la pertinence.
> (Merle, Fradin, Soinard 1994, p. 61)

> Ce n'est pas le texte que je donne à la machine mais les mots-clés. Le texte on n'en est pas forcément maître. (Merle, Fradin, Soinard 1994, p. 62)

C'est effectivement le réflexe technique des habitués des pratiques de recherche documentaire. Mais cette solution rencontre très vite des difficultés. Lorsqu'un chercheur EDF rédige une note, et qu'il lui est demandé d'indiquer des mots-clés, surgissent immédiatement une série de questions : peut-on mettre n'importe quel mot-clé ou faut-il les prendre dans une liste établie, un thesaurus ? Dans le premier cas, on n'est jamais sûr de choisir les termes appropriés, ou qui seraient utilisés par quelqu'un d'autre ou ultérieurement. Dans le second, il s'agit d'une part de disposer facilement de la liste de référence, d'autre part d'y trouver des termes représentatifs et de composer un ensemble de mots-clés équilibré vis-à-vis des autres documents. Bref, il s'agit davantage d'une compétence de documentaliste que de tout rédacteur de document.

> - Il faut mettre [des mots clés à sa note], mais pour quoi faire ? On ne sait pas comment mettre les bons mots-clés.
> [...]
> - Le thesaurus on ne connaît pas. A chaque fois qu'on nous demande des mots-clés, on respire un grand coup, on met cinq mots-clés mais on ne sait pas si c'est pris en compte... ou alors c'est le chef qui le fait[20]... et la plupart du temps ce ne sont pas les mots-clés que j'aurais mis.

---

[19] Le plan type des textes d'Action prévoit une rubrique *Partenariats envisagés*, mais cela ne recouvre donc pas nécessairement les collaborations en cours.

[20] La proposition « C'est souvent le chef qui met les mots-clés sur mes travaux » est unanimement désavouée par sept chercheurs EDF, en réponse à un questionnaire (Merle, Fradin, Soinard 1995, p. 36).

(Merle, Fradin, Soinard 1994, pp. 93-94)

Autres corpus : mais le corpus adéquat existe-t-il ?

Bien entendu, quels que soient les documents recueillis, ces écrits professionnels n'explicitent pas systématiquement et complètement l'activité du chercheur.

> La note est présentée comme l'objectif numéro un de l'ARD. Mais en fait, le processus d'acquisition de compétences est aussi important. Une partie du savoir-faire n'apparaît pas dans les notes (informations confidentielles pour la direction cliente).
>
> [Une part de cette information fait l'objet d'une] restitution opérationnelle lors de réunions. [Une part également est publiée sous forme de] communication lors de colloques : accord verbal de la hiérarchie, autorisation implicite du client.
>
> (Merle, Fradin 1994, §6.3.1, p. 29)

Ils laissent dans l'ombre les informations jugées trop confidentielles, dans certains genres sont omis celles qui seraient trop techniques. L'« évident » n'est pas redit. La plupart des documents *circulant* dans l'entreprise ne font pas non plus état des questions encore ouvertes, sur lesquelles on est en recherche : on a le plus souvent affaire à une « production finalisée » –des décisions, des résultats.

> Le compte rendu d'activité [du Département] MMN [...] contient les listes suivantes : notes, publications, comptes rendus, cours donnés, thèses en cours et soutenues, stages, séminaires. Il n'y manque qu'une chose : la liste des questions que les chercheurs se posent (et ce n'est pas une boutade, cette information est significative). (Merle, Fradin 1994, §6.3.2, p. 30-31).

Le choix du (des) document(s) pour caractériser un profil suppose d'avoir précisé quelle orientation l'on veut donner au profil : veut-on décrire les compétences acquises, celles à acquérir, les intérêts du moment... voilà autant de facettes qui arbitrent et décident de la représentativité d'un document. Ainsi, les documents que consulte un ingénieur donné et ceux qu'il rédige ne se ressemblent pas nécessairement : ce n'est donc pas sur un rapprochement avec les seconds que l'on pourrait retrouver les premiers.

b) De la confidentialité

L'utilisation de textes existants, comme source d'information pour construire les profils, a l'avantage d'apporter une information relativement riche, plus riche que la donnée de quelques mots-clés, génériques et épars, plus ou moins réfléchis. Les calculs effectués à partir de cette représentation peuvent donc fournir des résultats motivés et intéressants.

Il est important aussi de pouvoir éclairer et commenter le lien trouvé pour chaque rapprochement texte - texte effectué. Mais il n'est pas toujours souhaitable de donner accès à un document ayant servi à construire un profil : il peut s'agir d'une personne dont certains aspects de l'activité sont confidentiels. L'information textuelle doit être alors exploitée pour fournir des indicateurs satisfaisants, sans pour autant révéler, même indirectement, des aspects qui ne sont pas connus de l'utilisateur et qui n'ont pas à le devenir.

Une tactique de projection par exemple (dont on montrera une réalisation dans l'interface de DECID) permet de caractériser chaque rapprochement uniquement par rapport au document soumis, sans dévoiler des informations que l'utilisateur n'aurait pas déjà en sa possession.

B. UNE SESSION D'INTERROGATION DE DECID : FONCTIONNALITÉS ET INTERFACE

1. Présentation de la démarche suivie

Nous avons choisi de présenter ci-après l'ensemble des fonctionnalités de DECID, suivant un ordre (chrono-)logique. Cette manière de faire a l'avantage de donner une vue complète et cohérente de l'interface de l'application de diffusion ciblée. Elle rend compte de l'interface à laquelle aboutit le projet, fruit de la réflexion de toute l'équipe pendant toute la période de la thèse. Un certain nombre de fonctionnalités existaient déjà sur la version messagerie de l'outil de diffusion ciblée (en 1994), mais une part importante de l'ergonomie a été élaborée, et les anciennes fonctionnalités revues, à l'occasion de la mise au point de l'interface Web, inaugurée fin 1995.

Ce faisant, en choisissant de donner d'emblée une vue intégrale des fonctionnalités, on occulte délibérément le décalage entre l'interface effective et l'interface imaginée. Car il y a *(i)* DECID en exploitation, tel qu'il est connu des utilisateurs, *(ii)* le prototype intégré, qui prépare la version suivante, en test, et éventuellement présenté en démonstrations, et *(iii)* des modules non encore vraiment intégrés, et qui pour le moment sont des maquettes montrant la faisabilité de tel ou tel type d'interaction ou d'interface. Présenter successivement ces trois niveaux aurait été répétitif, car sauf exception les fonctionnalités d'un niveau sont reprises au niveau suivant. Cela aurait aussi donné une place excessive à des distinctions somme toute contingentes, car ce qui est à un moment donné en exploitation a été auparavant au stade de prototype. L'objectif de la thèse est de faire le point des connaissances acquises, en synthétisant ce que peut être, concrètement, l'interface d'une application de diffusion ciblée. La répartition entre les différents degrés d'avancement fait donc l'objet d'un tableau récapitulatif, qui situe, une fois toutes les fonctionnalités présentées, le statut de chacune.

Une autre distinction importante, et qui devient difficile à percevoir dans une description intégrale des fonctionnalités, est la part entre ce qui relève proprement du travail de cette thèse, et ce qui a été réalisé par ailleurs dans le projet. Cela correspond également à la mise en évidence des fonctionnalités concrétisant des principes de linguistique textuelle, par rapport aux autres fonctionnalités, liées à la problématique de l'application de diffusion ciblée mais moins à la textualité des données. Là encore, un bilan conclusif recense, en les groupant, les apports de la thèse qui ont donné lieu à des innovations au plan de l'ergonomie, pour l'affichage et la manipulation de données textuelles. Cette troisième lecture des fonctionnalités complète les précédentes et ne les remplace pas.

La vue chronologique complète est importante pour comprendre le *contexte* dans lequel les innovations proposées prennent sens, autrement dit pour en percevoir le fonctionnement et la portée. L'application constitue un tout cohérent, et tout démembrement comporte sa part de réduction et d'arbitraire. Par exemple, le mécanisme d'authentification et d'accès par mot de passe pourrait apparaître complètement hors du champ de préoccupation, centré sur le texte. En fait, en y regardant de plus près, c'est ce qui conditionne la visualisation et la navigation dans le texte intégral. Car pour que la diffusion ciblée ait une utilité réelle, il faut que les profils soient calculés sur des textes qui ont une certaine valeur, une charge informative importante. Pour pouvoir se baser sur ces textes, il faut être en mesure d'en assurer une gestion sécurisée. Donc : sans l'authentification, pas de « vrais » textes.

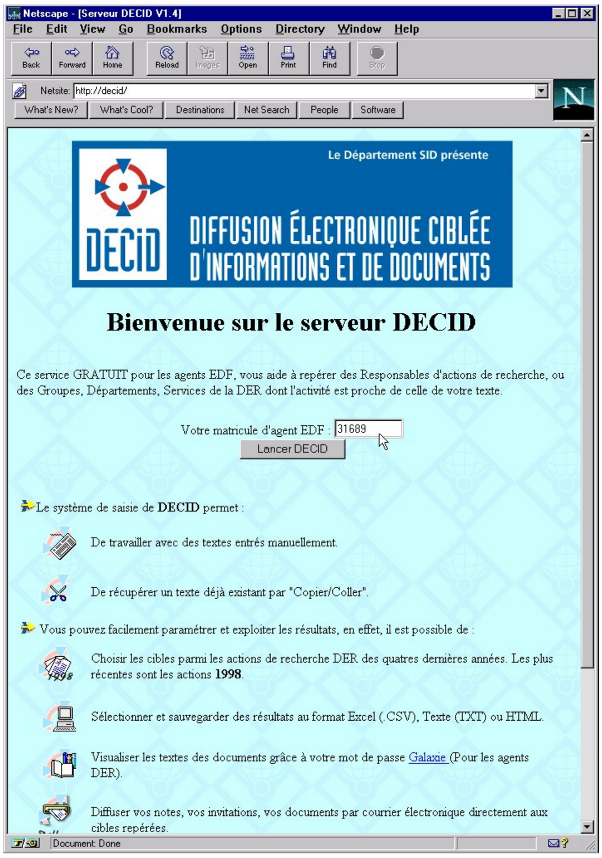

FIGURE 3 : Page d'accueil (Intranet EDF).

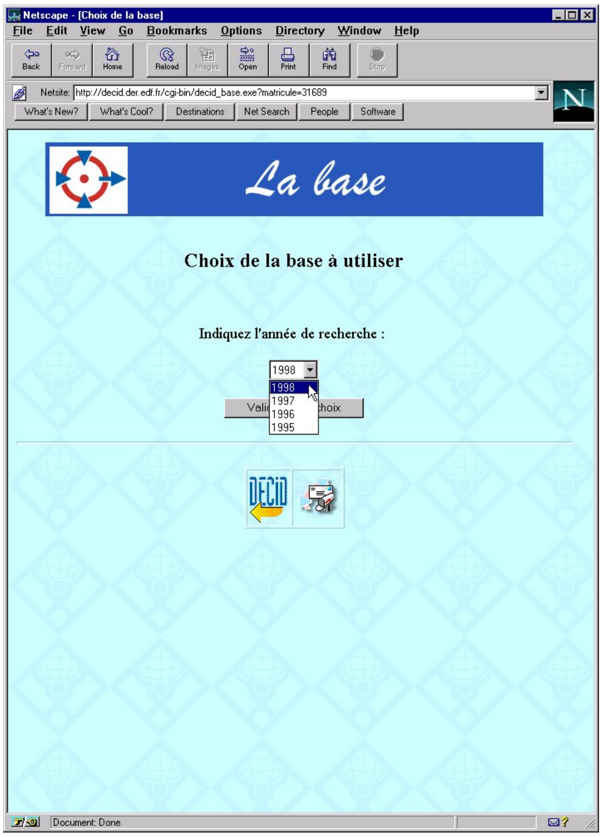

FIGURE 4 : Sélection d'une année ; la recherche des destinataires s'effectuera en fonction des activités des personnes à l'année indiquée.

2. Accès

a) Equipement

Navigateur Web

Il n'y a *pas besoin d'installer (et donc de déployer à grande échelle) un logiciel spécifique* sur les postes des utilisateurs pour donner accès à DECID. Les traitements de DECID se font sur un serveur, et les échanges suivent le protocole `http`. DECID peut donc être interrogé à l'aide de n'importe quel navigateur Web[21] : il suffit d'indiquer, dans son navigateur, l'adresse du serveur[22], pour que s'affiche la page d'accueil. Ensuite, l'utilisateur est guidé par l'interface et retrouve un cheminement hypertexte familier aux internautes. Il n'a pas besoin d'une expertise particulière en ce domaine, car il retrouve des modes d'interaction classiques, comme des menus déroulants, des boutons, un ascenseur en marge de la fenêtre pour faire défiler la page.

On considère que le navigateur Web fait partie de l'équipement logiciel standard en bureautique, et qu'il entre dans les outils suffisamment couramment utilisés et accessibles pour ne pas constituer un obstacle[23]. Les difficultés éventuelles peuvent venir de craintes à introduire le Web dans l'entreprise. Il s'agit peu de risques de sécurité (un « coupe-feu » s'interpose avec l'extérieur, contrôle

[21] Par exemple les logiciels *Netscape* ou *Internet Explorer*. Le terme français *navigateur* est ici l'équivalent de l'anglais *browser*.

[22] Opération qui n'est même nécessaire qu'aux premières interrogations, si l'utilisateur adopte DECID et pense à le « bookmarquer », c'est-à-dire à l'enregistrer dans le carnet d'adresses Web de son navigateur.

[23] DECID peut aussi être l'occasion de s'initier au Web : cela ne présente pas de difficultés majeures, et forme l'utilisateur à une nouvelle technologie.

tous les échanges et prévient les intrusions éventuelles dans le système d'information et les données de l'entreprise. Le facteur principal d'hostilité au Web, c'est la crainte de mettre à disposition un équipement, alors davantage considéré comme un instrument de divertissement (engendrant coûts et pertes de temps), que comme un outil puissant d'information et de communication (permettant des investigations larges, rapides, dynamiques et actualisées).

Remarque importante pour la conception de l'interface, il faut savoir que l'apparence à l'affichage d'une même page Web peut varier sensiblement d'un navigateur à l'autre, et d'une machine à l'autre (résolution adoptée pour l'écran, station Unix ou PC, etc.). Les incidences sont par exemple :

- *les couleurs* : un choix harmonieux dans une configuration peut se transformer en teintes agressives, ou/et modifier les contrastes écriture / fond et rendre la lecture pénible ;
- *l'échelle* : caractères devenant trop petits ou trop gros, troncature latérale du bandeau en entête, réduction du champ de ce qui est visible en premier lieu (il faut avoir l'initiative et faire l'effort de dérouler la page pour voir et avoir accès à certaines informations ou interactions essentielles) ;
- *l'effectivité* : il peut arriver que des différences d'affichage conduisent à des interprétations divergentes[24] ; les navigateurs ont également chacun des extensions fonctionnelles qui leur sont propres (il faut donc se garder de les utiliser si l'on veut avoir une application utilisable avec tous les navigateurs), et n'ont pas le même niveau de compatibilité avec d'autres logiciels, notamment des outils permettant une extension des possibilités de l'interface (par exemple *Tk/Tcl*, dont il est question au paragraphe suivant).

La conception de l'interface suppose donc des tests sur les principales configurations attendues, afin de garantir une bonne qualité de l'ergonomie dans les différents cas.

Option : Plug-in Tk/Tcl

Les capacités des navigateurs actuels (hors extensions spécifiques) sont suffisantes pour des affichages et des interactions classiques. La réalisation de certaines fonctionnalités innovantes, pour DECID, rencontre cependant les limites du standard actuel. En particulier, le langage *Java* ne permet pas[25] d'afficher un texte dans une fenêtre et d'y appliquer interactivement des opérations locales (par exemple, sélection et marquage de passages).

Le langage *Tk/Tcl* est connu et pratiqué pour la réalisation d'interfaces : il offre le moyen de réaliser les *affichages les plus évolués* conçus pour DECID. Son utilisation pour une application Intranet requiert néanmoins l'installation d'un module sur le poste client. Même si cette opération est facilitée par la possibilité de télécharger le module directement à partir de l'application DECID (voir par exemple FIGURE 11), c'est une procédure informatique et une intervention sur la machine à laquelle peuvent renoncer une part importante des utilisateurs. On prévoit donc un fonctionnement de base de DECID, ne faisant appel qu'aux fonctionnalités intrinsèques des navigateurs ; l'ajout du module *Tk/Tcl* est alors une extension optionnelle, pour les utilisateurs qui le souhaitent, donnant accès à des fonctions complémentaires.

Il y a effectivement des utilisateurs plus demandeurs que d'autres d'une version avancée. Tout d'abord, les gros utilisateurs de l'application, qui en ont un besoin quotidien dans leur travail, et qui sont prêts à faire une démarche particulière pour avoir un gain en puissance et en finesse. On peut aussi penser à ceux qui sont curieux des nouvelles technologies, qui ont l'habitude de les expérimenter, et qui sont coutumiers de l'installation de nouveaux outils.

---

[24] Une mésaventure réelle, est celle d'un texte dont certaines parties devaient être présentées comme rejetées : le navigateur utilisé lors de la conception affichait le texte comme barré ; mais un autre navigateur interprétait le fichier autrement et omettait le biffage prévu, ne distinguant plus en rien le texte « admis » et le texte « rejeté ».

[25] Plus exactement, le langage *Java* ne permet pas la réalisation des fonctionnalités souhaitées (d'interventions sur un texte) de façon simple et raisonnablement abordable : il faudrait avoir recours à une gestion graphique du texte, qui demanderait un lourd travail de développement.

b) Aide à l'utilisateur

Documentation en ligne

Une présentation de quelques pages, soigneusement rédigée et illustrée, est destinée au nouvel utilisateur. Elle détaille et explique ce qui ne peut faire que l'objet de mentions ou de rappels lors du traitement, pour garder un affichage agréable et efficace, non surchargé.

Cette documentation a pour vocation de *guider* les utilisateurs moins sûrs d'eux[26], et pour tous de favoriser une *bonne compréhension* du traitement effectué pour permettre un usage optimal. La documentation ne manque pas de souligner les caractéristiques fondamentales de la diffusion ciblée : elle précise les textes qui ont servi à construire les profils, et l'importance de préférer une requête textuelle à quelques mots-clés. Ces considérations d'ensemble sont complétées par un commentaire, étape par étape, d'une part des interactions possibles et de leur signification opératoire, d'autre part des informations apportées par l'affichage et de la manière de les interpréter.

Contact par formulaire

A tout moment d'une session (c'est-à-dire depuis toutes les pages de l'interface de DECID, on le voit par exemple en FIGURE 4), il est possible d'envoyer un *message à l'administrateur* du système[27]. C'est le moyen de demander un dépannage, d'apporter des suggestions, de faire part de ses réactions face à tel ou tel comportement du système.

Ce dialogue est donc important (si ce n'est vital) à plusieurs points de vue. Il permet de repérer des défauts encore inaperçus (problèmes techniques ou mauvaise communication sur certains points), et des limites du système dans des situations d'usage réelles. En même temps, les échos des utilisateurs sont des éléments pour mieux connaître leurs besoins et leurs désirs (« J'ai pu faire ceci, mais j'aurais bien voulu pouvoir faire cela en plus »). Cette forme d'information a l'avantage de faire ressortir les réactions fréquentes (par rapport aux réactions marginales ou personnelles), et de suivre l'avis des utilisateurs au fil de l'évolution de l'application. Enfin, certains utilisateurs sont motivés pour coopérer sur la mise au point de l'application, et le formulaire est pour eux un bon moyen de faire entendre leur avis. Ils peuvent ainsi entrer facilement en relation avec l'équipe de conception et d'administration de l'application.

c) Contrôle d'accès

Intranet

Dans le cadre d'une industrie comme EDF, la diffusion ciblée est une application sensible, dans la mesure où elle donne une vue sur les directions stratégiques auxquelles est consacré l'effort de recherche. Autant il est important que le personnel de l'entreprise ait une connaissance suffisante des activités internes pour assurer une bonne coordination et une bonne synergie des équipes, autant il est clair que l'entreprise est affaiblie si ses concurrents parviennent à identifier ses orientations et à connaître ses choix prospectifs.

Cette protection vis-à-vis de l'externe est assurée par l'utilisation du réseau Intranet (vs Internet), qui limite strictement l'accès à l'application aux postes de travail reliés à ce réseau, concrètement aux ordinateurs sur les *sites EDF*.

Identification

Parmi les personnes sur les sites EDF, toutes n'ont pas le même statut. Par exemple, à la Direction des Etudes et Recherches travaillent non seulement des chercheurs de l'entreprise, mais

[26] Les informaticiens habitués ont plutôt tendance à se servir directement d'une nouvelle application, en s'appuyant sur l'ergonomie qui, si elle est bien conçue, est « intuitive ». C'est un scénario tout à fait réaliste pour DECID, aussi les précisions d'usage les plus importantes sont évoquées sur les pages du déroulement de la session.

[27] Ces messages sont gérés par le module d'administration, et un exemplaire est automatiquement et immédiatement envoyé sur la messagerie du ou des administrateur(s).

aussi du personnel d'encadrement, de gestion, des prestataires de service et des intérimaires, des stagiaires et des doctorants, etc. Approximativement, les trois catégories que forment *(i)* les agents EDF (le personnel statutaire de l'entreprise), *(ii)* les extérieurs, *(iii)* les personnes qui ne sont employés que momentanément par l'entreprise, correspondent à une division en trois tiers de la population présente sur les sites de la Direction.

L'ouverture d'une session de recherche de correspondants avec DECID commence en demandant à l'utilisateur de s'identifier (FIGURE 3). En l'occurrence, *l'utilisateur doit indiquer son « matricule »*, clef unique d'identification dont disposent les personnes employées par EDF et elles seulement.

Cette information fournit un indicateur global sur la population des utilisateurs : combien sont-ils, comment se répartissent-ils dans l'entreprise (dans quels Services DECID est-il bien / mal connu, ou semble-t-il très / peu utile), quels types d'utilisateurs (ou segments) apparaissent en fonction de leur fréquence d'utilisation de l'application, quels sont les types d'utilisateurs majoritaires ou minoritaires, etc. Ce sont des indications précieuses pour un suivi marketing de la diffusion ciblée dans un contexte donné.

Authentification

Deux niveaux d'accès à l'information sont gérés dans DECID. Au premier niveau, sans contrôle supplémentaire, il est possible de repérer des personnes et de savoir par quels points elles sont concernées dans le document soumis. Une présentation générale permet de situer chaque destinataire repéré (son rattachement, l'intitulé de ses projets).

Pour en savoir plus sur la teneur des activités des personnes indiquées par le système, on peut visualiser les textes complets de leur programme de recherche et avoir donc plus de détail sur leur activité. Ceci est évidemment un niveau d'information supérieur, et l'accès à ces données doit être protégé. Lorsque l'utilisateur actionne la fonctionnalité de visualisation d'un texte d'Action (programme de recherche), il lui est demandé un *mot de passe* (FIGURE 15), sans lequel l'accès au texte intégral lui est refusé. Si l'utilisateur est bien habilité, ce mot de passe ne lui est plus redemandé lors de la session.

DECID n'impose pas à l'utilisateur de retenir un compte et un mot de passe supplémentaire, mais il utilise les authentifications du système GALAXIE, le système central d'information et de documentation, donnant ainsi un *accès unifié* à toutes ces applications[28]. Mieux : si l'utilisateur vient du site GALAXIE et y a déjà fourni son mot de passe, alors ceci est pris en compte par DECID, qui ne le lui redemande pas au moment de l'affichage d'un texte intégral.

d) Quelques mots sur l'administration du système

Tout cet exposé est centré sur l'application, telle qu'elle se présente à l'utilisateur. L'administration de l'application, qui correspond aux tâches assurant le bon fonctionnement du serveur, constitue la face cachée de DECID pour l'utilisateur. Certes, l'utilisateur n'est pas concerné par l'administration... pour autant que ses dysfonctionnements éventuels ne la rendent pas pesamment perceptible. Le soin apporté à mettre en place les moyens d'une administration de qualité nous semble donc mériter quelques précisions ici.

Interface Web

De même que l'utilisateur accède à DECID via son navigateur Web, de même l'administrateur dispose d'une interface Web conviviale pour gérer l'application. Les actions de l'administrateur sont essentiellement les mises à jour (installation de nouveaux profils, révision et enrichissement de la documentation, etc.) et les paramétrages (choix d'un moteur d'indexation[29], données affichées, etc.).

[28] Ceci n'a l'air de rien, mais exige un cryptage convenable des mots de passe, et une synchronisation complète avec les données d'accessibilité de GALAXIE.

[29] Cela a permis par exemple un test du nouveau module de caractérisation des textes, en alternative au découpage simple.

Gestion de configuration

L'ensemble du code source des programmes informatiques liés à DECID sont organisés et gérés avec un outil de gestion de configuration[30]. On dispose ainsi de façon centralisée de l'ensemble des sources cohérentes et à jour (permettant de compiler et générer les programmes exécutables adaptés à tout environnement d'exploitation) et de l'archivage systématique et organisé des versions antérieures.

3. Constitution de la requête

a) Sélecteurs

Les sélecteurs sont réalisés par des menus déroulants : ils servent à paramétrer la recherche en fixant une valeur parmi un ensemble de possibilités. Volontairement dans DECID, l'introduction de paramètres de choix est limitée à l'essentiel. De fait, pour que l'utilisateur soit vraiment en mesure d'indiquer une valeur pertinente pour sa recherche, et d'ajuster le traitement, il faut qu'il sache clairement interpréter la signification et les incidences de chaque choix. La multiplication des paramètres de réglage n'est pas bon signe : démagogique, elle donne un pouvoir illusoire à l'utilisateur, qui se retrouve comme face à un tableau de commandes complexes qu'il ne maîtrise pas, et que complique définitivement la combinatoire démultipliée des interactions entre les différents paramètres.

Choix de la base

Avant même de soumettre son texte de requête, l'utilisateur est invité à indiquer à quel ensemble de profils il veut le confronter (FIGURE 4). Actuellement dans DECID, cela consiste à *choisir une année* (1995, 1996, 1997, 1998, ou 1999) : les personnes sont alors sélectionnées en fonction des points communs entre le document soumis et leur programme d'activité à cette année-là.

L'année la plus neuve est l'année en cours (la recherche s'effectue en fonction des activités actuelles), ou l'année suivante (on s'intéresse alors aux projets de programmes d'activité).

Il est également intéressant de pouvoir rechercher des personnes rétrospectivement, en particulier s'il s'agit de trouver une personne ayant acquis une expertise dans un domaine. Cela peut être utile aussi pour essayer de trouver un interlocuteur si la recherche a été infructueuse sur les années les plus récentes.

Maintenant que la liste des années (et des bases disponibles) s'allonge, on perçoit plus clairement le besoin d'une interrogation rétrospective synthétique, permettant de rechercher, en une seule opération, des personnes ayant travaillé sur un domaine « récemment », sans être obligé de procéder année par année. La synthèse de plusieurs années n'est pas nécessairement une simple réunion des bases, mais pourrait être une combinaison plus élaborée prenant en compte le fait d'avoir travaillé plusieurs années sur un sujet, ou/et dans les années les plus récentes. Cette sélection autre qu'annuelle n'est encore qu'une piste de recherche.

Choix d'un type de destinataires

DECID permet de rechercher des destinataires individuels, mais également des équipes ou des divisions de la DER qui travaillent sur un sujet. Pour faire suivre un document, on veut trouver des noms de personnes ; mais pour rechercher une entité qui puisse être intéressée par certaines compétences (transmission de CV), le destinataire adapté est un Groupe ou un Département.

Le choix actuel dans DECID est entre quatre type de destinataires : *agent* (une personne), *Groupe* (de l'ordre de la dizaine de personnes), *Département* (de l'ordre de la centaine de personnes), *Service* (quelques centaines de personnes) (FIGURE 5). Les profils de structure (collectifs) ne sont pas un simple cumul , une juxtaposition, des profils des agents, mais sont calculés de sorte à rendre compte de façon synthétique des orientations importantes de la structure.

[30] L'outil de gestion de configuration utilisé est CVS.

FIGURE 5 : Types de destinataires.

FIGURE 6 : Seuillage du volume des réponses (nombre maximal de propositions de destinataires).

Seuillage du volume de réponses

Limiter à l'avance le nombre maximal de réponses souhaitées est utile pour ajuster la longueur de la page présentant les résultats. Même si une présentation étudiée permet de naviguer méthodiquement d'une proposition à l'autre, et de la liste récapitulative globale au détail de chaque rapprochement, il est plus confortable de manier une liste de taille adaptée à ses besoins. Si l'on cherche une ou deux personnes au plus, il est encombrant de recevoir une centaine de propositions (qui plus est, si l'on compte imprimer la liste des résultats). Inversement, si l'on faire une recherche aussi complète que possible de toutes les personnes concernées par un document (une annonce de séminaire par exemple), il serait gênant que le système ne puisse pas proposer plus de vingt noms, s'il y a effectivement plus de vingt personnes susceptibles d'être intéressées.

L'introduction de ce paramètre (FIGURE 6) est donc utile pour ne pas imposer une limitation du nombre des réponses arbitraire, alors que les besoins sont variables en fonction des circonstances de recherche. Cependant, ce n'est pas toujours un paramètre très significatif pour l'utilisateur ; par défaut, le système s'arrête de lui-même à un nombre moyen de propositions (une vingtaine).

La raison d'être de ce paramètre s'affaiblit avec un affichage non plus linéaire mais thématique des résultats. En effet, alors que la liste s'allonge avec le nombre de réponses, l'arborescence thématique permet de parcourir les résultats par étapes, en n'ayant à chaque fois qu'un nombre raisonnable de propositions, et en cernant ainsi peu à peu les quelques-uns souhaités.

Paramètres textuels envisagés

D'autres menus, qui n'apparaissent pas encore sur l'interface actuels, sont prévus. Ils feraient également partie des sélections (par menu) au moment de l'entrée de la requête.

Il s'agirait d'indiquer au système des paramètres pour optimiser l'analyse du texte de requête. Deux paramètres sont envisagés, avec dans un premier temps les valeurs suivantes :

- sélecteur de *langue* : français (par défaut), anglais.

- sélecteur de *genre* : non précisé (par défaut), texte d'Action, CV, mots-clés, résumé.

La liste des genres n'est pas établie. En tout cas, il ne s'agit pas de décliner les genres textuels possibles, mais uniquement de faire apparaître la désignation de familles de textes qui peuvent bénéficier d'un traitement adapté (sans préjuger avoir identifié un genre, du point de vue d'une théorie des genres textuels). Conformément à l'esprit de DECID, cette liste resterait ouverte, c'est-à-dire prévoirait toujours la possibilité de laisser le genre indéterminé, le système n'ayant pas à imposer une norme restrictive à l'utilisateur.

Concrètement, indiquer, en soumettant le texte d'un Curriculum Vitæ, qu'il relève du genre CV, aurait pour effet d'éviter des rapprochements malheureux, en particulier ceux dus à la remotivation arbitraire de ses termes de construction. Par exemple, un intertitre *Langues vivantes* dans le CV ne justifie pas sa retransmission aux quelques équipes faisant de la linguistique à la DER. Ou encore, le fait que le candidat soit passé par une Université ne le rend pas nécessairement plus proche de n'importe quelle équipe qui a noué une collaboration avec une (autre) Université.

Quant au sélecteur de langue, il n'a de sens que lorsque DECID sera en mesure de traiter et confronter aux profils des textes en anglais. Peut-être une valeur « *mixte* » serait-elle à ajouter, pour prendre en compte des textes de requête mêlant les deux langues.

Pour la langue (et dans une moindre mesure pour le genre), une autre stratégie consisterait à avoir un module de préanalyse qui identifie automatiquement, en s'aidant d'indices discriminant, une valeur correcte pour le paramètre[31]. Mais n'est-il pas plus simple et plus efficace de recevoir l'indication de l'utilisateur ?

b) L'introduction d'un texte

DECID prévoit trois modes d'entrée d'un texte de requête :

- par simple frappe de quelques lignes, *au clavier* ;
- par *copier / coller* (notamment depuis une autre application, un traitement de texte par exemple) ;
- par navigation dans l'arborescence des répertoires du poste et sélection d'un *fichier* contenant le document à soumettre.

Le *copier / coller* est souvent préférable à la frappe, car il favorise la soumission d'un texte plus riche (apport de contexte). Il est facilité aussi par la robustesse du traitement, qui tolère parfaitement des extraits approximativement ajustés (coupure au milieu d'une phrase par exemple, ou autres juxtapositions agrammaticales), ou les éléments périphériques collectés dans un *copier tout* (par exemple le copyright) (FIGURE 7et suivantes).

Cette capacité de l'application à tirer parti quasiment de n'importe quel matériau « textuel » n'est-elle pas en opposition avec une recherche fine sur la textualité ? Est-il cohérent d'autoriser et de traiter sans broncher une juxtaposition de mots ou de bribes de texte ?

De fait, l'approche est résolument descriptive plutôt que normative. L'objectif est de laisser la plus grande liberté expressive à l'utilisateur, sans préjuger de sa forme effective (dans la hâte d'un moment, il peut être plus expressif de juxtaposer quelques mots-clés, ou de coller quelques extraits, que de composer ou recopier quelques phrases de synthèse). Toutefois, l'utilisateur est responsabilisé quant à la pertinence de se requête : tout est permis, mais tout ne convient pas[32]. Le rôle du système est de faire au mieux avec ce qui lui est donné, éventuellement d'avertir l'utilisateur de la difficulté à exploiter sa requête dans le contexte de la base des profils et de son mode de traitement.

---

[31] Une tactique est de reconnaître les mots grammaticaux, différents d'une langue à l'autre, et fréquents donc présents dans tout texte (ce serait moins vrai pour analyser une liste de mots-clés), cf. par exemple (Wechsler, Sheridan, Schäuble 1997).

[32] Dans le cadre de l'étude des dialogues homme-machine, l'attitude de l'utilisateur (qui a besoin de recourir au système) est coopérative, il n'a aucun intérêt à chercher à se heurter aux limites de la machine :

« En règle générale, l'utilisateur n'adresse au système que des requêtes pertinentes, ou du moins auxquelles il peut raisonnablement attendre une réponse. Pour épargner son temps, il se montre coopératif et restreint le champ du dialogue aux domaines sémantiques représentés dans le système, pour autant qu'il les connaisse.

En revanche, les informaticiens se font un point d'honneur de tester les limites des systèmes, et de les faire échouer, sous le prétexte non futile de les améliorer. Ils agissent alors en expérimentateurs plutôt qu'en utilisateurs quelconques. » (Rastier 1991, §VI.3, p. 170)

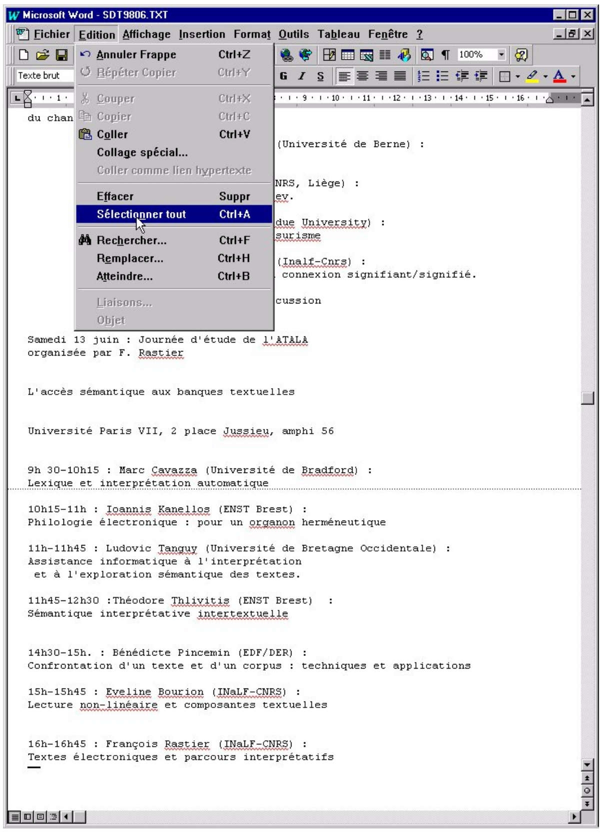

FIGURE 7 : Le document à diffuser.

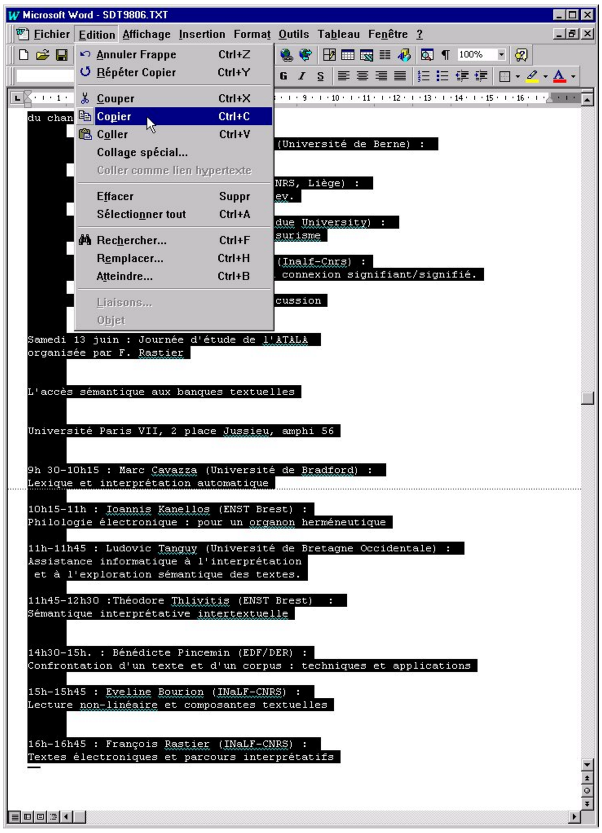

FIGURE 8 : Saisie par *Copier* standard.

La requête est textuelle non parce qu'elle devrait vérifier certains critères de bonne formation, qui fixeraient ce qui aurait le droit d'être appelé texte, mais parce que l'on choisit de considérer comme un texte le contenu de la requête. La textualité n'est pas une définition formelle, mais une décision herméneutique. Les quatre facettes textuelles identifiées lors de l'étude de la textualité peuvent éclairer cela. Comme l'y invite la première facette textuelle, le traitement va chercher à discerner et identifier des unités et structures linguistiques (des mots, le mode d'un verbe, etc.) : il ne « voit » pas simplement une suite de caractères ou une suite d'octets, qui pourraient coder tout autre chose. Avec la deuxième facette, le traitement se rend sensible à ce qui traduit une cohérence d'ensemble de la requête, il s'appuie éventuellement sur le rythme des divisions en paragraphes (qu'est-ce que chacun apporte), il peut y lire une progression si celle-ci enrichit son analyse. La troisième facette conduit à une lecture de la requête dans le contexte intertextuel de la base, ou peut-être en se référant à un genre connu de l'application (CV, texte d'Action...). La quatrième facette est déjà reconnue par le fait même d'accepter de considérer comme un texte ce que l'utilisateur soumet. Les procédés de surlignage et de relief sur le texte de requête, dont il est question plus loin, sont également une forme de mise en œuvre de cette facette.

Une manière aussi de respecter la textualité de la requête, est de rendre compte de sa disposition spatiale, de son volume, de l'équilibre (visuel) relatif entre les paragraphes (distinction entre des paragraphes plus gros ou plus petits). Pour ce faire, on a tenu à afficher le texte de requête en opérant une *césure automatique des lignes* : sinon, chaque paragraphe est figuré par une ligne démesurée, et la perception du déploiement du texte à travers les différents paragraphes est perdue.

Il a fallu prendre en compte la possibilité de voir soumis en requête des *textes « longs »*. Ce cas ne serait guère réaliste pour des requêtes entrées au clavier, mais le devient par l'intermédiaire du copier / coller. Des tests montrent que le mécanisme classique de communication (par cgi) ne permet pas de transmettre un texte de l'ordre d'une vingtaine de pages ou plus. Une manière de contourner cette limite est de communiquer au serveur non plus le texte lui-même, mais l'adresse d'un fichier où trouver le texte.

FIGURE 9 : Entrée de la requête par un *Coller*.

FIGURE 10 : Requête prête à être lancée.

Selon la source et l'usage du texte, celui-ci est codé selon un certain format, et DECID peut en l'occurrence avoir affaire à différents formats. Ceux qui sont reconnus et traités sont :

- le *texte simple* (correspondant aux fichiers traditionnellement suffixés en `.txt`) : le texte est mis en forme par quelques caractères de contrôle (tabulation, retour chariot).
- le texte balisé, selon une DTD SGML, en particulier *HTML* : on soumet par exemple à DECID des pages Web telles qu'elles sont récoltées par un « aspirateur Web », à savoir la forme source de ces pages (comprenant les balises) et non ce que l'on recueille par copier / coller à partir de l'affichage d'un navigateur (texte sans le balisage).

Le cas des fichiers au format du *traitement de texte* standard à EDF-DER[33]est étudié mais non résolu, notamment en raison de la diversité des formats en fonction des versions successives du logiciel. Pourtant, la prise en compte de ce cas serait utile, car c'est en effet le format de la plupart des requêtes faites en indiquant un fichier.

La recherche de la thèse a notamment consisté à tirer parti des informations de structuration textuelle, pour les deux premiers formats. Par exemple, au lieu de traiter le format balisé en effaçant tout ce qui a la forme d'une balise (ce qui est déjà mieux que de considérer les balises au même plan que des mots), il s'agit de reconnaître ces balises, puis de les interpréter, en en tirant et en gardant des informations que l'on juge sémantiquement importantes (ce qui est affiché comme titre, ce qui constitue un paragraphe, etc.)

c) *Actions sur le texte de requête*

D'une manière générale, les interventions de l'utilisateur sur le texte de requête (retouches, surlignage) ne sont pas accessibles dans le cas où le texte a été soumis par indication d'un fichier,

[33] Pour faciliter l'échange des documents, un cadre de cohérence à l'échelle de l'entreprise fixe un traitement de texte standard (actuellement, une certaine version du logiciel *Microsoft Word*).

pour qu'il n'y ait pas d'ambiguïté :car il ne serait pas clair de savoir ce qui est modifié, le fichier lui-même ou sa copie provisoire dans la fenêtre de requête.

Mises à blanc

Dès les toutes premières versions, DECID comportait l'action, simple mais très utile, d'*effacement du texte de requête* (bouton sous l'affichage du texte, FIGURE 10). Cette action sert à préparer la fenêtre à une nouvelle requête. En effet, le texte de la requête courante n'est pas automatiquement effacé, pour pouvoir être retouché ou complété au fur et à mesure de la recherche. De cette manière, l'utilisateur est pleinement maître de l'évolution de sa stratégie d'interrogation (par affinement ou par rupture). Il est donc important de comprendre que cette fonctionnalité va de pair avec la possibilité de retour au texte de requête après une recherche.

Avec le surlignage possible de la requête (présenté ci-après), un autre mode de mise à blanc est prévu. Il s'agit en effet de pouvoir *revenir au texte initial*, vierge de toute mise en relief, sans avoir ni à retirer chaque marque de surlignage une à une, ni à effacer le texte lui-même.

Surlignages

En soumettant un texte, l'utilisateur peut vouloir indiquer à DECID que c'est sur tel ou tel point qu'il recherche plus particulièrement des interlocuteurs ou destinataires. Le texte a un relief pour lui : tel paragraphe est central, ou problématique, ou innovant ; telle expression précise mentionne un produit sur lequel il voudrait en savoir plus.

C'est une sorte de surlignage ou d'annotation, qui s'ajoute au texte, qui rend naturellement compte de ces saillances. Il y a des marques qui repèrent un mot, un élément : un trait qui souligne ou encercle un mot, une flèche ou une croix dans la marge, qui repère la ligne. D'autres formes de traces indiquent des passages, typiquement un trait (droit, courbe, ondulé, double) dans la marge sur toute la hauteur du passage retenu. En somme, il y a un *surlignage horizontal*, de très courte étendue, qui signale un mot précis, et un *surlignage vertical*, qui porte sur plusieurs lignes, et marque la présence d'une idée qui retient l'attention et sur laquelle on souhaite revenir.[34]

L'interface traduit ces deux modes de surlignage dans la lignée des conventions d'ergonomie des traitements de texte.

Dans la fenêtre où est édité le texte de requête (FIGURE 11), cliquer (avec la souris) deux fois sur un mot le sélectionne : il est alors coloré[35] dans une teinte vive (le reste du texte est et reste en affichage normal, noir sur fond blanc). Le mot est marqué par un surlignage horizontal, qui le distingue dans le « flot » du texte comme une expression importante, remarquable. L'utilisateur indique ainsi qu'il est intéressé de trouver des personnes ayant mentionné cette expression précise dans les textes composant leur profil.

Un triple clic souris sélectionne le paragraphe : le paragraphe est coloré d'une autre couleur que le mot par un double clic, il reçoit une teinte sensiblement différente et éventuellement moins vive, puisqu'il couvre une zone plus étendue. Il s'agit là bien sûr de marquer un surlignage vertical, qui s'attache à un passage pour l'idée qui y est développée. Ce que l'utilisateur signifie alors, c'est que les points présents dans ce passage sont plus importants pour sa recherche que les autres parties du texte non surlignées, et il souhaiterait particulièrement retrouver des personnes ayant (eu) une activité en rapport avec ce qui est décrit dans le passage. Cela n'implique pas nécessairement que cela ait été formulé de la même manière, avec les mêmes mots, dans le texte de requête et dans les textes descriptifs des profils.

---

[34] On pourrait aussi concevoir d'autres types d'annotations (voir par exemple (Virbel 1994)), par lesquelles l'utilisateur remodèle le texte de requête à l'intention du système : l'indication de liens et la mise en relation de passages (représentables par les éléments `LIST` et `ITEM` de la DTD Corpus), l'insertion d'additifs sur certains points (*parties à fonctionnement infra-textuel* dans les termes de notre modèle).

[35] On pourrait préférer, pour imiter l'apparence d'un surlignage, de colorer le fond derrière le mot ; cela est plus difficile avec les technologies disponibles que de changer la couleur des caractères du mot.

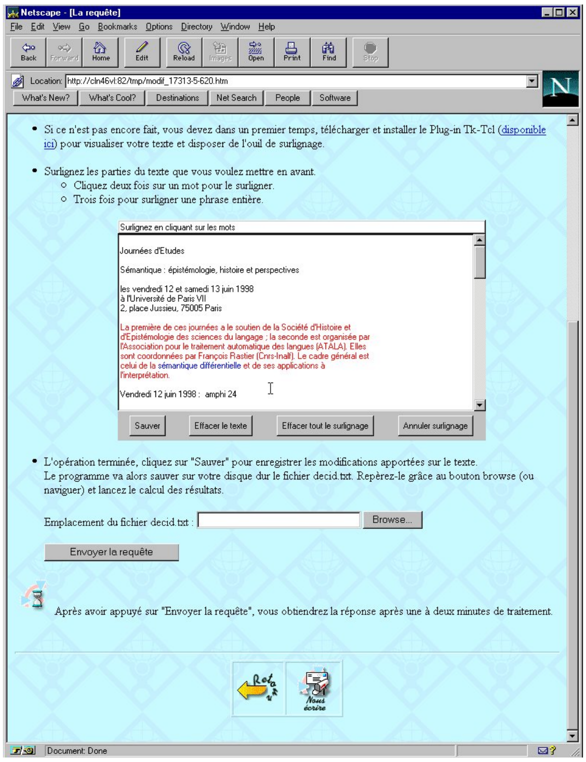

FIGURE 11 : Surlignage de la requête.

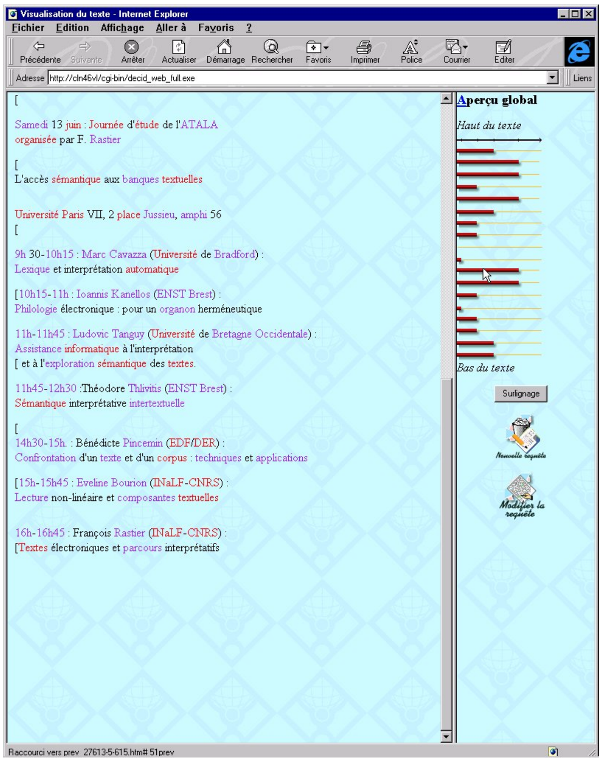

FIGURE 12 : Aperçu de la *Prise en charge* de la requête, avant la recherche de destinataires.

Surlignage horizontal et surlignage vertical se combinent. Si un ou quelques mots sont surlignés horizontalement et que l'utilisateur décide finalement de sélectionner verticalement tout le paragraphe où ils figurent, alors le paragraphe entier passe en surlignage vertical, le surlignage horizontal étant fondu dans le surlignage vertical. On veut ainsi rendre compte d'un utilisateur qui repère d'abord quelques mots, puis comprend progressivement que c'est finalement tout un passage qui l'intéresse, et l'idée exprimée dans ce passage davantage que les mots eux-mêmes. Si l'utilisateur souhaite maintenir qu'une expression particulière précise est importante, ou qu'il la remarque une fois le surlignage vertical d'un paragraphe effectué, il peut toujours lui appliquer un surlignage horizontal, « par dessus » le surlignage vertical.

Nous avons vu qu'un bouton permet d'effacer d'un seul coup tous les surlignages. Il est aussi possible de les défaire un à un, par la même opération que celle qui les a créé : un double clic annule le surlignage horizontal d'un mot, un triple clic enlève le surlignage vertical d'un paragraphe. Cela donne donc une certaine souplesse de travail et d'annotation sur le texte de requête. Une fonctionnalité envisageable, mais qui ne ferait peut-être pas tant gagner que cela[36], serait une commande de type *undo / redo*, capable d'effacer les surlignages (*undo*) dans l'ordre inverse de leur création (en commençant donc par enlever les derniers, les plus récents), et de rétablir les surlignages effacés (*redo*), dans l'ordre de leur création.

4. Informations sur le traitement

a) *La forme de la requête*

Un *message d'avertissement* est affiché *si la requête n'est que de quelques mots*, au lieu de se présenter comme un texte. Le but est de sensibiliser l'utilisateur au fait qu'il n'utilise pas DECID de

[36] Gain en charge mémorielle, mais perte en souplesse. *A priori*, les premiers surlignages sont les plus « sûrs », et sont donc effectivement les derniers sur lesquels revenir ; également, on se ravise habituellement plutôt sur ce que l'on vient de faire (récence). Cependant, l'effacement est chronologiquement systématique, et une annotation intéressante incluse dans une série que l'on veut effacer est prise dans le mouvement d'annulation.

façon optimale (en particulier sa requête risque de souffrir d'un manque de contexte). Le système avertit ainsi qu'il est prévisible que les résultats de sa recherche soient médiocres et que ceci n'est pas (entièrement) imputable au système. Plus positivement, il fait prendre conscience à l'utilisateur qu'en enrichissant quelque peu sa requête, il peut faire une recherche beaucoup plus satisfaisante.

Ce rappel contextuel du caractère textuel de la requête DECID paraît une manière opportune d'éviter le cas, sinon trop fréquent, d'utilisateurs déçus parce que n'ayant jamais interrogé DECID que par un ou deux mots. Le poids des habitudes acquises avec les systèmes documentaires et moteurs de recherche est tel, que les utilisateurs ne « voient » pas la demande qui leur est faite, dans DECID, d'utiliser un texte. Le message d'avertissement est alors un moyen de l'informer de l'opposition, importante pour DECID, entre mots-clés et texte : DECID attend de façon privilégiée une requête textuelle, les mots clés ne sont pas ici un moyen d'interrogation efficace mais marginal.

b) La prise en charge : adéquation de la base à la requête

L'utilisateur a opté pour une base de profils (par exemple, celle de l'année courante), et il a entré son texte de requête. DECID est alors en mesure de lui donner des indications sur la manière dont sa requête entre dans le cadre général de la base ou non. *L'utilisateur voit, dans son texte, ce qui correspond à la base et ce sur quoi il n'y a pas de réponse à attendre* (FIGURE 12). C'est en effet une information importante pour éventuellement renoncer à la recherche telle qu'elle est définie (le passage du texte qui était le plus intéressant pour l'utilisateur n'a aucun écho dans la base choisie), ou pour mieux interpréter les résultats (inutile de parcourir longuement la liste des destinataires suggérés à la recherche d'une personne spécialement concernée par un passage du texte qui n'est de fait pas couvert par la base). La prise en charge de la requête est donc une information générale et anticipée sur le potentiel de la recherche.

D'un point de vue herméneutique, il est intéressant aussi de noter que c'est une lecture nouvelle, du point de vue de la base, du texte fourni par l'utilisateur. Le texte de requête est pour l'utilisateur un texte qu'il connaît bien, par rapport auquel il a ses propres repères, ou bien au contraire ce peut être un document qu'il a à retransmettre et dont il a fait une lecture rapide, sans en percevoir tous les détails significatifs. Dans ces deux cas, la confrontation du texte à la base apporte un autre point de vue, qui interagit avec celui de l'utilisateur. Par exemple, celui-ci s'aperçoit que son texte a des passages particulièrement « attracteurs » par rapport à la base des profils, et que d'autres passages se démarquent nettement des préoccupations de la DER (pour l'année choisie) (passages trop théoriques, domaine ou discipline hors du champ de travail de la DER, etc.).

Dans les premières versions de DECID, avant qu'il y ait cette fonctionnalité de *prise en charge*, il avait néanmoins semblé intéressant de donner à l'utilisateur la forme prise par sa requête après analyse. Concrètement, cela prenait la forme d'une liste, ordonnée alphabétiquement, des descripteurs attribués au texte après indexation automatique, ou bien des mots issus du découpage du texte et restant après élimination des « mots vides ». L'utilisateur avait ainsi des éléments supplémentaires pour comprendre sur quoi s'était basé le système pour le calcul des rapprochements. Dans le cas de l'indexation automatique en particulier, il pouvait constater si cette représentation était complète (le thesaurus peut manquer de descripteurs pour décrire certaines parties), si elle comportait des erreurs (par exemple des contresens, se répercutant ensuite dans les résultats). Avec le passage à un traitement uniquement par découpage, cet affichage de la représentation interne du texte s'est révélé inutile et même nuisible. Inutile, car le découpage est prévisible et sans surprise. Nuisible, car l'utilisateur voit le caractère fruste de la représentation, et s'en inquiète. Or le mécanisme de calcul des rapprochements est relativement puissant (pondérations et valorisation du contexte via les groupements de mots), et l'utilisateur « oublie » d'apprécier les résultats, alors qu'il y a souvent de bonnes propositions de destinataires. L'enseignement tiré de cette expérience, est qu'il n'est pas bon de fournir une représentation *interne* à l'utilisateur, puisqu'il n'est pas à même de savoir la signification et la portée qu'elle prend dans le cadre du traitement ; il est « impressionné » (en bien ou en mal) et cela occulte la valeur réelle que peuvent prendre pour lui les résultats. La *prise en charge* évite donc cette erreur d'ergonomie : elle donne des indications à travers le texte lui-même, sous la forme « externe » qu'il a pour l'utilisateur, et non sous la forme « interne » d'une représentation destinée aux calculs[37].

L'indication de prise en charge est donc un moyen approprié pour que l'utilisateur perçoive de façon aussi juste que possible (car en se plaçant dans sa vision du texte) sur quoi vont porter effectivement les calculs, et par conséquent comment il peut réajuster sa recherche[38]. Sans ce genre de guidage, l'utilisateur serait capable d'aller à rebours de ce qui aiderait sa recherche (par exemple, par suppression maladroite de contexte), car l'appréciation ou l'intuition de l'utilisateur ne correspondent pas nécessairement à ce qui est efficace pour le système. Avec l'affichage de la prise en charge, l'utilisateur prend conscience des éléments les plus influents, des modifications qui remanieraient les résultats de la façon la plus sensible ; cela donne à l'avance une idée de la stabilité des résultats (à quoi tient-elle) et oriente éventuellement une nouvelle interrogation du système.

Précisons donc quels types d'informations apporte maintenant la *prise en charge*, et comment elles s'affichent. Le texte n'est en fait pas confronté à tous les profils (c'est ce que fait le traitement lui-même), mais, ce qui est plus rapide, est évalué dans le cadre de l'univers d'unités descriptives disponibles (autrement dit, le « dictionnaire » ayant servi à caractériser tous les profils). Le texte de requête présente des unités[39], qui vont servir à construire sa représentation dans le calcul, et ce sont ces unités qui sont étudiées au regard de l'univers d'unités associé à la base.

La première information que l'on peut avoir, c'est de signaler ce qui dans le texte de requête est *inconnu* de l'univers de description, et donc sera ignoré du traitement. L'utilisateur est invité à interpréter cette information en contexte, pour comprendre si c'est un mot particulier qui est ignoré, ou si c'est manifestement tout un aspect du thème qui sort du champ de la base.

La seconde information de prise en charge consiste à mettre en évidence ce qui, dans ce qui est reconnu, est également particulièrement *susceptible de contribuer significativement* à des rapprochements. Grossièrement, les mots grammaticaux, les termes généraux, les formulations standard[40] sont connus du système mais n'ont pas un rôle « intéressant », alors qu'un terme technique par exemple peut se présenter comme ayant une certaine valeur dans la base.

Le texte de requête se voit donc appliquer un code de deux couleurs, qui traduit trois valeurs :

- dans une couleur froide, les mots ou expressions non connues du système, pour la base considérée : il peut être important pour l'utilisateur de voir qu'ils sont « hors d'atteinte » ;
- dans une couleur chaude, les mots ou expressions susceptibles d'être actifs dans la sélection : on fait ici ressortir les « points chauds » du texte, vers lesquels vont s'orienter les résultats ;
- les portions de texte laissées telles quelles (écriture en noir) est ce qui est connu du système, mais est perçu comme sans intérêt particulier ; cela reste comme fond, sans saillance ni donc couleur particulière.

Le texte est donc affiché dans une fenêtre classique, avec une écriture alternativement noire ou colorée. Il est en outre bordé par un histogramme, qui est conçu plus généralement accompagner

---

[37] (Gros, Herviou-Picard 1996) rencontrent la même distinction, entre présentation interne et présentation externe, dans le cadre de la constitution et de l'exploitation de terminologies. Si l'extraction de termes est utilisée directement par la machine (présentation *interne*, non vue de l'utilisateur mais employée pour des calculs), le bruit (i.e. la présence de mauvais candidats termes) est peu ou moins gênant : l'influence des termes en trop est généralement réduite, et la machine n'est pas submergée par un volume important de données. La perspective est totalement différente si les termes extraits doivent être présentés à un expert pour validation : la présentation (*externe*) doit être outillée par une interface adaptée et accompagnée d'aides à l'interprétation.

[38] Dans (Veerasamy & Belkin 1996) qui s'intéressent à des requêtes de quelques mots-clés, ce genre d'information est donné *a posteriori*, dans une représentation suggestive (*ibid*., p. 89) qui affiche parallèlement l'ensemble des profils d'influence de chaque terme de requête sur la collection des documents sélectionnés. Cela fait clairement ressortir : les termes de requête qui ont un rôle faible et uniforme, ceux qui ont une influence forte et contribuent de façon décisive à la sélection des résultats, ceux qui sont associés à un sous ensemble particulier de la de la sélection.

[39] Il s'agit ici des unités élémentaires, à la « surface » du texte ; les comparaisons se font sur la base des unités élémentaires, puis en opposant les unités élémentaires qui contribuent à des unités descriptives sémantiquement riche, à celles qui s'avèrent jouer un rôle peu significatif dans la constitution des unités descriptives.

[40] Tout ceci est relatif, à un corpus, ou à la description d'un genre (opérée elle-même à partir d'un corpus).

tout texte potentiellement long et auquel on a associé un certain relief (la mise en valeur d'éléments, une mesure qui varie au fil du texte, etc.).

L'histogramme marginal : articulation global / local dans le parcours du texte

La lecture et l'interprétation d'un texte fait appel à une représentation à la fois de l'ensemble du texte (son volume, sa dominante générale,…) et du passage que l'on est en train de déchiffrer. Les deux s'articulent, par exemple la localisation d'un passage, son étendue, peuvent jouer sur le statut que le lecteur accorde à tel ou tel élément.

DECID accompagne la fenêtre de visualisation du texte (locale) par un *histogramme, associé à l'ascenseur (indicateurs globaux)*. Cet histogramme apparaît dans l'affichage de la *Prise en charge* (FIGURE 12), de projections d'un texte sur un autre (FIGURE 17).

Les barres de l'histogramme correspondent à des tranches régulières dans le déroulement linéaire du texte[41] (comme les feuillets d'un livre). L'ensemble de l'histogramme figure donc comme l'épaisseur d'un livre fermé –le livre du texte–, chaque barre ébauchant le fin sillon d'une page, à laquelle ouvrir le livre et trouver un certain passage (au début, au milieu, à la fin,...).

L'histogramme est associé à l'affichage (local) du texte dans la fenêtre. *Cliquer sur une barre positionne le texte de la fenêtre à l'endroit correspondant* (on ouvre le livre à un endroit choisi dans l'épaisseur de sa tranche). De plus, si l'on ajuste la hauteur de la fenêtre à la base de l'histogramme, les barres en face de l'ascenseur sont celles en relation avec le passage affiché. De fait : et la colonne dans laquelle se déplace l'ascenseur de la fenêtre, et la base de l'histogramme, sont des *représentations homothétiques du déroulement linéaire du texte*. Autrement dit, la position haute de l'ascenseur, comme la première barre de l'histogramme, correspondent au début du texte, idem pour la fin du texte (position basse, dernière barre). Et la taille de l'ascenseur figure bien la proportion du texte qui est visible dans la fenêtre : si c'est une portion importante, l'ascenseur est un rectangle allongé, et les barres de l'histogramme qui sont en face de lui détaillent en plusieurs tranches ce qui est contenu dans la fenêtre.

La taille des barres de l'histogramme traduit la *valeur d'une mesure au fil du texte*. Chaque barre représente, par sa longueur, la valeur de la mesure pour la tranche du texte qu'elle décrit.

Dans le cas de la *prise en charge* (du texte de requête par rapport à la base) plus la barre de l'histogramme est longue, mieux la partie correspondante du texte est prise en charge par le système. Les vallées de l'histogramme soulignent les « zones d'ombre », échappant à la suite des traitements. Le choix actuel est que la longueur de la barre exprime la proportion d'occurrences d'unités connues et actives par rapport au nombre total d'occurrences saillantes (actives ou bien inconnues) dans le passage : une barre complète, maximale, reflète un passage entièrement pris en charge ; une barre très courte, quasi inexistante, est le signe d'un passage mal couvert par la base. Le cas (théorique ?) d'un passage qui n'aurait aucune unité surlignée se voit affecter une barre de longueur intermédiaire, ni minimale ni maximale (valeur neutre de 50 %).

c) Le temps de traitement

Il est essentiel pour l'utilisateur de savoir un ordre de grandeur du temps de traitement : s'agit-il de quelques secondes (immédiat) ? d'une vingtaine de secondes (semi-interactif) ? de plusieurs minutes (différé rapide) ? etc. Si l'utilisateur s'attend à avoir une réponse en une dizaine de secondes et que le temps de traitement réel soit de quelques minutes, la situation est inconfortable (l'utilisateur s'inquiète et s'impatiente, il reste mobilisé par l'attente du résultat) et inefficace (il y a des chances que l'utilisateur interrompe le traitement en supposant qu'il s'est « planté », alors même que le traitement était en train d'aboutir). Il a donc besoin d'un bon ordre de grandeur pour s'organiser (attendre un résultat immédiat / mettre à profit les quelques minutes prévisibles).

[41] Donc en particulier, la répartition du texte par rapport aux barres ne se calcule pas à partir des lignes au sens informatique du terme (suite de caractères terminée par un retour chariot). En effet, la ligne informatique ne rend pas compte des différences de taille entre les alinéas, alors que ces tailles sont généralement très contrastées sur les textes de quelques pages soumis à DECID.

Les différences entre temps de traitement sont très sensibles. Avant d'être accessible par une interface Web, la diffusion ciblée fonctionnait par messagerie. La requête était envoyée dans le corps d'un message adressé à l'automate de diffusion ciblée, qui répondait en fournissant une liste de destinataires potentiels. La messagerie, mode de communication asynchrone, est tout à fait compatible avec des temps de traitement de quelques minutes. L'interface Web instaure, elle, une communication synchrone, interactive : l'utilisateur est face à son écran, écran qui est en grande partie occupé par la fenêtre de l'application, et pour que l'interaction soit efficace l'utilisateur a besoin d'un retour quasi immédiat. Pour DECID, l'objectif est alors d'assurer des temps de traitement toujours inférieurs à la minute, voire toujours de moins de trente secondes[42]. Une *estimation du temps de traitement* est toujours fournie, et la durée effective du traitement est donnée avec les résultats.

La formule estimant le temps de traitement dépend des opérations à effectuer (l'algorithme), des ressources de la machine, mais aussi de la longueur et de la composition de la file d'attente. Les problèmes de file d'attente sont surtout sensibles si le traitement fait appel à un serveur externe, effectuant un traitement complexe potentiellement long, et susceptible de recevoir un flux de requêtes venant d'autres sources. Pour DECID, ce peut être le cas s'il fait appel à un serveur de Traitements Automatiques du Langage Naturel, qui peut très bien être en train de travailler sur un corpus volumineux au moment où DECID a besoin de lui.

Si le système disposait d'informations sur le déroulement du traitement, et que le traitement soit généralement de plus de quelques secondes, l'estimation du temps de traitement pourrait être complétée par un suivi de l'avancement du traitement, montrant de façon visuelle quelle portion du traitement est déjà effectuée.

5. Communication des résultats

a) Informations de présentation de chaque personne

Un destinataire est identifié par son *nom* et son *prénom* (FIGURE 13). L'indication *M.*, *Mme* ou *Melle* est utile, car l'utilisateur ne connaît *a priori* pas toutes les personnes, et certains prénoms ne permettent pas de savoir si l'on a affaire à un homme ou une femme ; or il est préférable pour l'utilisateur de le savoir s'il doit joindre la personne, par courrier ou par téléphone (pour lui envoyer un document, lui demander un avis). Enfin, à EDF, l'indication du *matricule* départage le cas échéant les homonymes.

La personne est située par son positionnement dans l'entreprise, ici à l'intérieur de la Direction des Etudes et Recherches. DECID donne son *rattachement* complet (*Service*, *Département*, *Groupe*), avec le *code* et le *libellé* de chaque unité. L'utilisateur perçoit ainsi rapidement dans quelle optique la personne aborde le thème du document soumis.

Une information plus précise est ajoutée avec les intitulés des projets dont la personne est responsable : ce sont les *titres de ses textes d'Action*, qui ont servi à calculer son profil.

Tout ceci constitue des indications brèves et est systématiquement affiché, pour tout destinataire proposé (FIGURE 14). Des informations plus détaillées, par exemple des textes qui serviraient à présenter la personne et son activité, seraient plutôt à proposer en consultation, parallèlement à la liste générale des destinataires.

Remarquons qu'ici, est superposé ce qui pourrait être distingué dans une autre application de DECID : les textes qui servent à construire les profils, et les textes qui peuvent être affichés pour présenter la personne. Nous continuerons ci-après à parler des textes d'Action, mais à ce stade ce n'est plus en tant que source des profils, mais en tant que *textes de présentation*, qu'ils sont utilisés.

[42] Eventuellement, un temps de traitement de quelques minutes, voire d'une dizaine de minutes, serait envisageable, pour une version « avancée » de DECID, qui justifierait cette attente par une qualité de traitement très supérieure. Quoi qu'il en soit, une version rapide et interactive est une base indispensable.

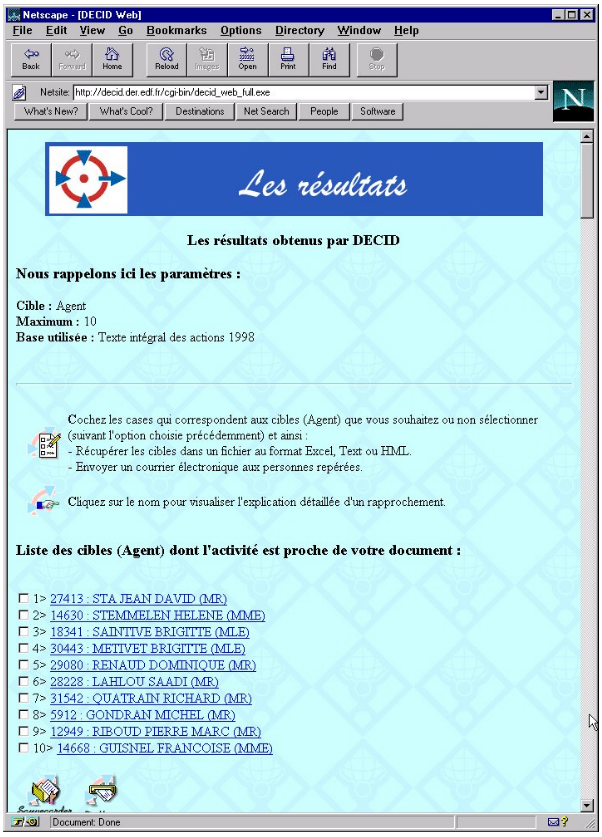

FIGURE 13 : Liste des destinataires proposés.

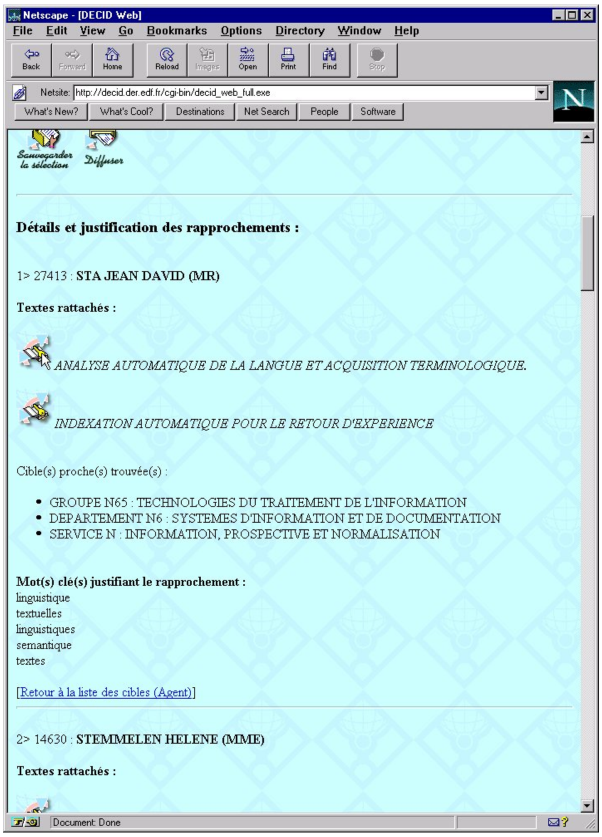

FIGURE 14 : Informations sur un rapprochement.

b) Explication : motifs d'un rapprochement requête-profil

Mots (ou unités) ayant le plus contribué au rapprochement

Le profil et le texte de requête ont été rapprochés grâce à un certain nombre de mots[43] qu'ils ont en commun. DECID affiche les *cinq mots* en commun qui ont eu la plus forte contribution au rapprochement.

Cinq est un maximum, il peut arriver qu'il y ait moins de cinq mots « significativement » en commun. D'ailleurs si la requête était quelques mots-clés, il est fort probable qu'il y en avait moins de cinq, et alors le nombre de mots en commun est au plus le nombre de mots de la requête.

La liste complète des mots en commun, même après retrait des mots grammaticaux, est souvent relativement longue (une vingtaine de mots ou plus) ; ordonnée par valeur de contribution au rapprochement décroissante, elle comporte une queue, constituée d'une accumulation de mots ayant une faible contribution, et souvent un apport sémantique faible dans le contexte de la base. Il serait donc peu utile et visuellement lourd d'afficher cette liste complète.

Prendre les cinq premiers mots convient pour donner un aperçu immédiat de la qualité et de la teneur principale du rapprochement (FIGURE 14). Bien sûr, ces mots révèlent en partie les thèmes en commun. Mais aussi, si l'on a affaire à quelques mots peu saillants, et sans relation significative, on conclura rapidement à une proposition sans valeur. En revanche, une forte cohésion de ces termes explicatifs autour d'un centre d'intérêt signale un rapprochement très vraisemblablement intéressant. Ce que l'on peut linguistiquement traduire comme suit : *l'isotopie entre les quelques mots centraux dans le rapprochement est un bon indice de la motivation sémantique effective du rapprochement*.

[43] Avec le nouveau moteur de caractérisation des textes, ce ne sont plus directement et nécessairement des mots qu'ont en commun les deux textes, mais des unités descriptives, qui peuvent s'être manifestées sous une forme différente dans chaque texte.

La projection : une lecture de la requête selon le point de vue du profil

Pour aller plus loin qu'une liste de quelques mots-clés, il faut rendre à ces malheureux mots leur(s) contexte(s). Le premier contexte contexte envisageable est celui de la requête. Une aide à l'interprétation du rapprochement est alors la *présentation du texte de requête, dont les mots ou les passages en relation avec le profil considéré sont mis en valeur*.

Cette projection (FIGURE 17) repère, de façon très visuelle, si le texte de requête concerne entièrement ou partiellement le destinataire envisagé, et dans le second cas, quelle partie du texte pourrait davantage rejoindre ses préoccupations. Comme la projection est d'autant plus riche que le texte de requête est assez développé et détaillé, elle encourage la soumission d'un texte plutôt que de quelques mots-clés.

C'est également pour l'utilisateur un jeu de relectures de son texte : « et si untel lisait ce texte, qu'y verrait-il peut-être plus particulièrement ? ... Tiens, il y a ceci qui ressort, je ne pensais pas pourtant que cela valait attention... ».

Enfin, c'est également la seule contextualisation à laquelle on ait accès, si le texte des profils n'est pas consultable pour des raisons de confidentialité, et que l'on ne dispose pas de textes de présentation associés aux profils. Dans ce cas, la projection peut même être volontairement légèrement moins précise, en signalant les passages où il y a des points communs mais sans indiquer précisément les mots.

L'accès au texte intégral, guidé par le texte de requête (index contextuel)

Les textes ayant servi à construire les profils ont un intérêt certain, pour comprendre les thèmes d'activité du destinataire proposé et mieux percevoir en quoi il est concerné par le texte de requête. L'idée est donc non seulement de pouvoir consulter un texte descriptif d'activités de la personne[44] mais aussi de le lire selon la perspective de la requête.

Pour ce faire, *tous les mots ayant significativement contribué au rapprochement sont mis en valeur[45] au fil du texte* : on voit donc où se distribuent les « atomes crochus » (FIGURE 16).

Cependant, ces indices de pertinence sont alors diffus, et il faut parcourir tout le texte pour vraiment en prendre connaissance. L'ajout d'un *index*, généré dynamiquement et adapté à la description de ce rapprochement, apporte une solution en offrant une autre forme d'accès et de parcours. Cet index est une liste, en marge du texte, de *tous les mots ayant significativement contribué à la sélection* du profil par rapport à la requête, en commençant par les plus influents (FIGURE 16). Pour chacun des mots de l'index, son *nombre d'occurrences* est indiqué, et des *liens hypertextes* permettent de localiser instantanément chaque occurrence, au sein de son contexte dans le texte. La lecture s'organise alors par des aller-retour entre l'index et le texte, le premier donnant des points d'entrée personnalisés (correspondant aux points communs entre le texte du profil et la requête), le second replaçant ces termes dans leur contexte d'évocation. L'index n'existe pas indépendamment du texte et de la requête : il est *transitoire* et *a posteriori*.

La lecture pourrait se prolonger en s'élargissant à l'*intertexte* du texte présenté. Un lien hypertexte est ainsi envisagé *de DECID vers LEADER* (*Livre Electronique des Actions de la DER*), qui réunit et organise l'ensemble des textes utilisés pour la construction des profils, et offre une palette d'outils de navigation sur ce corpus[46]. DECID est de fait une manière puissante de trouver des points d'entrée intéressants sur LEADER (LEADER a des accès par table des matières et par équation booléenne, mais pas d'interrogation souple par un texte), et LEADER un moyen approprié pour mieux connaître les activités d'un chercheur, une fois ce chercheur identifié.

---

[44] Il est essentiel de se conformer à sa présentation canonique, ou à défaut d'adopter une mise en page soignée et respectueuse du texte original (Merle, Fradin, Soinard 1994, p. 24).

[45] Le concepteur et réalisateur de cet affichage (Laurent LUCIANI, société Décilog) a trouvé une manière élégante de mettre en valeur les mots dans le texte : l'astuce consiste à jouer non seulement sur la couleur mais aussi sur un très léger grossissement de la police de caractères.

[46] Par exemple : les références d'une Action à une autre sont traduites par des liens hypertextes ; l'Action est présentée dans l'environnement des Actions du même Groupe ; il est possible de voir les Actions rattachées au même projet ; etc.

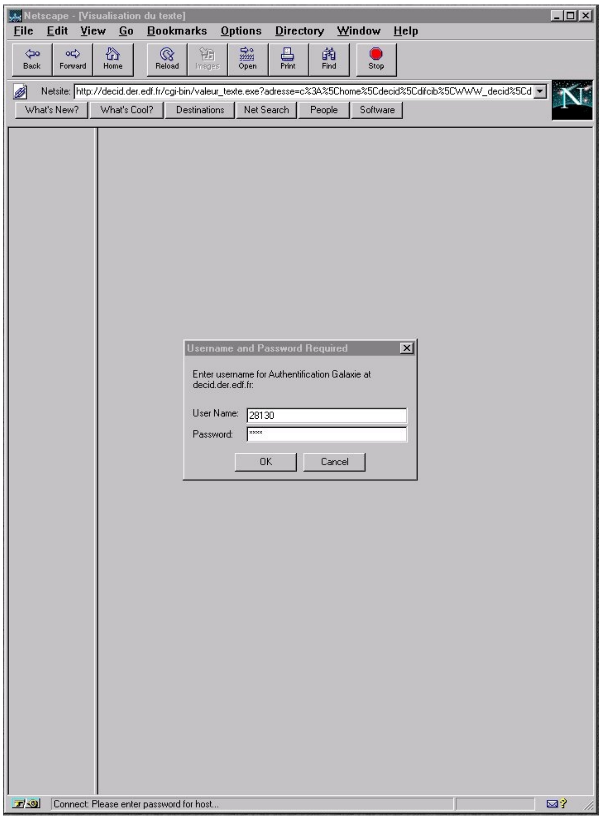

FIGURE 15 : Contrôle d'accès pour le texte intégral.

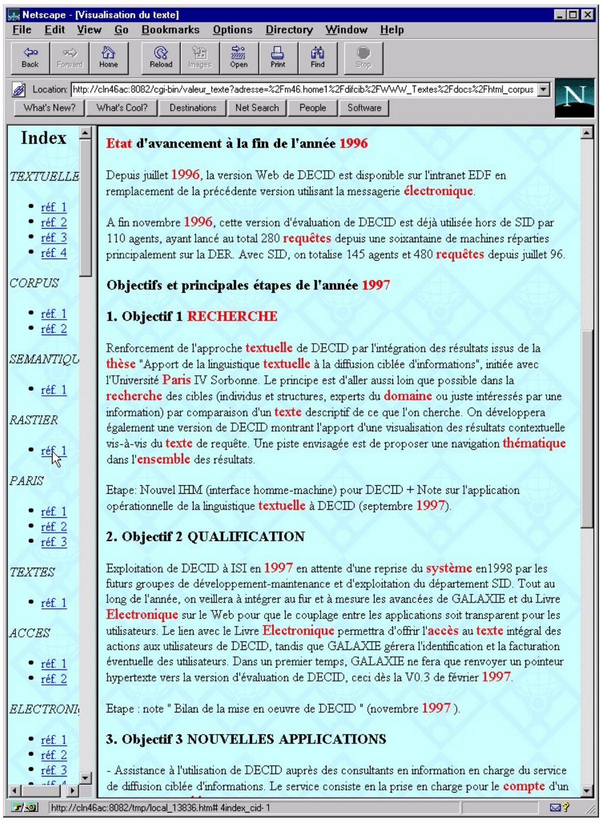

FIGURE 16 : Texte d'Action avec index contextuel.

c) *Organisation de l'ensemble des propositions*

Les systèmes documentaires habituels aident à sélectionner... mais malheureusement pas à lire. L'informatique documentaire s'arrête à la réquisition, elle saisit des documents sans en faire un corpus ; au contraire, il importe d'en donner une intelligibilité d'ensemble.

Ordonner les unes après les autres les propositions de destinataires n'est pas très naturel : telle personne retient l'attention pour tel aspect, telle autre pour une raison différente mais tout aussi valable. L'ensemble des propositions que l'on garde a sa propre logique, les noms retenus sont complémentaires. S'en tenir à un affichage linéaire des résultats, par « ordre de pertinence décroissante », ne s'avère donc pas pleinement satisfaisant.

L'organisation thématique en Pistes / Originalités, à géométrie variable

L'utilisateur de DECID *construit* sa lecture des résultats (et donc la pertinence réelle qu'ils prennent pour lui), en s'orientant dans une arborescence dépliable[47] et stable[48] qui organise l'ensemble des propositions. Le choix a été fait d'une représentation semi-graphique, plus légère et rapide pour son calcul, sa transmission, et son affichage. L'arborescence se subdivise d'abord en *pistes* ; chaque *piste* se subdivise en *originalités* ; à chaque *originalité* est associée une liste ordonnée

[47] L'arborescence dépliable a l'allure familière d'un gestionnaire de fichiers : c'est une arborescence, et descend dans le détail des branches en fonction de ses besoins.

[48] L'arborescence pistes / originalités est un référentiel à partir duquel l'utilisateur peut s'orienter et se repérer. Elle s'oppose en ceci à l'affichage proposé par Semiomap, élégant mais désorientant, qui se réorganise « élastiquement » (et quelquefois un peu brusquement) dès que l'utilisateur veut connaître le détail d'une des directions d'exploration des résultats.
Pour voir *Semiomap* (des molécules thématiques spatiales, maillées, élastiques et interactives, plongées dans une nuit étoilée) :
`http://www.semio.com/`
Un des concepteurs de *semiomap* est Claude VOGEL.

de destinataires potentiels. Chacun de ces niveaux a sa signification dans la démarche de dépouillement des résultats.

DECID propose donc d'abord plusieurs *pistes*, qui correspondent à autant de facettes détectées dans le texte. Par exemple, en ce qui concerne cette thèse, on pourrait trouver une piste « hypertexte, interface », une piste « textes, sémantique », etc. Les pistes rejoignent la pratique de l'utilisateur et se situent au niveau des domaines de travail, des différents usages auxquels peut conduire le texte de requête. L'utilisateur a ainsi une cartographie des grandes régions dessinées à partir du texte, et peut élaguer, par pans entiers, ce qui ne l'intéresse pas. La division en pistes donne d'emblée une vue d'ensemble des résultats de la recherche (sans avoir besoin de parcourir une liste, dans laquelle d'ailleurs on ne sait où s'arrêter). C'est également un excellent moyen pour ne pas être gêné par d'éventuelles erreurs d'analyse du système : si un contresens est fait sur un aspect du texte, tous les rapprochements liés sont regroupés dans une même piste, et ne viennent pas « polluer » les autres propositions.

Au sein d'une piste, on contraste des aspects particuliers, de même qu'un documentaliste recommande à la fois tel ouvrage pour sa clarté, tel autre pour son glossaire, et tel autre pour ses développements sur des questions récentes. Ici, ce sont les différentes spécialités des destinataires potentiels qui sont soulignées : les uns font le lien avec telle théorie, d'autres présentent l'intérêt particulier de connaître tel contexte d'application, etc. Pour que les résultats de la recherche présentent du relief et de l'intérêt, il faut en effet éviter de banaliser chaque présentation, et de s'en tenir à une description vague et générale. L'idée est de dire « le plus de ceux-ci, c'est... », de signaler des spécialités distinctives, sans chercher à donner une vision complète (et soit trop simplifiée, soit trop longue). Si l'*originalité* indiquée retient l'attention de l'utilisateur, alors il peut chercher à en savoir plus et compléter sa connaissance de l'activité de la personne. C'est en quelque sorte une approche différentielle qui souligne les points forts de chaque proposition et (peut-être) donne envie d'aller voir, de prendre contact. Alors que les pistes sont une manière de décliner la requête sous différentes perspectives, les originalités se rendent proches des propositions concrètes tirées de la base. Le mouvement est bien, à partir d'un document initial, de (re)découvrir la réalité dont rend compte une base.

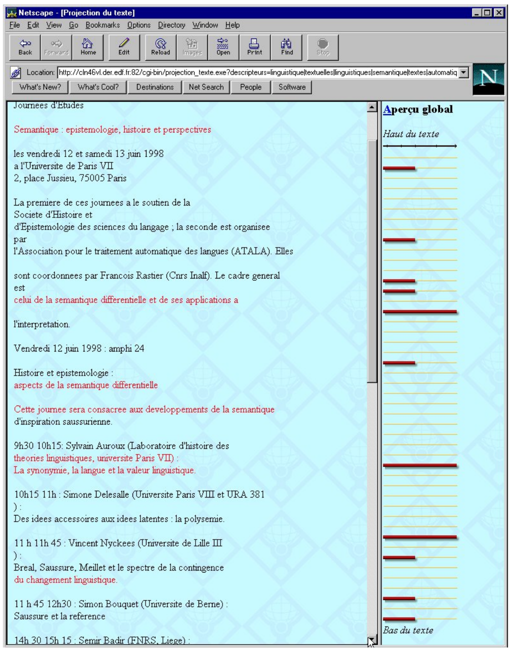

FIGURE 17 : La *projection* d'un profil sur le texte de requête ; le curseur de l'ascenseur indique la portion de l'histogramme correspondant à la partie du texte affichée dans la fenêtre.

FIGURE 18 : Arborescence *Pistes / Originalités* ; à droite, le descriptif des Pistes ; en bas, la liste des propositions avec les informations détaillées.

Ensuite, à un troisième niveau, pour les personnes qui se situent dans une même *piste* et se caractérisent par une même spécialité ou *originalité*, on peut garder une notion de proximité, pour fournir une *liste ordonnée* de sélections en commençant par celle qui est *a priori* la plus proche du texte de départ. La proximité entre les textes correspond approximativement à la fiabilité de la proposition : plus les textes s'écartent, plus la proposition est hasardeuse. S'il s'agissait d'une autre base avec un autre corpus, pour une application de recherche documentaire par exemple, d'autres critères d'ordre peuvent être envisagés[49], par exemple selon la chronologie (le plus récent est présenté en premier), ou selon un indicateur de popularité (pour un corpus de pages Web, sachant le nombre de visites de chaque page, ou le nombre de liens pointant vers elle).

Dans chaque liste associée à une piste puis une originalité, ne figurent que les noms des destinataires proposés. Pour en savoir plus sur chaque proposition, il suffit de *cliquer sur le nom* : la *liste détaillant la présentation* des personnes avec l'explication du rapprochement avec la requête est alors immédiatement *positionnée* sur la personne désignée. La vue d'ensemble donnée par l'arborescence reste donc limitée à l'essentiel pour s'orienter dans le parcours des résultats. En effet, l'information détaillée n'a besoin d'être affichée que pour la proposition qui est en train d'être examinée par l'utilisateur, elle est focalisée à son point d'attention courant. Un autre effet intéressant de cet affichage, qui demande de cliquer sur le nom pour connaître le détail du rapprochement, est que l'on garde trace des propositions déjà examinées et des propositions qui restent à parcourir. En effet, les navigateurs Web affichent conventionnellement les ancres des liens activés d'une autre couleur : ici donc, les noms des personnes pour lesquelles ont a déjà demandé le détail du rapprochement sont marqués comme « *déjà vus* ». Cela facilite l'exploration méthodique et systématique des résultats.

Pistes, *originalités* : comment étiqueter chacun de ces points de choix, pour guider l'utilisateur ? Une première indication à fournir est une *indication de volume* : combien de destinataires potentiels correspondent à chaque piste, puis à chaque originalité. Cela donne un aperçu des dominantes et des points de vue plus rares ou minoritaires, par rapport au document soumis et pour la base de destinataires considérée. D'autre part, pour caractériser chaque piste / originalité, lui sont associés les quelques *mots (ou unités descriptives) les plus représentatifs* de l'ensemble des textes sous-jacents. Une information plus riche serait de remplacer ces KWOC (*KeyWord Out of Context*, on a en effet juste une liste de mots sans autre précision, et qui ne se contextualisent qu'entre eux, par la liste qu'ils forment) par des KWIC (*KeyWord In Context*, on aurait alors une sélection d'extraits centrés sur les mots caractéristiques) ou, plus lisiblement peut-être, par des KWAC (*KeyWord And Context*, chaque mot représentatif étant suivi de quelques exemples d'emplois caractéristiques)[50].

L'arborescence est consultée dynamiquement. Au départ, seules les pistes sont présentées. L'utilisateur choisit alors de suivre celles qui l'intéressent. Pour explorer une piste, il la « *déplie* » (en cliquant sur le conventionnel symbole « + », qui se transforme alors en « - » pour indiquer la possibilité de « repli »). Sont alors détaillées les diverses originalités d'approche du thème. A nouveau, l'utilisateur fait son choix, et décide de déplier et consulter ce qui lui semble convenir. Il accède alors seulement à une liste ordonnée de noms. Un « *limiteur de liste* » est prévu pour s'en tenir à la tête de liste si celle-ci est trop longue. Grâce à tout cela, l'écran n'est pas surchargé par l'ensemble des propositions, mais n'affiche que celles qui paraissent motivées dans le cadre des besoins de l'utilisateur. La représentation n'est pas uniforme, elle ne détaille et n'approfondit que ce qui est au centre d'intérêt et d'attention de l'utilisateur, sans perdre de vue un environnement toujours présent mais moins précis (Fischler & Lahlou 1995, §4.4, p. 22).

Il n'est pas exclu qu'une même personne (destinataire potentiel), du fait de la diversité de ses activités, puisse relever de plusieurs pistes. Cette personne est alors nommée à différents endroits

---

[49] La multiplicité des critères d'ordre ou de catégorisation intéressants et complémentaires (rang ou valeur de similarité, nombre et nature des mots de la requête retrouvés, auteur, genre, année, longueur de la bibliographie, référence ou code de classement, volume,...) a conduit à la construction de représentations capables de traduire un grand nombre de dimensions (coordonnées selon deux axes dans le plan, mais aussi caractéristiques des éléments positionnés dans ce plan : taille, couleur, saturation de la couleur ou niveau de gris, forme géométrique, étiquette), cf. (Nowell & *al*. 1996).

[50] Nous empruntons les désignations KWOC, KWIC, KWAC à (Salton & MacGill 1983, §3.5.A, p. 78).

dans l'arborescence. Il est intéressant pour l'utilisateur de le savoir, et de voir rapidement où elle apparaît. Pour cela, le *nombre de fois où la personne est mentionnée* est indiqué en face de son nom, et des liens hypertextes permettent de faire le tour (*navigation cyclique*) de tous ses rattachements piste / originalité.

Il est classique aussi qu'il reste quelques destinataires inclassables, isolés dans le cadre des autres destinataires trouvés, et ne relevant pas des principales pistes qui ont pu être définies[51]. Ces *inclassables* (ou *trouvailles*, terme évocateur mais convenant mieux à la recherche documentaire qu'à la recherche de personnes) font l'objet d'une pile (liste) complémentaire, qu'ira explorer l'utilisateur désireux d'élargir sa recherche.

La liste générale ordonnée

L'arborescence piste / originalité renvoie donc à une liste complète des destinataires proposés, pour l'information détaillée sur les rapprochements. L'ordre linéaire est ergonomique dans la mesure où il aide à un parcours systématique des résultats, sans oublis et sans redondances. Il offre d'emblée une dynamique, un sens de lecture.

L'ordre adopté est classiquement l'ordre suivant la *pertinence décroissante*. La pertinence est ici comprise au sens du score numérique donné par la mesure de similarité (proximité) entre le profil et le document soumis. C'est d'ailleurs ce type d'affichage qui existe depuis l'origine dans DECID (en exploitation actuellement, il est sous la forme présentée FIGURE 13, FIGURE 14). Il est toujours présent (en bas de la FIGURE 18), mais maintenant l'arborescence piste / originalité se propose pour médiatiser le parcours et la composition de la liste. En effet, l'ordre de pertinence force un ordre qui n'a pas toujours lieu d'être, et hiérarchise des propositions qui ne sont pas directement comparables. Les différentes thématiques y apparaissent mêlées et par paquets épars, sans grande logique[52]. La lecture d'une liste de propositions par score de pertinence décroissante ne donne jamais le sentiment de faire le tour des résultats, à moins de la parcourir en entier[53], car un motif de rapprochement différent des précédents peut avoir été doté d'un score de pertinence relativement faible, et les propositions correspondantes se retrouvent perdues, en queue de liste, parmi les propositions médiocres selon les motifs principaux.

L'arborescence piste / originalité aide au parcours thématique de la liste par ordre de pertinence. Elle en donne une vue personnalisée et filtrée, si l'utilisateur choisit d'écarter certaines pistes qui ne l'intéressent pas.

L'ordre de pertinence, qui n'est en définitive qu'un ordre heuristique et imparfait selon un certain score numérique, peut se voir préférer, pour certains usages, un autre ordre des résultats, complémentaire. Pour la diffusion ciblée, on peut penser à :

- une *liste alphabétique, par nom*, des destinataires trouvés : l'utilisateur peut tout de suite vérifier la présence de certaines personnes auxquelles il pense, et pour lesquelles il souhaite que le système l'aide à évaluer leur intérêt possible pour le document ;
- une liste selon l'*organisation hiérarchique* en Services, Départements, Groupes : elle permet par exemple d'organiser la diffusion en veillant particulièrement à une bonne répartition de l'information entre les différentes équipes.

Dans tous les cas (ordre de pertinence, liste alphabétique, organisation selon l'organigramme), la liste linéaire gagne à être préalablement parcourue, recomposée et sélectionnée thématiquement, à l'aide de la vue générale en pistes et originalités.

---

[51] L'organisation en pistes correspond à notre mode de pensée et de repérage : l'homme éprouve le besoin de classer, de catégoriser ; mais l'expérience montre aussi que le résultat est toujours insatisfaisant, comme en témoigne l'inévitable pile divers ou à classer (Fischler & Lahlou 1995, §4.5, pp. 23-24).

[52] C'est ainsi qu'une présentation par *quorum-level* (Salton 1988), telle que celle réalisée par SPIRIT, est moins fine mais ergonomiquement plus efficace qu'une présentation par ordre de pertinence, telle celle des moteurs de recherche sur Internet. Dans la présentation par quorum-level, les résultats sont groupés en fonction des mots de la requête présents dans les documents.

[53] Ce qui est rarement possible pour les recherches sur les bases volumineuses ; la solution du seuillage des résultats est généralement insatisfaisante, car brutale et toujours susceptible d'écarter quelques propositions intéressantes, sur des aspects (mots-clés) que le calcul a moins valorisés que les autres.

Les possibilités de parcours et d'organisation des résultats restent ouvertes, grâce à l'export des résultats (FIGURE 20), fonctionnalité présentée ci-après.

6. Exploitation de la sélection

a) Retour sur la requête et affinement itératif

L'utilisateur peut rapidement s'apercevoir qu'un aspect qu'il jugeait important a été insuffisamment pris en compte, ou que le document qu'il a soumis est composite et gagnerait a être analysé partie par partie.

DECID lui permet de revenir sur son texte de requête, tel qu'il a été soumis (FIGURE 10), et de le modifier (sauf s'il s'agissait d'un fichier), aussi souplement qu'il a pu composer sa requête[54] : ajout de quelques lignes au clavier, ajout d'un complément technique par copier / coller, effacement d'une partie formelle peu liée à la problématique concernée, mise en relief par surlignage d'un aspect important pour cette recherche, etc.

Il peut bien sûr aussi effacer instantanément le texte soumis pour repartir sur une nouvelle expression de la requête à partir d'un autre texte.

b) Export : récupération de la liste des propositions retenues

Les propositions calculées par DECID couvrent un large éventail de destinataires, et c'est une partie d'entre eux seulement que choisit de retenir l'utilisateur : de fait, c'est bien lui, l'utilisateur, et non la machine, qui peut savoir ce qui a une pertinence pour lui, en fonction de ses connaissances antérieures, des circonstances de sa recherche, de l'usage escompté de la liste des destinataires.

Au fur et à mesure de son exploration des propositions retournées par le système, l'utilisateur *coche*, sur une liste générale, les destinataires qu'il veut retenir (FIGURE 21). Il est prévu d'ajouter la possibilité d'entrer, en face de chaque nom, un *commentaire bref* : l'utilisateur peut ainsi ajouter ses propres repères, catégoriser et classer les résultats à sa façon, noter une observation. On pourrait encore ajouter la métaphore[55] d'une *étagère*, sur laquelle ranger les propositions retenues. Cette étagère comporterait par exemple par défaut trois rayons, avec les étiquettes « à traiter (urgent) », « à retenir (important) », « à voir (intéressant) ». L'étagère serait aménageable selon les souhait de l'utilisateur, qui pourrait mettre le nombre de rayons qui l'arrange et étiqueter ces rayons à sa guise. Il disposerait ainsi de deux modes complémentaires d'enregistrement d'informations au fur et à mesure du dépouillement des résultats : une première organisation en différentes catégories, et des annotations mémorisant des particularités significatives. Toutes ces informations prépareraient une exploitation ultérieure des propositions finalement retenues. Elles contribuent à un équilibre nécessaire entre repérages standard et repérages personnels : l'information tirée du parcours des résultats est réorganisée et transformée pour être transcrite de façon opérationnelle (Fischler & Lahlou 1995, §5.2.1, p. 34).

---

[54] La modification après examen des résultats pourrait même être rendue plus souple que la composition de la requête initiale, ceci afin de pousser l'utilisateur à prendre d'abord connaissance des résultats avant de remanier le texte soumis, car en le reformulant selon ses idées il s'écarte peut-être d'un traitement optimal. En effet, les résultats sont potentiellement meilleurs que ce qu'il croit possible avec un traitement contraire à ses intuitions (texte au lieu de mots-clés, analyse apparemment fruste).
Cette tactique avait été adoptée pour l'application de gestion de profil : elle « obligeait » à d'abord examiner comment se comporte le profil directement calculé à partir d'un texte, avant de permettre de le retoucher.

[55] A propos de métaphore dans les systèmes d'information et de documentation : l'équipe travaillant sur la bibliothèque électronique à EDF avait imaginé de représenter les profils de la diffusion ciblée (ou ceux d'une DSI) comme un *aimant*, attirant sélectivement les documents passant dans son voisinage... pourquoi pas ?

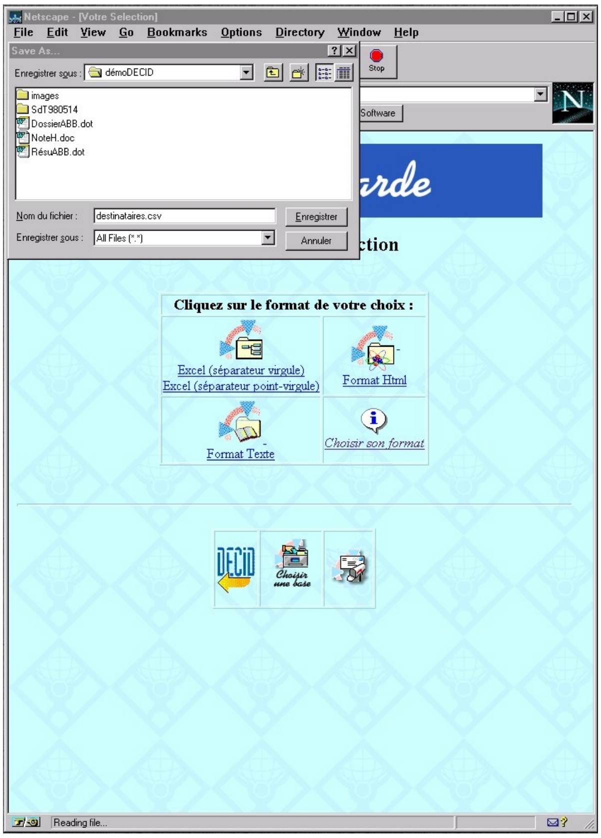

FIGURE 19 : Sauvegarde de la liste des destinataires retenus par l'utilisateur parmi ceux proposés.

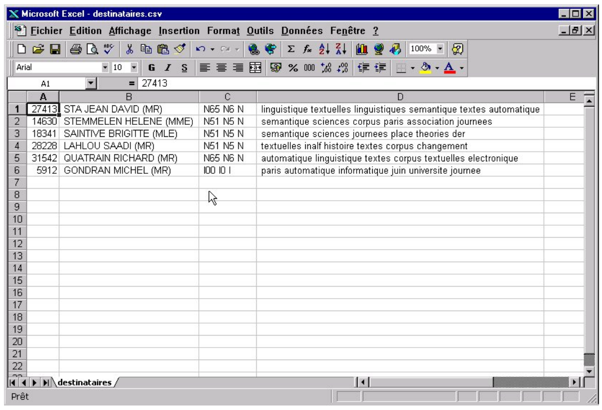

FIGURE 20 : Le fichier créé par DECID peut être ouvert dans un tableur standard et faire l'objet de tris et de retraitements.

Il faut en effet prévoir un mode de sauvegarde des résultats d'une recherche. Une fois enregistrée dans un fichier qui appartient à l'utilisateur, la liste des personnes trouvées par rapport au document peut être réexaminée ultérieurement (en dehors de la session d'interrogation, pour laquelle on ne dispose pas toujours du temps adéquat). La sélection sauvegardée ouvre de nouvelles possibilités : consultation plus approfondie (impression dans un format choisi, multiplication des possibilités de vues et de tris), et utilisation, intégration à ses propres documents et données, par exemple comme base pour la constitution et la mise à jour d'une liste de diffusion de référence.

Trois formats de sauvegarde sont proposés (FIGURE 19), qui reflètent bien la diversité des modes d'exploitation de la sélection retenue :

- le format *texte simple*, qui présente l'avantage d'être un standard simple, non dédié à un logiciel ou à une application spécifique, et aisément visualisable dans tout éditeur de texte. Il est en revanche peu structuré : chaque ligne commence par le nom d'une personne, suivi des différents renseignements sur le rapprochement avec le document de requête. Ce fichier généré automatiquement suit une présentation régulière, et peut être parsé pour faire l'objet d'une exploitation particulière, en devenant l'entrée d'un nouveau traitement automatique.
- le format *CSV*, qui écrit les résultats sous une forme tabulée telle qu'elle est directement exploitable par un *tableur* (par exemple *Excel*). Bien entendu, chaque information associée à un rapprochement est l'objet d'une colonne : une colonne pour le nom de la personne, une colonne pour son code de rattachement, une colonne pour les mots-clés explicatifs du rapprochement, une colonne pour l'annotation éventuelle de l'utilisateur, etc. La liste des destinataire bénéficie alors de la souplesse de manipulation du tableur familier à l'utilisateur, et peut être triée, filtrée, etc., sur tous les critères souhaités (FIGURE 20).
- le format *HTML*, pour affichage dans n'importe quel navigateur. Il fournit une présentation soignée, appropriée à la lecture et à la consultation.

Une page d'explication est proposée à l'utilisateur qui connaîtrait mal ces formats et a besoin de savoir les usages de chacun.

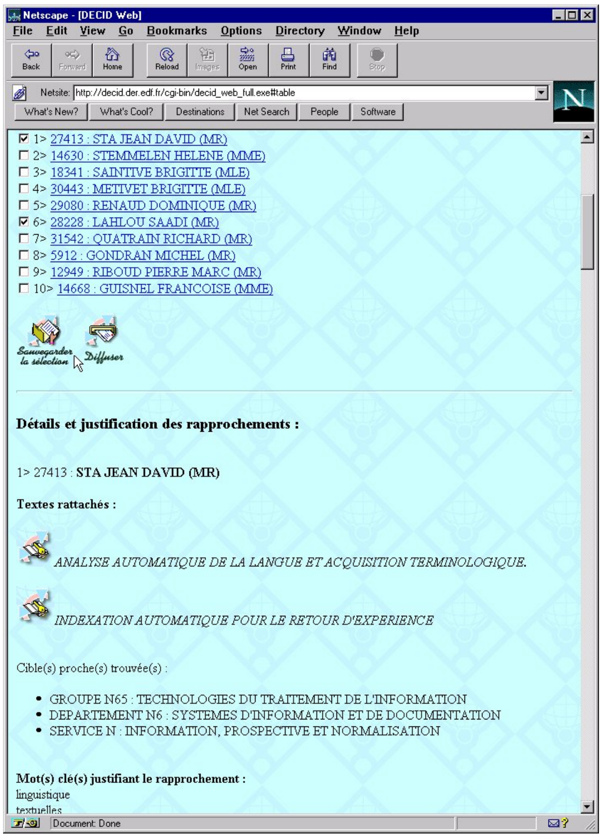

FIGURE 21 : L'utilisateur sélectionne les noms qui l'intéressent parmi les propositions de DECID.

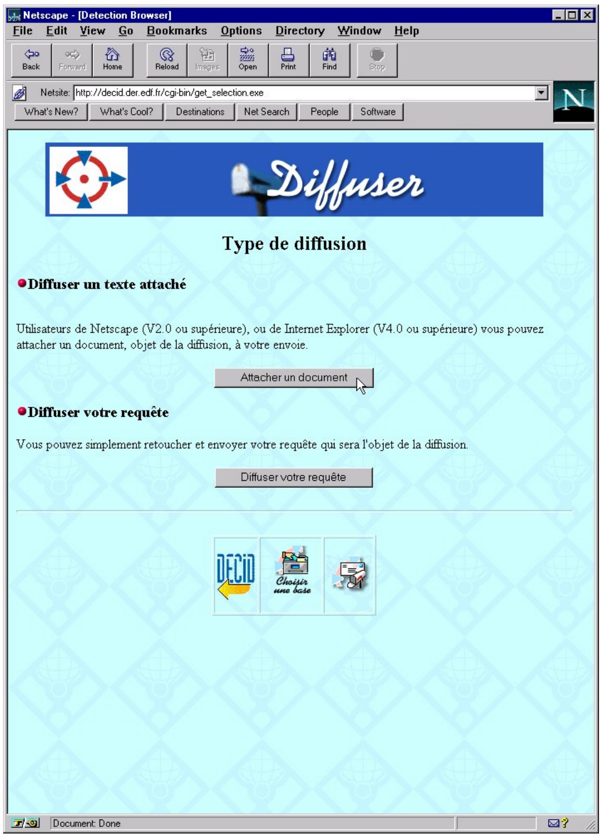

FIGURE 22 : Utilisation de DECID pour une diffusion effective, celle du texte soumis au système ou du document original correspondant.

En fonction des formats de sauvegarde, DECID rappelle dans le fichier les différents paramètres généraux de la recherche. L'utilisateur garde ainsi *trace du contexte d'obtention* de la liste : la *base de profils* dont sont tirés les destinataires, un rappel du document soumis (par exemple ses *premiers mots*, mais on pourrait aussi permettre à l'utilisateur d'indiquer la *désignation* qui lui convienne), et l'on pourrait ajouter *la date et l'heure* de la recherche.

c) Diffusion du document

Un des objectifs de soumettre un document au système de diffusion ciblée peut être... de diffuser effectivement le document. Or, une fois une liste de destinataire établie, il reste encore toute une série de tâches, assez lourdes et rébarbatives. Le rôle de DECID est d'alléger, autant qu'il est possible, cette charge de travail. A la machine revient d'effectuer ce qui est répétitif et déterministe, et d'aider l'utilisateur pour les gestes qu'il doit raisonnablement assumer, et ceux qu'il peut préférer réaliser lui-même.

La diffusion opérée par DECID est une *diffusion électronique, par la messagerie interne*. Elle suppose donc soit que l'on dispose du document à diffuser sous forme électronique, soit que l'on en diffuse un signalement, en indiquant l'endroit où le document est disponible.

Ce que la machine est capable de traiter automatiquement, c'est par exemple la *recherche des coordonnées électroniques des destinataires* et le remplissage du champ correspondant. L'utilisateur a toujours la possibilité de modifier une adresse, si besoin.

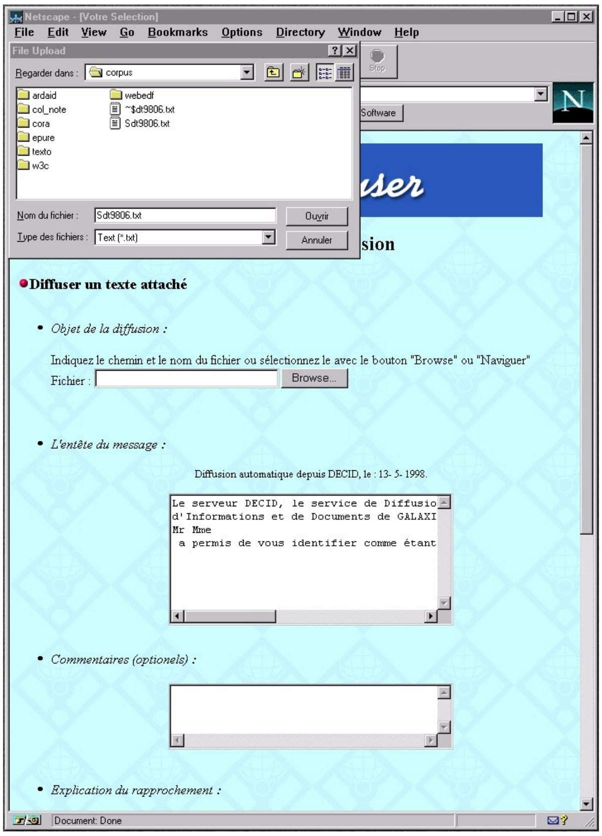

FIGURE 23 : Préparation du message électronique de diffusion.

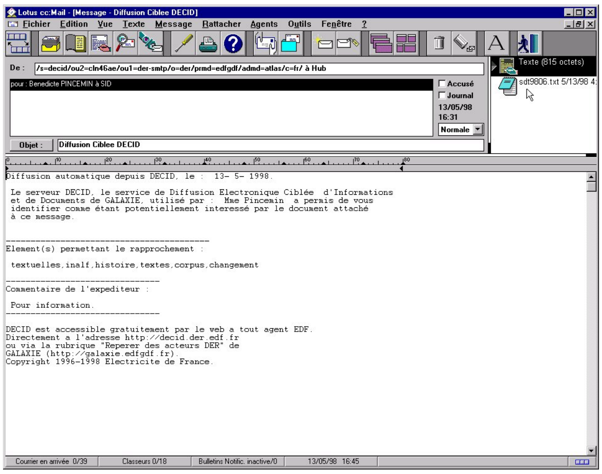

FIGURE 24 : Le document arrive dans la messagerie du destinataire ; il est en pièce jointe à la lettre d'accompagnement générée par DECID.

DECID prépare également une *lettre d'accompagnement* (FIGURE 23). Il épargne ainsi à l'utilisateur la rédaction d'un message et sa frappe. L'ensemble des indications nécessaires au destinataire sont méthodiquement inclues dans la lettre, pour qu'il puisse efficacement tirer parti de l'envoi[56]. Sont prévus : le nom de l'expéditeur, le fait que la suggestion d'envoi ait été obtenue par un traitement automatique (l'expéditeur est en partie déchargé de la responsabilité d'un envoi finalement jugé non pertinent par le destinataire), le principe de fonctionnement de DECID et sur quelle base le profil de la personne a été déterminé, la donnée des explications des points communs entre le document transmis et le profil du destinataire. L'utilisateur (expéditeur) peut bien sûr ajouter ses propres remarques ou un mot personnel, dans une zone prévue à cet effet.

Ce qui est effectivement envoyé est, au choix de l'utilisateur (FIGURE 22), *le texte qui a servi comme requête* pour interroger DECID et chercher les destinataires, *ou bien un fichier*, expédié sous forme de fichier attaché au message contenant la lettre. Le fichier est indiqué par navigation dans les fichiers accessibles sur le poste de l'utilisateur. Cette distinction est utile dans le cas où la recherche s'est effectuée sur une synthèse ou un extrait représentatif du document.

Une fois tout ceci préparé, l'utilisateur lance la réalisation de l'envoi. DECID lui confirme, par une liste récapitulative des noms et adresses, l'expédition d'un courrier électronique pour chacun des destinataires prévus. La FIGURE 24 montre le courrier tel qu'il est reçu par un destinataire.

[56] Cette explication et contextualisation de l'envoi est capitale, du moins c'est ce qui apparaît dans les études du syndrome de saturation cognitive : la nécessité pour le destinataire d'analyser chaque information reçue et d'identifier ce qu'il a à en faire est, avant même tout travail d'utilisation de l'information, un facteur important du sentiment de surcharge.

7. Récapitulatifs des fonctionnalités

a) Bilan selon l'état d'avancement et d'intégration à l'application

Toutes les fonctionnalités présentées ont été validées au plan de la faisabilité technique, au sens où elles ont toutes été implémentées.

Une fonctionnalité réalisée en tant que maquette signifie qu'un module existe pour l'effectuer. Ce module peut faire l'objet de démonstrations, en faisant abstraction du fait qu'il n'est pas encore paramétré pour travailler comme prévu sur des données réelles du traitement.

Lorsque la fonctionnalité appartient au prototype, elle acquière un rôle réel, et entre en jeu dans une véritable session d'interrogation de DECID. C'est une anticipation sur le futur (proche) de l'application.

Il n'y a en définitive pas de différence très sensible, pour une fonctionnalité, entre le fait de relever du prototype ou bien d'être dans l'application en exploitation, car l'évolution de l'application s'est faite progressivement. Les fonctionnalités les plus innovantes ont été intégrées récemment, si bien qu'elles ne bénéficient pas encore de retours utilisateurs significatifs. La seule différence à souligner, c'est que les fonctionnalités en exploitation ont démontré leur fiabilité (DECID est en service de façon ininterrompue, cf. annexe II.1) et sont entrées sans heurt dans la pratique des utilisateurs.[57]

Le tableau ci-après reprend les étapes qui viennent d'être présentées, avec les numéros des paragraphes correspondants.

---

[57] Il est bien sûr déjà possible de faire une première évaluation par rapport à une grille de critères systématiques. Par exemple, une méthodologie d'évaluation très complète de l'interface homme-machine de logiciels multimédia, dans le cadre des logiciels pédagogiques, a été mise au point à l'Université de Technologie de Compiègne (travaux de P. TRIGANO, O. HÛ et S. CROZAT) ; l'interface de DECID répond de manière satisfaisante aux critères de qualité proposés.
Voir par exemple le document de travail suivant (extrait d'un polycopié de cours) :
`http://www.hds.utc.fr/~ptrigano/lo18_index.html`
DECID est également conçu dans l'esprit des trois principes d'interaction indiqués par (Denos 1997, §II.1.3.1, pp. 53-55, §III.3, pp. 91-93, & §V.4, p. 231) :
1. *Détecter* : « facilité à détecter la pertinence (par exemple, 'ce document est pertinent seulement si le corpus ne comprend pas d'autres documents plus pertinents' ou encore 'je ne peux évaluer la pertinence d'un si grand nombre de documents') ». Ce principe est pris en considération par l'organisation et la lisibilité des résultats, les informations et explications fournies, la visualisation de l'adéquation de la base des profils à la requête (la 'prise en charge').
2. *Comprendre* : « clarifier la sémantique de l'interaction », « accord sur le sens du langage de requête (par exemple, 'le terme que j'emploie a-t-il bien le même sens pour le système') ». Ce principe est pris en considération en rendant compte du comportement du système, de ce sur quoi il s'appuie (mots les plus actifs dans les rapprochements), pour percevoir éventuellement là où intervenir pour approfondir ou réorienter la recherche. En outre, la requête textuelle, *(i)* en se plaçant directement au niveau du texte intégral et non au niveau d'un méta langage plus générique (comme les mots-clés usuels), *(ii)* sans dépendre de choix de terminologie et d'usage comme dans le cas d'une indexation contrôlée, et *(iii)* robuste grâce à sa contextualisation, épargne un certain nombre de difficultés liées à la formulation de la requête.
3. *Reformuler* : « mise en évidence d'éléments pour améliorer la formulation du problème d'information ('comment modifier la formulation de mon problème d'information alors que la réponse ne me fournit aucun indice pour le faire') ». Ce principe est pris en considération en aidant l'utilisateur à identifier les facteurs perturbateurs et correctifs, lorsqu'il perçoit un décalage entre ses attentes et les résultats obtenus. Là encore, tout ce qui permet de repérer les parties actives ou au contraire inertes dans les rapprochements va dans ce sens : par exemple, prise en charge et projections (avec histogramme), affichage du texte intégral (avec index contextuel).

	DECID en exploitation	**Prototype intégré (version en test)**	**Maquette (modules)**	**Projet, idées.**
2. Accès				
2.a. Equipement	- navigateur Web		- plug-in Tk/Tcl (option)	
2.b. Aide à l'utilisateur	- documentation en ligne - contact par formulaire			
2.c. Contrôle d'accès	- Intranet, identification, authentification.			
2.d. Administration	- interface Web - gestion de configuration			
3. Constitution de la requête				
3.a. Sélecteurs				
- base	- bases annuelles			- interrogation rétrospective pluriannuelle
- type de destinataires	- profils : agent, Groupe, Département, Service			
- seuillage du volume de réponses	- nombre maximum de réponses			
- paramètres textuels				- langue, genre
3.b. Introduction d'un texte	- clavier, copier / coller, fichier - césure automatique - non troncature des textes longs - formats : texte, HTML			- format Word
3.c. Actions sur le texte d e requête				
- mise à blanc	- effacement requête		- effacement surlignage	
- surlignages			- horizontal, vertical.	
4. Informations sur le traitement				
4.a. La forme de la requête		- message d'avertissement si moins de 5 mots		
4.b. La prise en charge : adéquation de la base à la requête		- codes de couleur : inconnu, fort potentiel de contribution - histogramme marginal		
4.c. Le temps de traitement	- estimation indicative			- (éventuellement, graphique de suivi)
5. Communication des résultats				
5.a. Informations de présentation de chaque personne	- nom et prénom - M. / Mme / Melle - matricule - rattachement (code			- textes de présentation autres que les Actions

	et libellé) - titres des textes d'Action			
5.b. Explication : motifs d'un rapprochement profil-requête	- 5 mots ayant le plus contribué au rapprochement - l'accès au texte intégral des Actions, guidé par le texte de requête (index contextuel)	- la projection : une lecture de la requête selon le point de vue du profil (avec histogramme marginal)		- accès à l'intertexte : lien de DECID vers LEADER (le *Livre Electronique des Actions de la DER*)
5.c. Organisation de l'ensemble des propositions				
- liste générale ordonnée	- score de pertinence décroissant - par rattachement, par nom : grâce à l'export au format CSV			- filtrage par piste / originalité
- organisation thématique en pistes / originalités			- dépliage - lien avec liste et présentation détaillée - trace des déjà vus - étiquetage des pistes - nombre de mentions d'une personne et navigation cyclique - inclassables (trouvailles)	- limiteur de liste - indication de volume de chaque piste ou originalité - étiquetage des originalités
6. Exploitation de la sélection				
6.a. Retour sur la requête et affinement itératif	- (retour sur la page de requête par la fonction *backtrack* du navigateur)			
6.b. Export : récupération dans un fichier de la liste des propositions retenues	- cochage des noms - formats : texte simple, CSV (tableurs), HTML. - rappel de la requête par ses premiers mots			- annotation possible en face de chaque nom - étagère de rangement - autres rappels de paramètres : base, date et heure ; possibilité de donner un nom à la requête
6.c. Diffusion (électronique, via messagerie interne)	- recherche des coordonnées électroniques - préparation d'une lettre d'accompagnement - envoi : texte de requête ou fichier attaché - compte-rendu d'envoi			

b) Le point sur les fonctionnalités plus spécialement textuelles, et issues du travail de thèse

Avec les nouveaux moyens d'affichage graphique et la puissance de calcul maintenant disponible sur les machines, les interfaces élaborées et attrayantes ne manquent pas[58], pourtant peu de propositions concernent l'affichage des textes (potentiellement longs), y compris dans le cadre des systèmes de recherche documentaire. Dans ce contexte, les innovations apportées par la thèse en matière d'ergonomie textuelle sont[59] :

- la définition de deux modes de surlignage d'un texte, le *surlignage horizontal* et le *surlignage vertical*, comme expressions de deux formes de mise en relief complémentaire, l'une attachée à une formulation précise ou à une mention, l'autre intéressée par l'idée sous-jacente au passage sélectionné. Ces deux modes sont réalisés de façon intuitive, par des clics souris multiples. (Voir § 3.c. ci-dessus).
- la notion de *prise en charge* d'un texte de requête par rapport à une base (§ 4.b.), qui peut plus généralement s'interpréter comme la confrontation d'un texte à un corpus dans son ensemble. Le texte se trouve alors décrit à travers trois valeurs : ce qui échappe au champ du corpus, ce qui est dans le champ du corpus et saillant, ce qui prend une valeur neutre et effacée (le fond sur lequel se détache les formes précédentes).
- l'*histogramme marginal* qui, associé à l'ascenseur, permet d'allier une vue globale du texte (long) à la vue locale qui s'inscrit dans la fenêtre d'affichage. Les pics et les vallées de l'histogramme permettent de repérer immédiatement les passages remarquables du texte (par rapport à une mesure donnée) et leur étendue. Un simple clic souris positionne le texte visualisé au point d'intérêt choisi sur l'histogramme. (L'histogramme marginal est présenté au § 4.b.)
- la *projection* d'un texte sur un autre (présentée dans le cadre du § 5.b.), qui donne une lecture du second texte selon le point de vue du premier. Il y a bien là une portée herméneutique, car cette fonctionnalité donne à voir un texte selon un éclairage significatif et parfois révélateur. Par ailleurs, la projection est un mode de caractérisation des affinités entre deux textes supérieur à la donnée des mots en communs, puisqu'il est pleinement contextuel. Notons enfin que l'affichage d'une projection se munit avantageusement d'un histogramme marginal.
- l'*arborescence pistes / originalités* (§ 5.c.), permettant une navigation thématique, et une focalisation interactive dans l'exploration d'un ensemble de textes (dans DECID, d'un ensemble de propositions de destinataires). Cet affichage n'exclut pas une présentation par liste ordonnée : il médiatise l'accès à la liste. Ce faisant, sont résolus la plupart des défauts inhérents à ces listes linéaires quand elles sont le seul mode de représentation de la pertinence, défauts théoriques (une pertinence préétablie et monodimensionnelle) et pratiques (dépouillement long et austère de propositions mêlées). Le but de l'arborescence pistes / originalités est une construction dynamique de la pertinence (l'utilisateur choisit sa manière de parcourir les résultats) et un examen efficace des propositions du système.

Une autre interface très intéressante pour la navigation dans un texte est l'*index contextuel* (associé dans DECID à l'accès au texte intégral des Actions, § 5.b). Sa conception et sa mise au point reviennent en premier lieu à Laurent LUCIANI (société Décilog).

La présentation intégrale de l'interface de DECID a également montré l'incidence de la prise en compte de la textualité dans la réalisation d'autres fonctionnalités, principalement : l'importance de la césure automatique des lignes ; l'attention portée à la non troncature des textes longs.

Les qualités recherchées seraient finalement d'obtenir des vues : synoptiques (voir tout), synthétiques (grandes lignes, dominantes), caractérisantes (spécificités, originalités, saillances),

[58] Voir l'étude très riche de (Cokburn & Jones 1997), qui ne manque pas de recenser HotSauce, WebCore, DeckScape, Webmap, Mitre, Hyperbolic Space, WebViz, HyperSpace...

[59] L'apport de la thèse réside dans la *conception* de ces interfaces ; elles ont été implémentées par Laurent LUCIANI (société Décilog), à qui l'on doit aussi quelques enrichissements astucieux des propositions initiales. Essentiellement : la possibilité de cliquer sur les barres de l'histogramme ; l'utilisation des double et triple clics souris pour le surlignage (plutôt que de faire glisser le curseur).

systématiques (avoir une règle de parcours), mnémoniques (par regroupements et sélections d'unités, par enchaînement et divisions organisant les rapports), contextuelles (sensibles aux variations mais fidèles, stables relativement à un contexte).