INSA de LYON – Doc'INSA

Serveur de thèses en texte intégral

Rapport d'étude préalable

Rédaction :	Marc-Etienne HUNEAU
Version :	2.45
Révision :	vendredi 3 avril 1998

1 Table des matières

2 Table des illustrations

3 Contexte de l'étude

L'INSA de LYON et sa bibliothèque scientifique et technique Doc'INSA souhaitent mettre en place un service d'accès à des thèses numérisées.

L'accès à ce service se fera via internet, sous la forme d'une interface « web ».

Doc'INSA est dépositaire de toutes les thèses produites dans les laboratoires de l'INSA. Ceci représente actuellement environ 130 thèses par an. Ces thèses devront être « numérisées », ou plus précisément converties en documents électroniques propres à être publiés sur internet.

La rétro-conversion des thèses reçues antérieurement à 1997 par Doc'INSA n'est pas envisagée.

Le projet recouvre deux objectifs principaux : la conception du service mis à disposition des futurs utilisateurs (ou FrontOffice), et l'élaboration de la chaîne de production des documents numériques.

Le projet a pourtant des implications plus larges au niveau organisationnel : c'est une procédure complète, débutant par l'information des auteurs et finissant par la publication des documents, qu'il faut élaborer. Ceci comprend des aspects juridiques incontournables (élaboration d'une décharge à faire remplir par l'auteur), des aspects techniques (création de modèles, transformation des documents), de nouvelles habitudes à faire passer par le biais de sensibilisations auprès des laboratoires et de la DED.

Comme on peut le voir, ce projet a des ambitions plus larges que la mise en place d'un service informatisé. Cependant, l'informatique ayant des implications à presque tous les niveaux de ce futur itinéraire des thèses, ces aspects ne sortent pas du cadre de cette étude.

4 Expression des besoins

Cette partie décrit de manière générale les besoins motivant la réalisation d'une chaîne d'édition électronique.

Avant de présenter les besoins du service, il faut rappeler le cadre du dit service : les thèses électroniques seront présentes dans la base de données de Doc'INSA au même titre que tout autre ouvrage. L'indexation et le catalogage de ces documents électroniques seront donc les mêmes que ceux de leurs équivalents papier.

4.1 Production des documents *(BackOffice)*

4.1.1 Format source

Les thèses sont fournies par leurs auteurs sous plusieurs formats. En effet, suivant l'application utilisée pour l'élaboration du document, les format reçus se partagent généralement entre Microsoft Word, Tex/LaTex et PostScript.

De plus, certaines thèses peuvent comporter en sus des photographies, illustrations et autres documents papier, non intégrés au document électronique. Ces documents doivent évidemment être intégrés à la forme numérique de la thèse avant sa publication électronique.

Enfin, il est envisageable d'intégrer à cette thèse électronique tout type de données multimédia, reliées ou intégrées au document électronique.

4.1.2 Conversion

Ces différents éléments devront être traduits en un document propre à être publié via le serveur WWW. Etant donné le volume important de thèses à transformer de la sorte, cette procédure doit être automatisée au maximum.

Il faudra cependant garder à l'esprit que le choix des formats évoluera probablement avec les progrès de la documentique et des technologies de l'internet ; par conséquent cette chaîne de production doit rester souple et adaptable.

4.1.3 Format cible

Le choix du format Acrobat (de Adobe), dérivé de PostScript adapté à l'édition électronique de documents hypermédia, semble s'imposer de lui-même au vu des autres expériences ou services existant. Ce format est adapté aux outils web (disponibilité d'un *Plug-ins* pour les navigateurs courants, logiciel de visualisation disponible sur de très nombreuses plates-formes), permet la recherche en texte intégral, et permet à l'auteur de protéger son document. *Ces deux derniers points seront précisés dans l'étude détaillée.*

Il est également envisageable, à plus long terme, de publier des documents tels que des thèses dans un format dérivé de HTML, par exemple XML. XML est apparenté à SGML, mais semble plus simple d'utilisation.

Si un changement de format est envisagé, ou même si le format Acrobat (propriétaire) connaissait des avatars, il importe de conserver les thèses dans un format « source », conservant la structure logique du document et propre à l'archivage.

4.1.4 Archivage

Afin de conserver les thèses, et dans le but d'une éventuelle évolution technique du service, il convient de choisir un (ou plusieurs) format(s) de stockage (Word, SGML) afin de regrouper les documents fournis par les auteurs. Ces documents (et la numérisation des éventuels ajouts imprimés, voir ci-avant) seront archivés (par exemple gravés sur disque optique). Cet archivage devra être soumis à quelques règles afin d'être cohérent et facilement réutilisable.

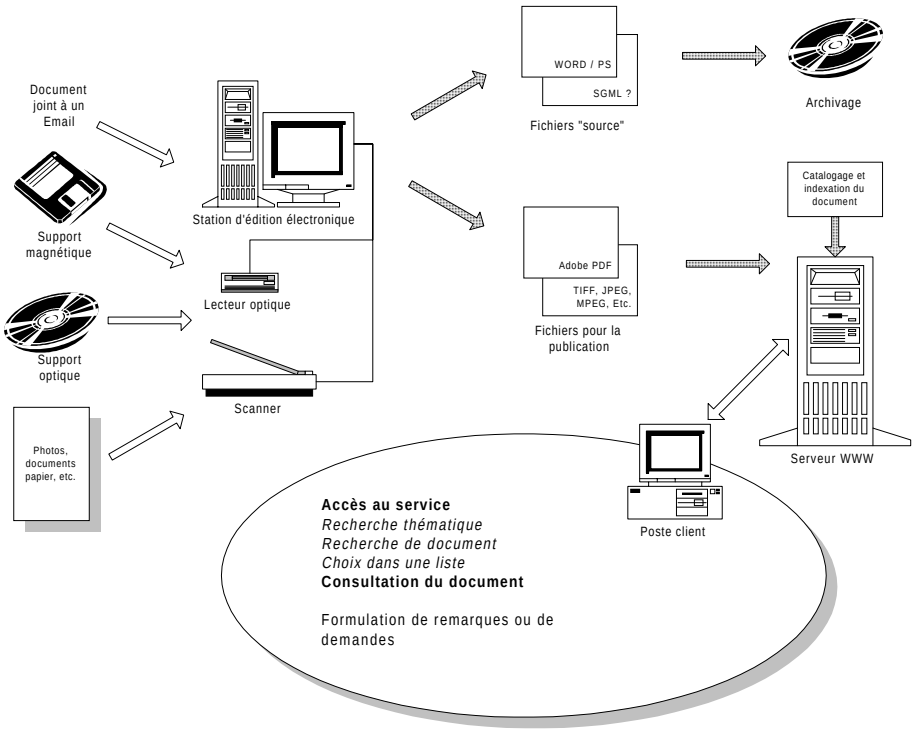
4.1.5 Fonctionnement général

Pour résumer l'itinéraire d'une thèse reçue à Doc'INSA en vue de sa publication électronique :

1. La thèse est fournie par son auteur, le support et la nature des documents pouvant varier.

2. Une « chaîne de production » permet de générer le document publiable, mais aussi un document de référence qui sera archivé.

3. Le document publiable est ajouté à la base de thèses et devient consultable.



• Figure 1 : Fonctionnement général du service, du document d'origine à sa consultation

4.1.6 Documentation technique

La chaîne de production sera documentée dans plusieurs buts :

Une documentation technique constituera une aide à la maintenance et à d'éventuelles évolutions du système.

Une documentation utilisateur et un support d'auto formation permettront une prise en main rapide et une utilisation efficace des outils de publication électronique.

4.2 Consultation des documents *(FrontOffice)*

Les documents électroniques seront publiés via un serveur. Ce serveur sera accessible par internet, l'interface en sera une interface web.

Cette interface reprendra les aspects rencontrés dans les autres services du même type (cf. : Etat de l'art). Typiquement, l'utilisateur a à sa disposition une recherche par mots clés, mot du titre, auteur ; un accès à la liste des documents (classés suivant différents critères, au choix de l'utilisateur), et la consultation proprement dite.

4.2.1 L'accès au service

Les modalités d'accès au service sont encore à définir. Si la quasi totalité des thèses sera accessible librement, il est possible que certaines d'entre elles soient soumises à des restrictions d'accès. En effet, des clauses de confidentialité ou des accords en vue d'une publication peuvent impliquer de limiter l'accès à certaines thèses.

Dans ce cas, le serveur web propose les outils de contrôle d'accès, mais ceux-ci devront être intégrés au système (interface d'administration des documents).

4.2.2 Documentation du service

Il importe de documenter le mieux possible le service. En effet, l'accès aux thèses doit être ouvert au plus grand nombre. De ce fait, l'interface doit être claire et intuitive, et comporter une aide efficace.

Par ailleurs, le service comprendra obligatoirement un certain nombre de notices légales, mises en garde ou décharges.

La présence d'une éventuelle documentation à destination des futurs auteurs est développée plus loin (4.2.5).

4.2.3 Aspect de l'interface

Comme il a été spécifié ci-avant, l'interface devra être intuitive, et proposera les outils habituels d'accès à l'information (listes, recherches).

Plus accessoirement, cette interface devra permettre à l'utilisateur de se procurer les outils lui manquant éventuellement, via les sites internet des éditeurs concernés (par exemple le plug-ins' Acrobat Reader).

Le succès d'un service internet tient non seulement à l'intérêt du service ou à la richesse de son contenu, mais aussi à son aspect.

Par aspect, il faut entendre bien plus qu'aspect visuel ou choix graphiques : l'aspect d'un site internet rassemble les notions d'ergonomie, de rapidité, de confort visuel - entre autres caractéristiques. Ces points devront être soignés, par la conception d'une interface ergonomique, par la spécification d'une charte graphique cohérente et sobre, et par des choix de réalisation judicieux et basés sur le fonctionnement des outils de l'internet, permettant une rapidité maximale.

4.2.4 Accès à l'information

Le public désireux de consulter une thèse est un public recherchant une information précise : recherche sur un sujet précis, ou d'un document précis.

Notre service doit dont fournir à l'utilisateur les moyens de cette recherche. Ceci passe par l'emploi de systèmes de recherche (par mots clés, par un ou des mots du titre, par le nom de l'auteur, par année, etc.), par la présentation de listes (listes thématiques, listes alphabétiques par auteur ou chronologique).

Certaines listes ont pour intérêt de présenter « a plat » l'ensemble des documents (liste chronologique ou par auteur). D'autres (listes thématiques) ont pour ambition supplémentaire d'orienter la recherche. Une liste hiérarchisée par disciplines, spécialités, etc. (comparable aux listes de UMI) peut en être un bon exemple.

Concernant également les listes, la forme de celles-ci est importante. Les informations à afficher pour chaque entrée de liste peuvent être diverses. Les rubriques des listes thématiques peuvent afficher le nombre de document concernés. Des informations telles que le format du document, sa taille ou ses caractéristiques particulières (présence de vidéo, etc.) peuvent avantageusement orienter le lecteur.

Enfin, les serveurs HTTP se voient actuellement adjoindre des outils permettant de pallier au caractère « sans connexion » de la consultation web. Il devient dès lors possible de conserver une *session* regroupant les recherches effectuées par l'utilisateur. Par exemple, un bouton de l'interface peut présenter à l'utilisateur les recherches qu'il a effectuées précédemment (leurs critères, leur résultat), afin d'ajouter une nouvelle souplesse au service.

4.2.5 Retour d'information

Le *feed-back*, ou retour d'information, est un point important pour l'élaboration et la maintenance d'un service ayant la double ambition d'être efficace et apprécié.

La solution du courrier électronique simple (simple lien vers une adresse mail depuis les pages web) est certes efficace. Cependant, il est possible que l'utilisateur n'ait pas accès à son compte de messagerie depuis le poste de consultation. Dans ce cas, la création de formulaires permettant de faire une remarque, de signaler une anomalie précise ou d'adresser une doléance à Doc'INSA est envisageable.

Par extension, il est imaginable que les auteurs puissent soumettre leurs documents source via le service web. Ceci implique que quelques pages présentent la procédure et les règles ou modèles à respecter par les auteurs afin que leur document soit publiable dans les meilleures conditions possibles.

5 Analyse du domaine

Cette analyse utilise les méthodes et les formalismes de OMT (technique de modélisation orientée objet). Ces formalismes sont d'une lisibilité relativement facile et leur compréhension ne devrait pas poser de problème à un lecteur connaissant le domaine de l'étude.

5.1 Données

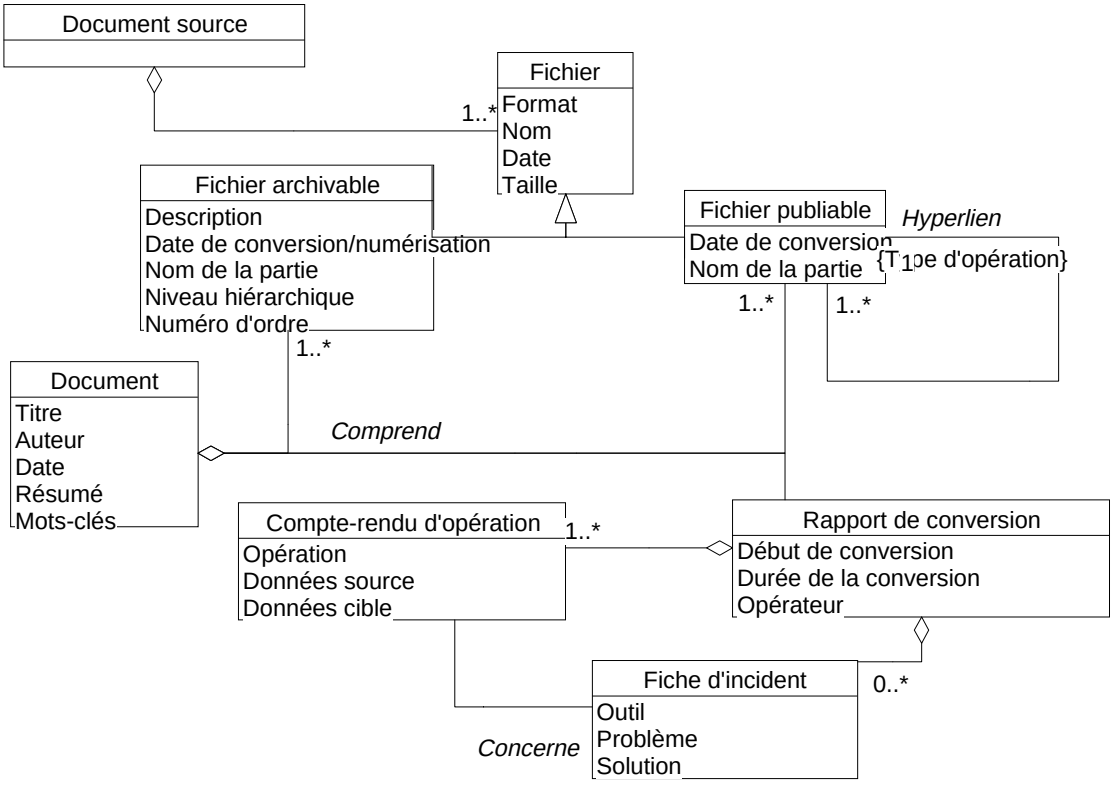
Cette partie décrit les modèles de données sur lesquels s'appuie la chaîne d'édition.

5.1.1 Modèle objet du domaine

Le domaine de l'étude a été modélisé, ce modèle constitue une première approche formelle du domaine de l'étude.

Les objets de la publication électronique sont principalement les documents (documents source, documents publiables ou archivables). Ces documents sont des documents électroniques : ils sont contenus dans des fichiers.

L'application à réaliser est la chaîne de traitement des documents en vue de leur publication électronique. Cette chaîne va rassembler des interventions sur la forme des documents et sur leur format. Toutes les opérations de la chaîne seront inscrites dans un rapport de conversion, afin d'assurer une traçabilité de la production mais également pour constituer une base de connaissances.



- Figure 2 : Modèle objet du domaine

5.2 Traitements

Cette partie présente les modèles des traitements de la chaîne d'édition.

Dans cette partie, les traitements marqués comme le présent paragraphe seront effectués 'manuellement'.

5.2.1 Vue d'ensemble de la procédure

Le backoffice rassemble les fonctions suivantes :

- Le traitement *avant conversion* des documents reçus
- La numérisation des documents papier (photos, etc.)
- L'archivage des documents retraités et des documents numérisés
- La conversion des documents en PDF
- L'insertion des annexes, photos, vidéos, dans les documents produits
- L'ajout éventuel d'hyperliens aux documents PDF
- L'optimisation des fichiers PDF pour la lecture *en ligne*
- La publication des documents électroniques résultats des opérations suscitées

Une part de ces opérations (traitement avant conversion, conversion PDF, archivage, etc.) sera automatisée (autant que faire se peut).

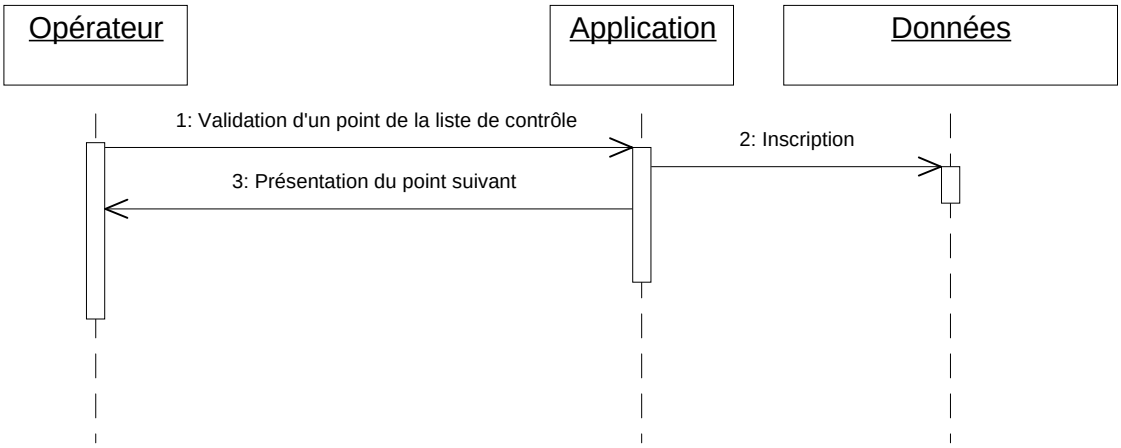
L'application d'édition électronique, objet de cette étude, fournira à l'opérateur le maximum d'aides au traitement des documents (listes de contrôle, lancement des outils, base de connaissance et journalisation des opérations et incidents).

Les modèles de scénarios suivants décrivent le comportement général de l'application.

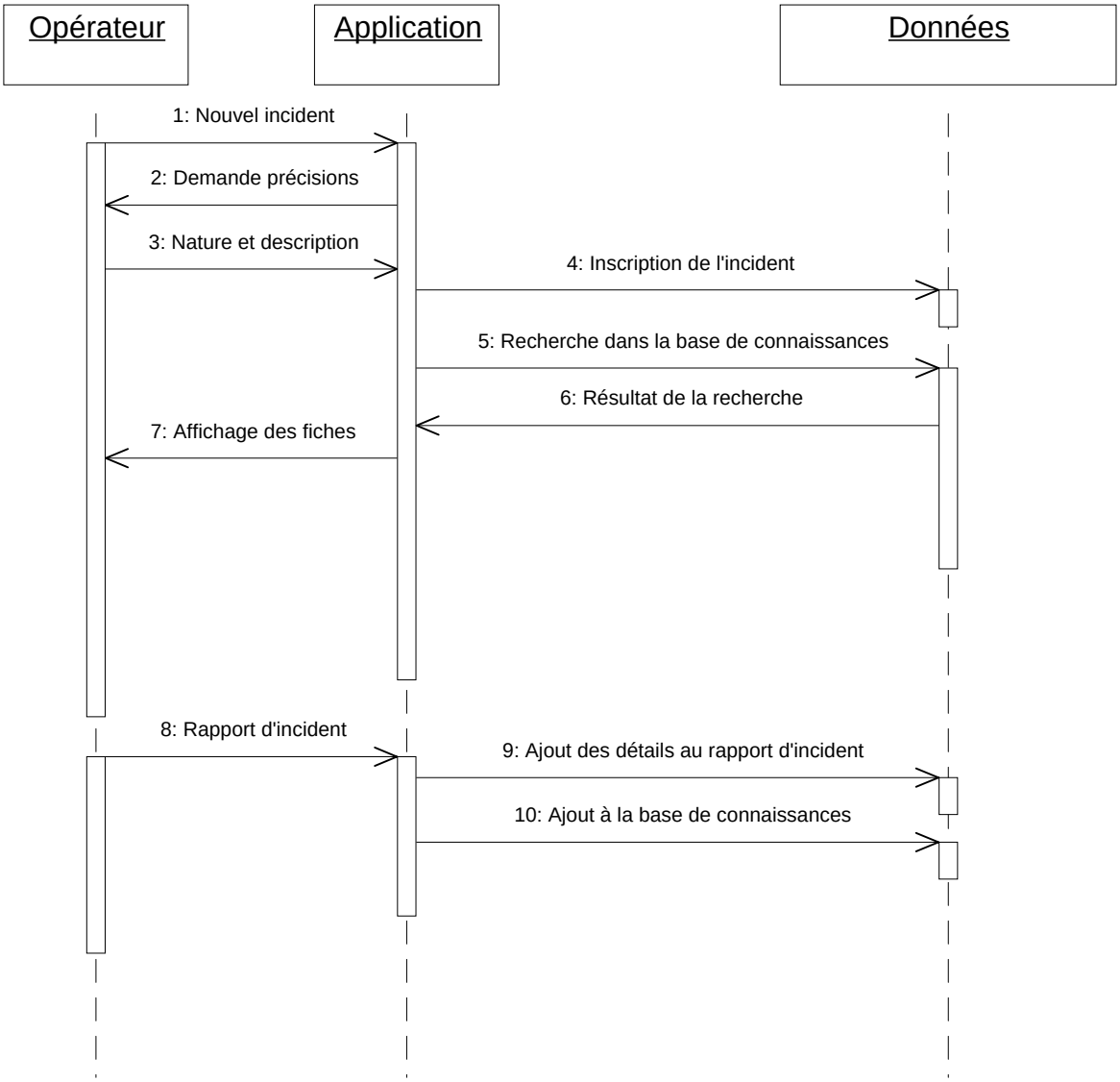
5.2.1.1 Fonctionnement général de l'application

L'application, outre les différentes aides à l'édition électronique proposées à l'opérateur au fur et à mesure de la 'genèse' du document publiable, offre un accès permanent à quelques fonctions :

- Liste de contrôle (pointée par l'opérateur, elle permet à l'application de suivre l'avancement de la procédure). Elle fournit une aide 'contextuelle' à cet opérateur.
- Base de connaissances. Formée sur la base de rapports d'incidents, elle permet à l'opérateur de savoir si un incident (problème en général lors de l'édition électronique) est répertorié, et s'il a été résolu. L'opérateur peut ajouter des rapports d'incident (résolus ou non) à cette base.



- Figure 3 : Scénario de validation d'un point de la liste de contrôle



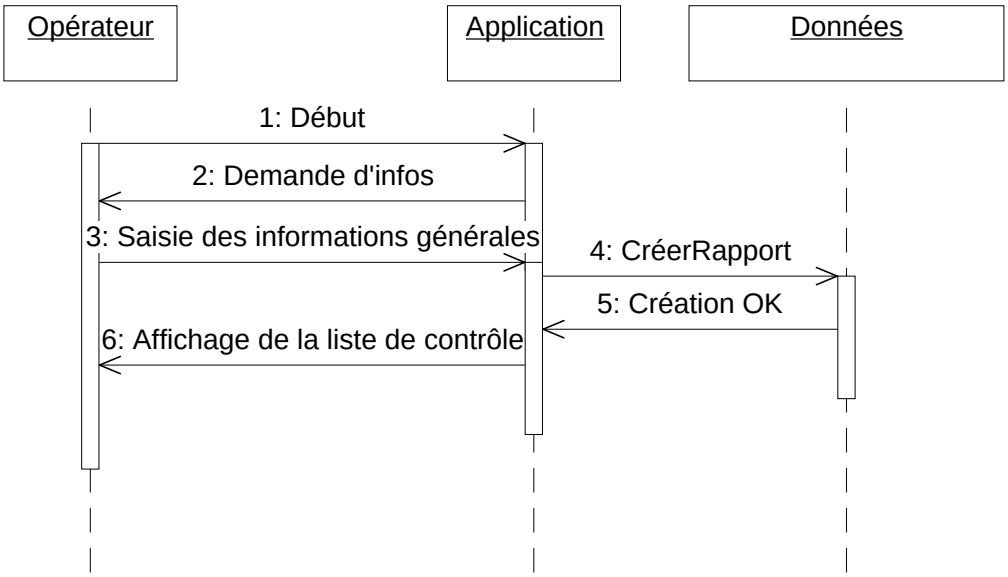
- Figure 4 : Scénario de traitement des incidents

5.2.1.2 Traitement avant conversion

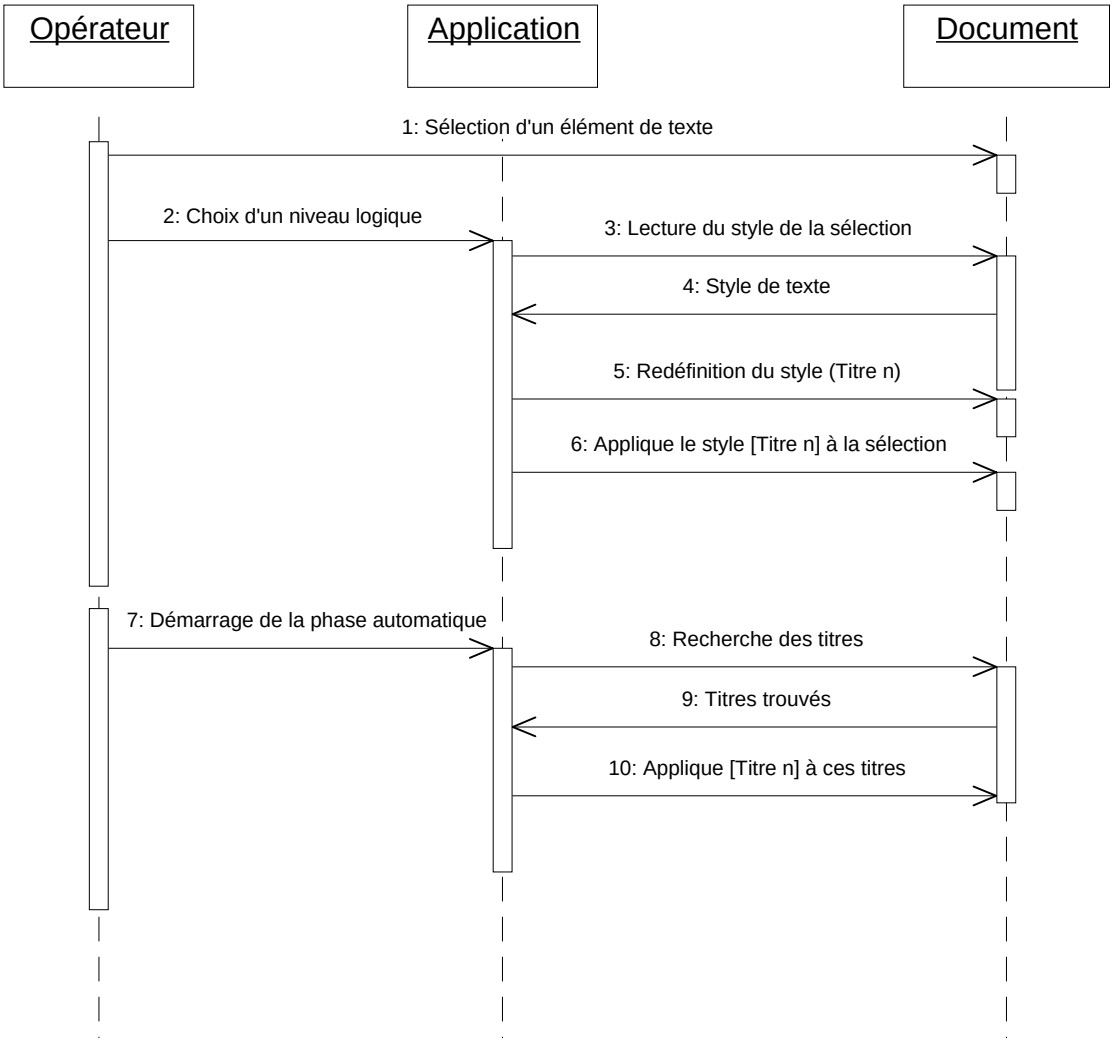
Les documents reçus (sur le serveur FTP, sur CD-ROM ou autre archive) seront copiés sur le disque de travail du poste d'édition.

Ils devront alors être 'vérifiés' suivant une courte procédure, et préparés si besoin est à la conversion en document(s) propre(s) à être édité(s).

- Démarrage de la procédure
    - Création du rapport de conversion
        - Saisie des caractéristiques intrinsèques
    - Elaboration automatique d'une liste de contrôle (*check-list*) de conversion
- Vérification de la forme du document reçu :
    - Organisation des fichiers (arborescence, nombre et format, système de noms)
    - Format (et éventuellement version du logiciel) des fichiers
- Analyse du document avant conversion :
    - Identification du sommaire général et de la structure complète du document
    - Rassemblement ou découpage des documents par chapitres et annexes
    - Vérification de la mise en forme (marges, styles, etc.)
    - Identification des documents à numériser
    - Rapport sur la structure et les composantes du document (validation dans la liste de contrôle)
- Révision du document source
    - Détection de la structure logique si elle n'est pas définie (document Word) et extraction de cette structure pour la génération des liens intra document



- Figure 5 : Scénario de démarrage de la procédure d'édition électronique



• Figure 6 : Scénario de révision de la structure logique d'un document

5.2.1.3 Numérisation des documents papier

Les documents papier devant être intégrés au document final ont été identifiés durant la phase décrite en [5.2.1.2]. Ils seront numérisés sur le poste d'édition.

Des consignes précises régiront la numérisation. *S'il est possible d'effectuer directement ces numérisations à partir des outils d'Adobe Acrobat, cette solution ne sera pas retenue pour des raisons de format d'archivage.*

- Identification de la forme de chaque document, qui décidera des paramètres de numérisation
- Numérisation
- Elaboration d'un rapport de numérisation (listant les documents numérisés, leur format, etc.)

La liste des documents à numériser étant issue de l'étape précédente, et le format de numérisation dépendant de la forme de l'original, l'opérateur n'aura qu'à saisir le format des documents.

L'application apportera alors à l'opérateur les directives de numérisation, qui n'aura qu'à accepter ou corriger ce format suivant ses propres choix.

5.2.1.4 Archivage des documents retraités et des documents numérisés

Un archivage systématique des documents reçus par Doc'INSA sera réalisé. Les documents archivés seront issus du retraitement décrit en [5.2.1.2] et d'éventuelles numérisations [5.2.1.3]. L'ensemble des archives constituera par conséquent un ensemble 'relativement' homogène, rassemblant pour chaque document : le(s) document(s) source retraités, les numérisations éventuelles, et les rapports correspondants.

5.2.1.5 Conversion des documents en PDF

Le format de publication est le format Acrobat d'Adobe.

Les choix de format ou résolution concernant ces fichiers 'cible' sont dictés par les considérations suivantes :

1) Ces documents ont pour destination la consultation en ligne via internet,

2) Le document électronique doit permettre une impression comparable à son homologue papier.

Le respect de la pagination d'origine n'est pas garanti par notre chaîne de production.

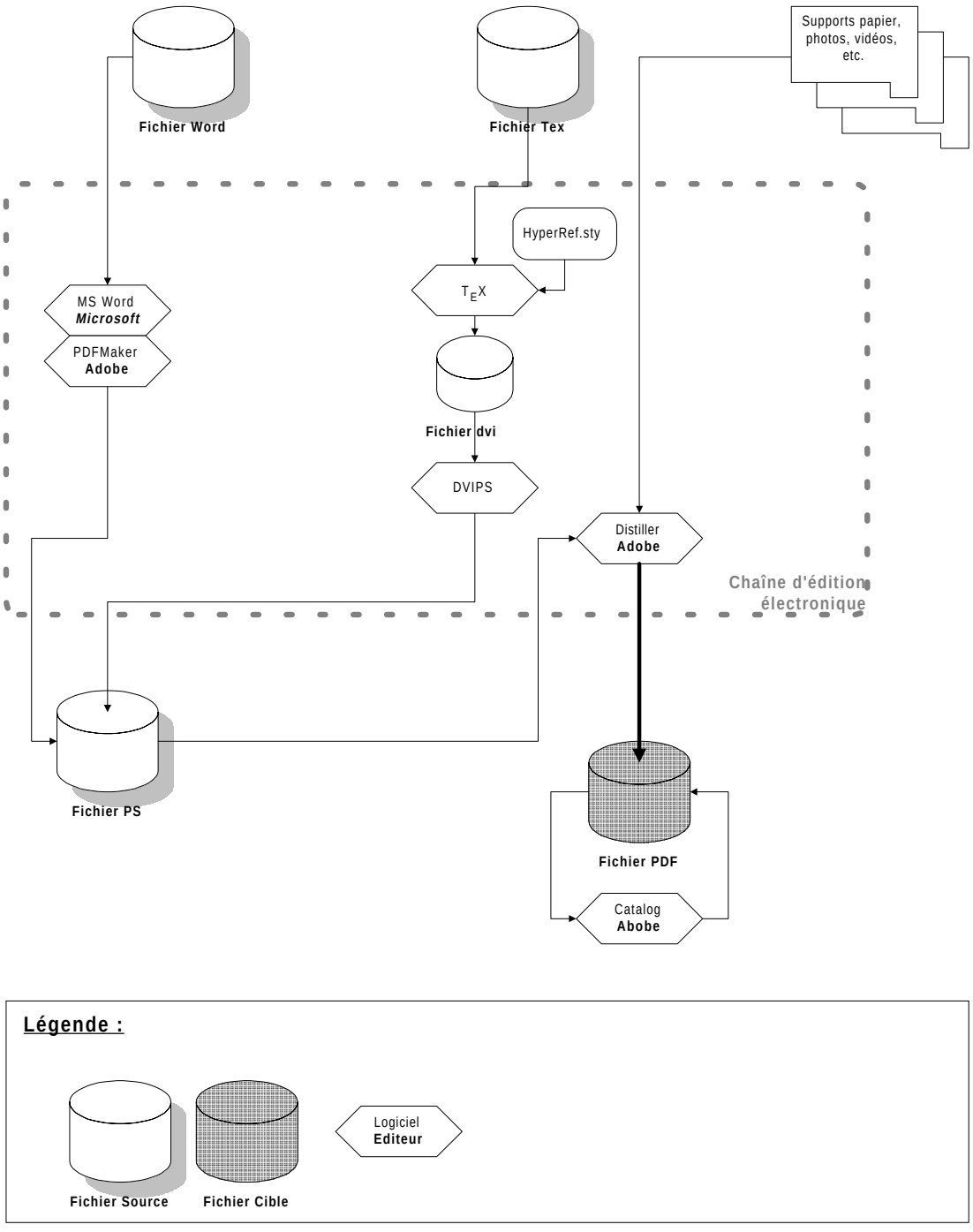
Les documents seront donc optimisés pour une consultation en ligne, mais la résolution des images ou autres éléments non vectoriels sera suffisante pour une impression laser de qualité (par exemple 300 ppp.).

Les outils de la conversion en PDF (format Acrobat) différent suivant le format source. Cependant, il est une constante facile à imposer et respecter : le format PostScript constituera un format intermédiaire.

En effet, les outils de création de PDFs 'avancés' (conversion automatique des hyperliens, des renvois, du sommaire du document source) produisent des fichiers PostScript contenant des codes supplémentaires à l'intention d'Acrobat Distiller.

- Génération de fichiers PostScript
    - Conversion automatique au format PostScript des fichiers constituant le document, identifiés lors du traitement avant conversion [5.2.1.2]
        - Les fichiers MSWord sont convertis par PDFMaker
        - Les fichiers Tex sont convertis via hyperref et DVIPS
- Ajout d'informations *Distiller* aux fichiers PS
    - Liens

    - Auteur, titre, etc..
- Conversion automatique de PostScript en PDF par Acrobat Distiller
- Ajout au rapport de conversion (incidents, durée)



- Figure 7 : Synoptique de la conversion des documents

5.2.1.6 Insertion des annexes, photos, vidéos

Les fichiers Acrobat étant prêts, ils peuvent être 'retraités' par un opérateur. Dans le module Exchange de Acrobat, il est possible d'ajouter des liens vers des fichiers extérieurs (vidéos, etc.) ou d'insérer des images (converties en PDF) dans les pages existantes.

- D'après les rapports établis lors du traitement avant conversion (et listant les documents numérisés à insérer), insertion de ces documents
- Inscription des insertions dans le rapport de conversion

5.2.1.7 Optimisation des fichiers PDF

Les fichiers PDF version 3 peuvent être 'optimisés' pour la lecture en ligne. Plus précisément, des marques sont insérées dans le fichier, permettant au serveur de n'envoyer que les pages demandées par le client.

Cette optimisation est effectuée par Acrobat Exchange. Elle sera facilement automatisée.

- Optimisation de l'ensemble des fichiers PDF de la thèse
- Inscription dans le rapport de conversion

5.2.1.8 Publication des documents

Les documents produits doivent être copiés sur le serveur HTTP, suivant une organisation stricte et dans un répertoire facilement identifiable par la suite (numérotation systématique des répertoires, par exemple).

L'inscription dans le catalogue (indexation dans la base Doris) et la création (ou l'extraction) de la page de résumé interviennent après la copie des documents, leur URL étant fixée.

- Copie des fichiers sur le serveur
- Inscription dans Doris
- Clôture du rapport de conversion

6 Forme des documents électroniques à produire

Cette annexe décrit sommairement la forme qu'auront les documents produits par la chaîne d'édition objet de la présente étude.

6.1 Liens et hyperliens au sein d'une thèse

La navigation entre les pages d'une thèse peut se faire de façon séquentielle. Cependant, l'ajout d'hyperliens et de "repères" Acrobat permet un accès plus rapide à l'information (et est rendu obligatoire par le découpage en chapitres).

6.1.1 Types de liens utilisables dans le format PDF

6.1.1.1 Repères

Les repères d'Acrobat sont les étiquettes texte affichables à gauche de la fenêtre d'Acrobat Reader. Ces repères peuvent pointer sur des "vues" au sein du document courant, mais aussi vers d'autres fichiers (PDF ou non). En fait, toute une panoplie d'actions peut leur être attachée.

6.1.1.2 Liens

Acrobat permet de définir des zones sensibles, superposées au contenu d'une page, auxquelles il est possible d'associer différentes actions.

Ces "liens" peuvent ainsi pointer vers une "vue" dans le document en cours, ou bien lancer l'ouverture d'un autre document (PDF ou non).

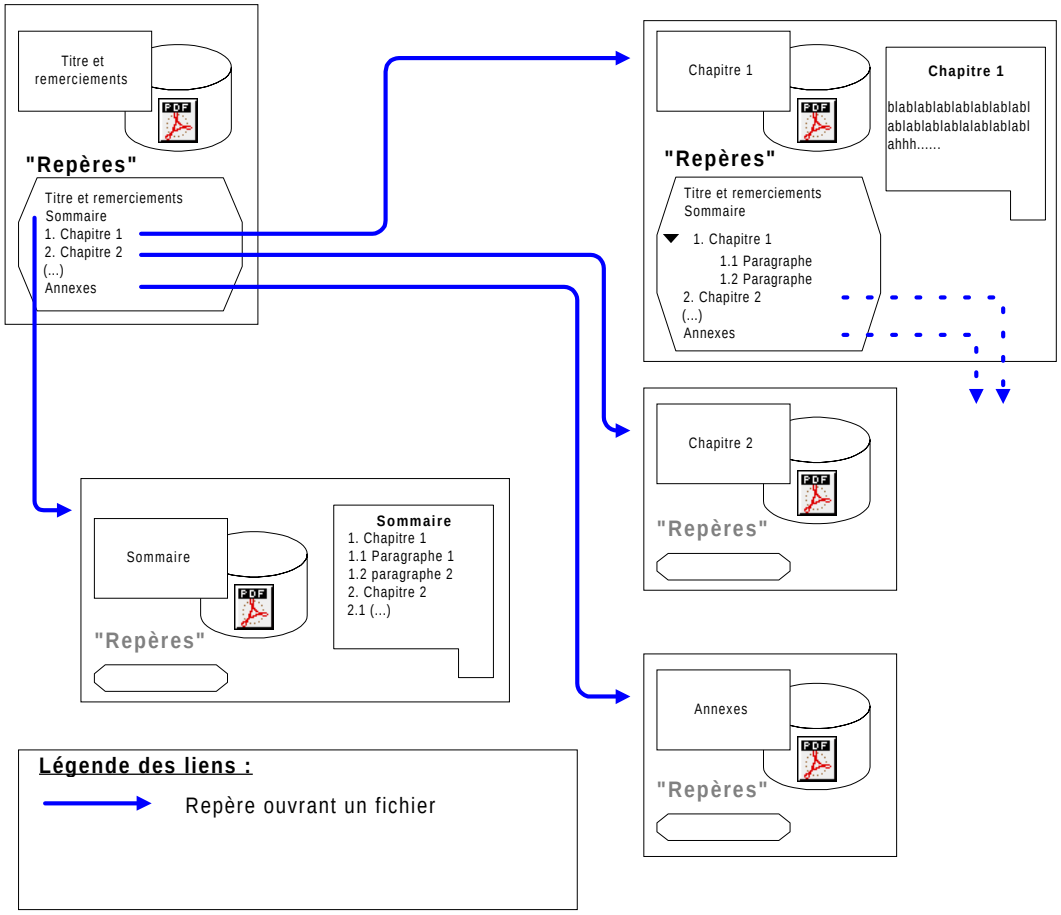
6.1.2 Structure type d'une thèse

Afin que l'ensemble soit cohérent, les thèses proposées sur le serveur auront la même structure (qui sera également une conséquence naturelle de l'automatisation partielle de la production des PDF). Cette structure est présentée rapidement ci-après.

6.1.2.1 Repères entre les documents constituant la thèse

Les différents fichiers PDF constituant une même thèse auront en commun le premier niveau hiérarchique des repères, chaque étiquette de ce niveau correspondant à un fichier PDF.

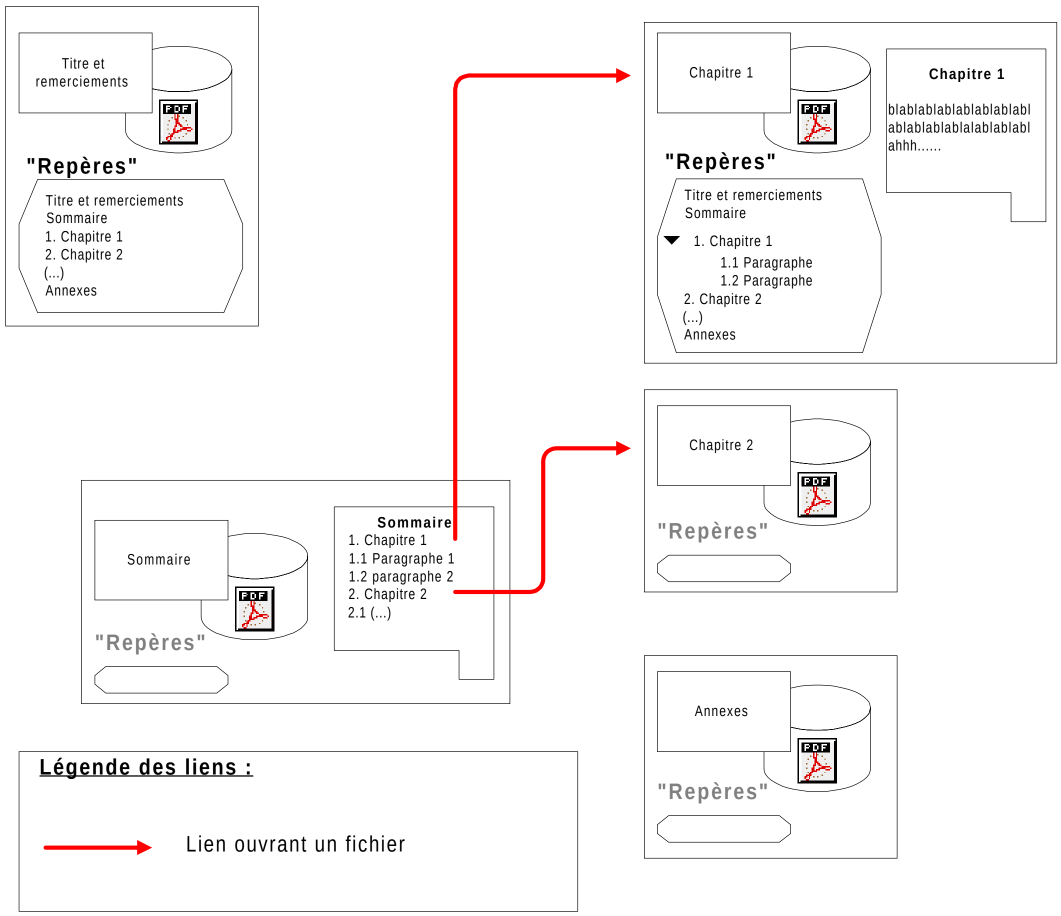
Ainsi, la navigation entre ces fichiers devient transparente pour l'utilisateur.



• Figure 8 : Liens et hyperliens – repères entre les fichiers

6.1.2.2 Liens vers les différents chapitres depuis le texte du sommaire

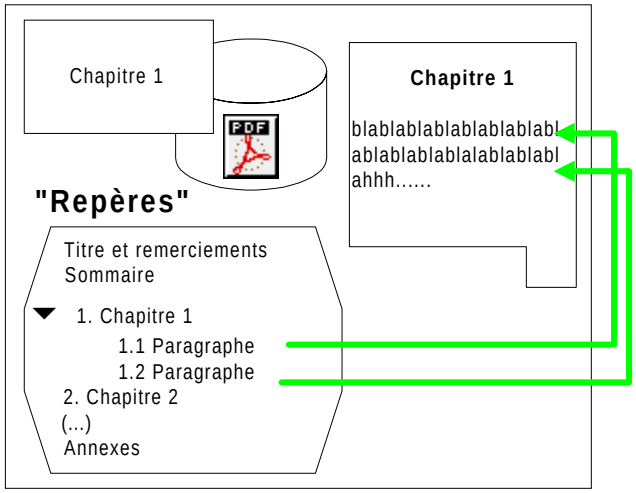
A la lecture du sommaire, l'utilisateur pourra cliquer sur un titre et accéder directement au chapitre correspondant. Il n'atteindra pas le titre choisi, les liens ne pointant que sur les débuts de chapitre. Cependant, il pourra ensuite retrouver le titre à consulter dans les repères de ce chapitre (voir ci-après).



• Figure 9 : Liens et hyperliens – Liens entre fichiers

6.1.2.3 Navigation dans chaque chapitre à l'aide des repères

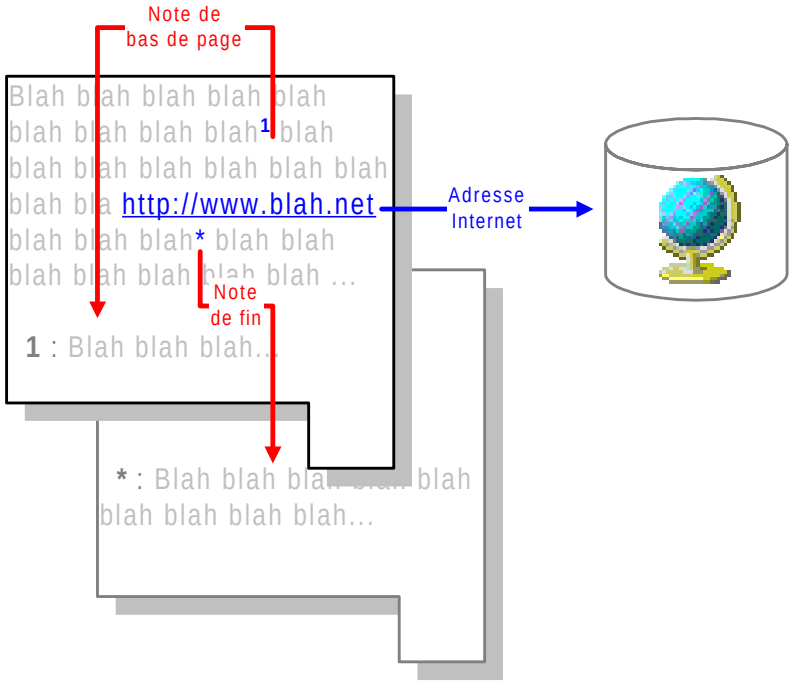
La structure logique de chaque chapitre est reprise en détail par les repères Acrobat. Ainsi, la hiérarchie des titre y est reprise et un clic sur un repère "saute" au texte correspondant.



• Figure 10 : Liens et hyperliens – Repères dans le même fichier

6.1.2.4 Autres liens au sein d'un document

D'autres liens sont générés automatiquement lors de la conversion au format PDF : les renvois aux notes de fin ou de bas de page sont "sensibles" et renvoient aux libellés correspondant. Les adresses internet (URL) sont également converties en liens accédant à l'adresse à l'aide du programme approprié.



• Figure 11 : Autres liens dans les documents