FICHES TECHNICO-COMMERCIALES

DOCUVISION™

La segmentation

Version 1.0

Référence: CT/JMM/90001
Révision 1.0 du Samedi 13 Novembre 1993
Auteur: Jean-Michel MARCASTEL
Diffusion: Direction Commerciale

Révisions

Date	Version	Révision
Sam 13 Nov 1993	1.0	Version préliminaire

DOCUVISION™ est une marque déposée de Litton/MC2.

La loi du 11 mars 1957 n'autorisant, aux termes des alinéas 2 et 3 de l'article 41, d'une part, que les "copies ou reproductions strictement réservées à l'usage privé du copiste et non destinées à une utilisation collective" et, d'autre part, que les analyses et les courtes citations dans un but d'exemple et d'illustration, "toute représentation ou reproduction intégrale, ou partielle, faite sans le consentement de l'auteur ou de ses ayants droit ou ayants cause, est illicite" (alinéa 1er de l'article 40).

Cette représentation ou reproduction, par quelque procédé que ce soit, constituerait une contrefaçon sanctionnée par les articles 425 et suivants du Code Pénal.

Présentation

Les récents efforts de normalisation des représentations informatiques des documents, qui ont donné naissance aux normes internationales SGML et ODA/ODIF, permettent d'espérer qu'à terme les documents seront échangés selon des formats électroniques reconnus par tous les systèmes de manipulation de documents.

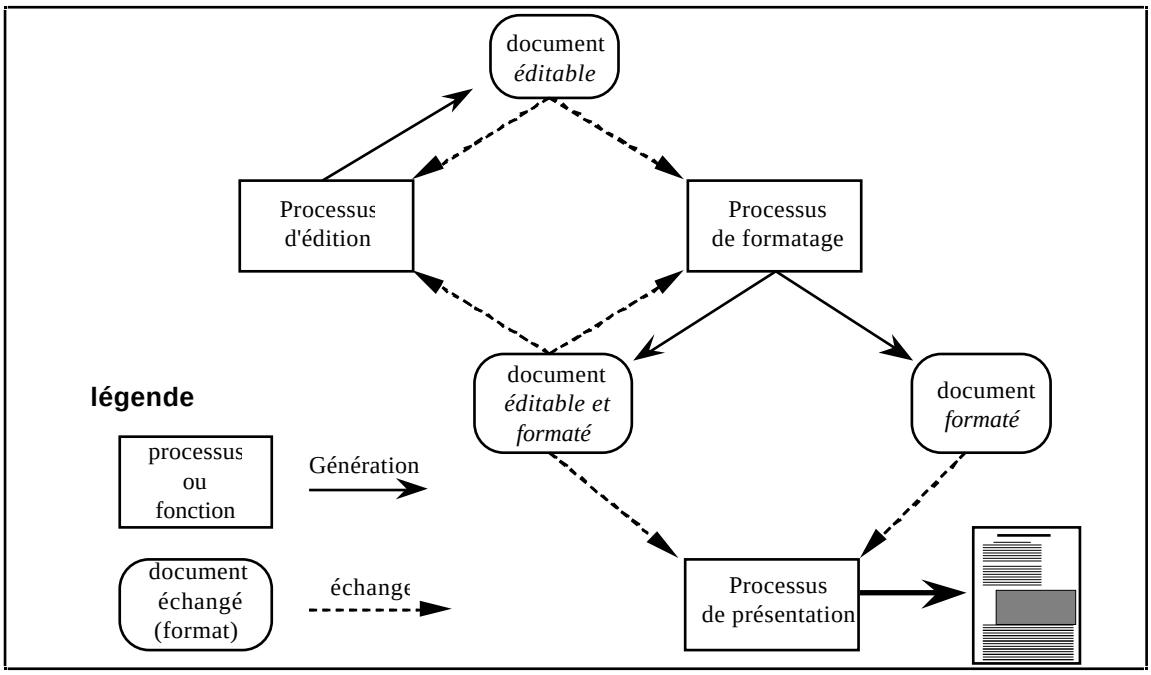

Figure XX. Présentation fonctionnelle d'un serveur de documents ODA.

Conformément à l'approche définie par la norme ODA, un document comprend deux structures distinctes et complémentaires.

- ◊ la *structure physique* précise comment les contenus sont disposés dans le document; un document est constitué d'un ensemble de pages, elles-mêmes structurées en une hiérarchie de blocs qui peuvent représenter des colonnes, des pavés de contenu, etc.,
- ◊ la *structure logique* organise le contenu d'un document en une hiérarchie d'objets logiques, tel que le titre, chapitre, tableau, paragraphe, résumé, etc.

Selon cette approche, la segmentation MC2 est séparée en deux opérations disjointes:

◊ la *segmentation physique* permet de retrouver en premier lieu des éléments liés aux règles d'édition,

◊ la *segmentation logique* permet d'obtenir une séparation des informations suivant les règles d'organisation.

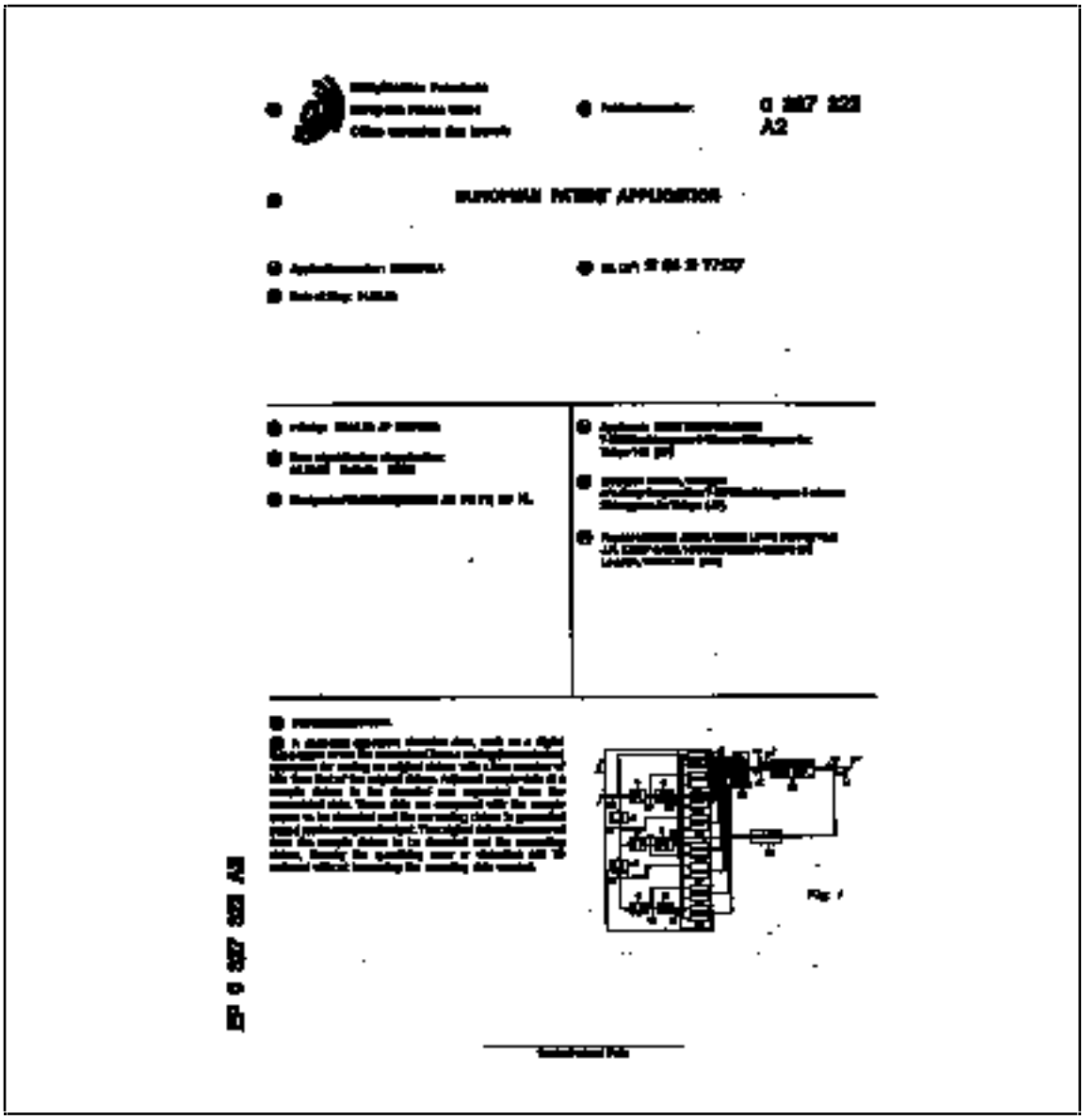

Figure XX. Exemple de document technique.

Segmentation physique

La segmentation physique tend à retrouver la fonction inverse de la fonction d'édition qui, à partir du contenu de l'information présentée, des règles typographiques et des contraintes d'impresion, a abouti au placement de chaque bloc et à l'organisation de ceux-ci les uns par rapport aux autres.

L'approche retenue est une méthode ascendant qui, partant des objets élémentaires - amas de points ou contours connexes, permet la construction de blocs structurés suivant les règles d'édition dans le sens blocs simples -> page. Elle permet notamment une analyse sans contraintes a priori sur la page analysée dans la mesure où le nombre de décisions arbitrraires est minimisé.

La segmentation physique peut être décrite comme l'enchaînement des traitements suivants:

◊ recherche de tous les objets graphiques ,

◊ agglomération d'objets selon des règles de proximité, de dimensions, de nature, etc.,

◊ validation des lignes par contrôles de conformité (homogénéité, rectitude, …),

◊ regroupement des lignes en paragraphes,

◊ validation des zone de texte.

Chaque traitement implique l'emploi de règles paramétrables déduites de l'observation préalable d'un grand nombre de documents, ou dont la nécessité est apparue pour résoudre de manière la plus générale possible le maximum de cas particuliers.

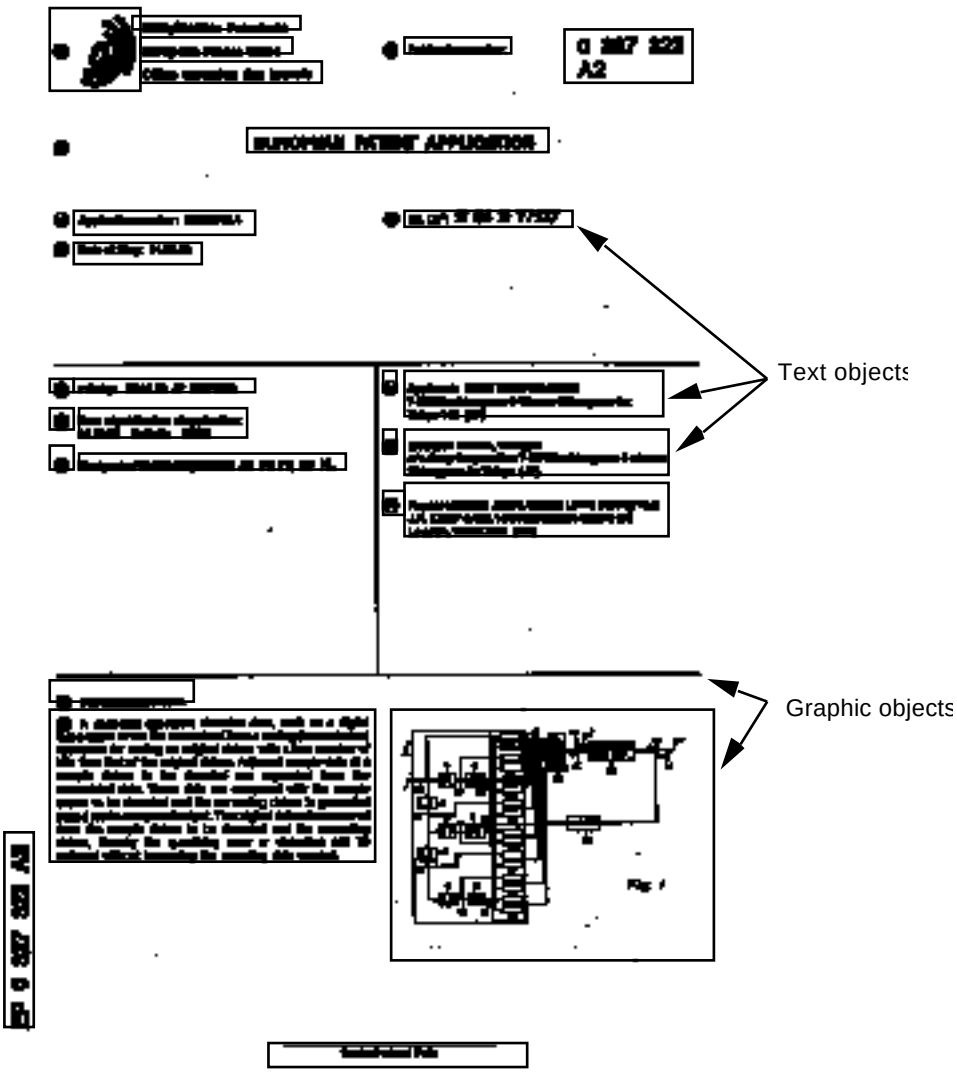

Figure XX. Structure physique après segmentation.

Segmentation logique

La segmentation logique permet d'identifier des suites d'unités logiques d'un document à partir de l'organisation physique des éléments d'information, et des règles de définition de reconnaissance des unités logiques associées. Elle traduit en term macroscopique la structure physique obtenue.

Les unités logiques sont décrites par des grammaires dont les éléments de vocabulaire terminaux sont les blocs fournis par la segmentation physique et dont les éléments de vocabulaire non terminaux représentent les sous-parties d'une unité logique.

Par exemple, une marque internationale est définie dans l'ordre par:

◊ une date: ligne de texte fer à gauche,

◊ une durée: ligne de texte centrée, de même niveau que la date,

◊ une référence: ligne de texte fer à droite, de même niveau que la date,

◊ un logo: bloc graphique centré ou bloc de texte dont la taille de fonte est supérieure à un seuil,

◊ un descriptif: suite de blocs de texte,

◊ un séparateur: trait horizontal.

Les notions de "date", "durée", "référence", "logo", "descriptif", "séparateur" sont des mots non terminaux.

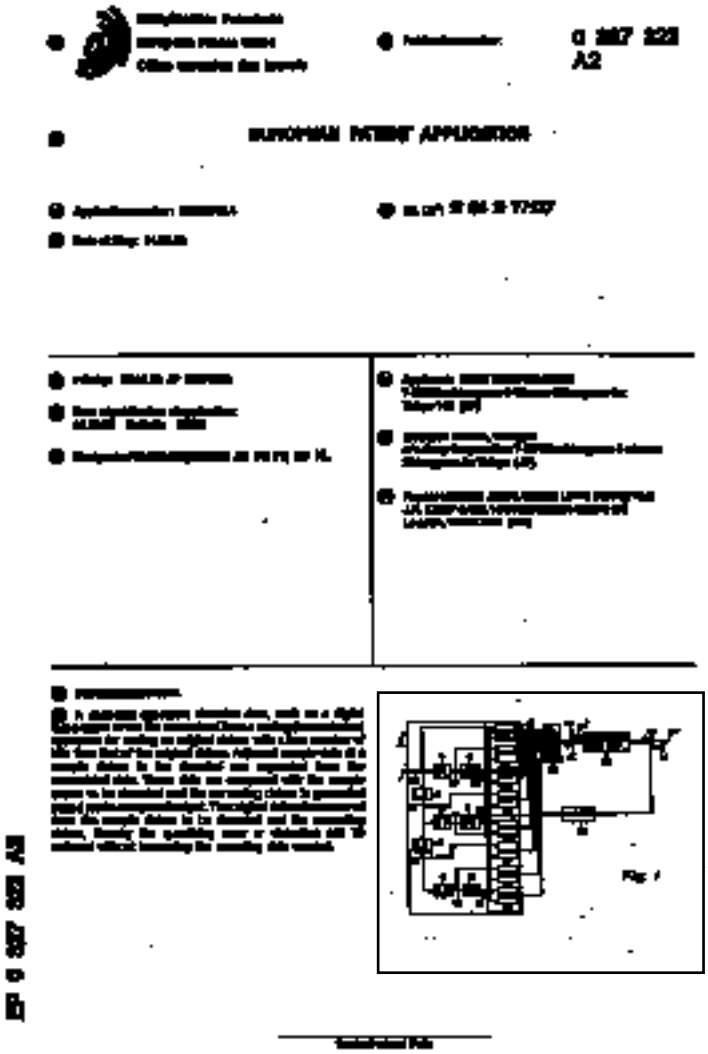

Figure XX. Structure logique après segmentation.

Configuration et apprentissage

La segmentation physique décrit la structure sous la forme d'éléments d'un vocabulaire terminal à partir duquel l'organisation cliente a défini des phrases types en introduisant des éléments d'un vocabulaire non terminal; la segmentation logique effectue une étape de reconnaissance syntaxique pour former une phrase non ambigüe.

Toute erreur de syntaxique est une faute de la segmentation physique au sens de la détection; toute erreur grammaticale est soit une erreur de la segmentation au sens du regroupement abusif en blocs soit une erreur dans les règles ou encore un cas particulier. Dans tous les cas une étape de validation/correction permet la correction des fautes (par simplification, ajout, ect.) des mots dans la phrase.

Mise en œuvre

La finalité de la segmentation est donc de permettre une interprétation du contenu logique des documents techniques de l'entreprise. Cette interprétation, moyennant l'adjonction de modules adéquats tels que la reconnaissance optique de caractères ou de reconnaissance de primitives graphique, permet d'envisager toutes les possibilités d'une simple redisposition ou réorganisation de l'information à sa réutilisation dans les circuits informatiques et de conception et de réalisation de la documentation technique.

L'interprétation, phase utlime de la segmentation appelé formattage client est nécessairement adapté cas par cas.

Ce chapitre présente une architecture type à base de la segmentation MC2. Son intérêt est de présenter les configurations matérielle et logicielle requise.

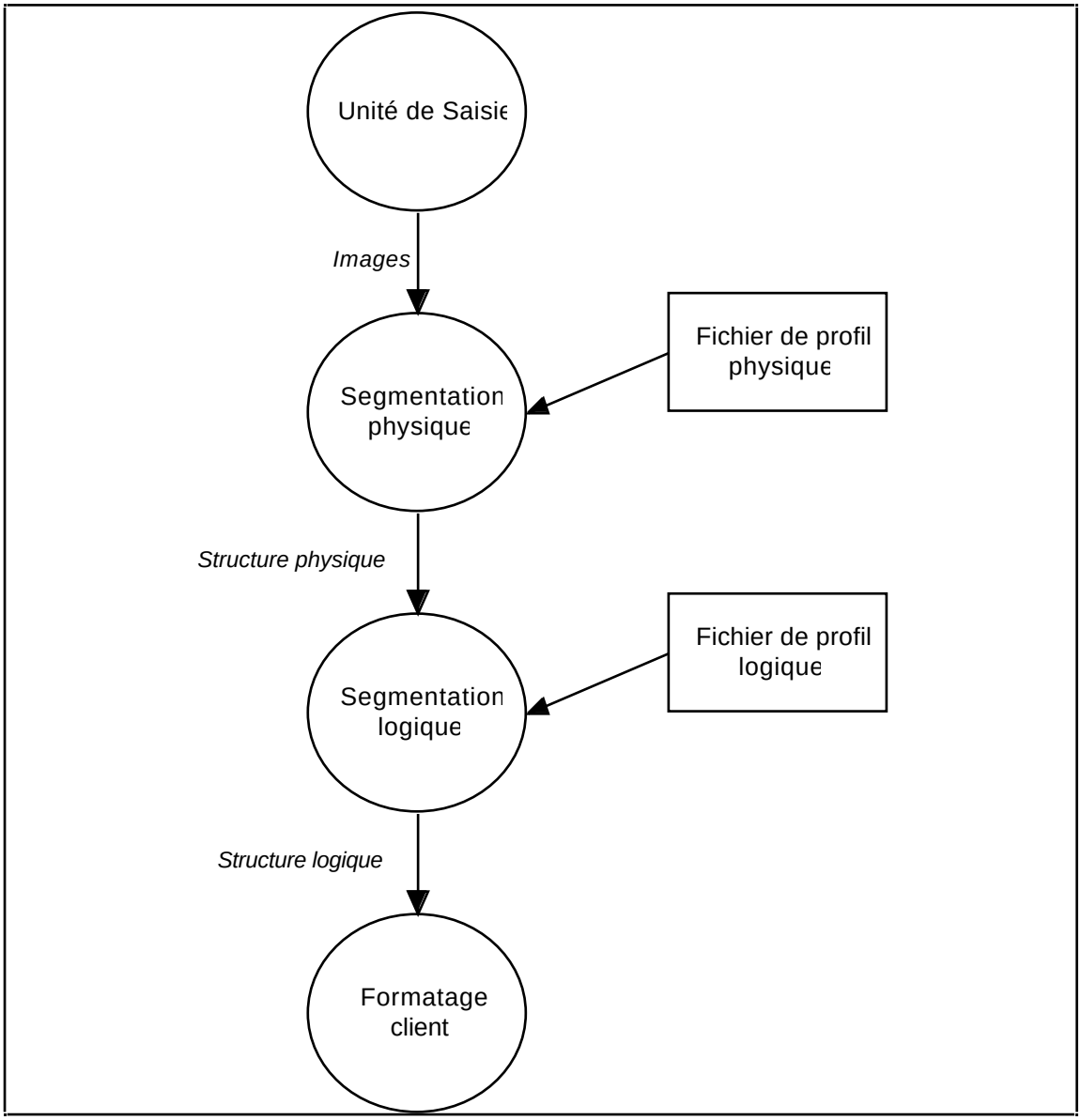

Figure XX. Organisation fonctionnelle de la segmentation MC2.

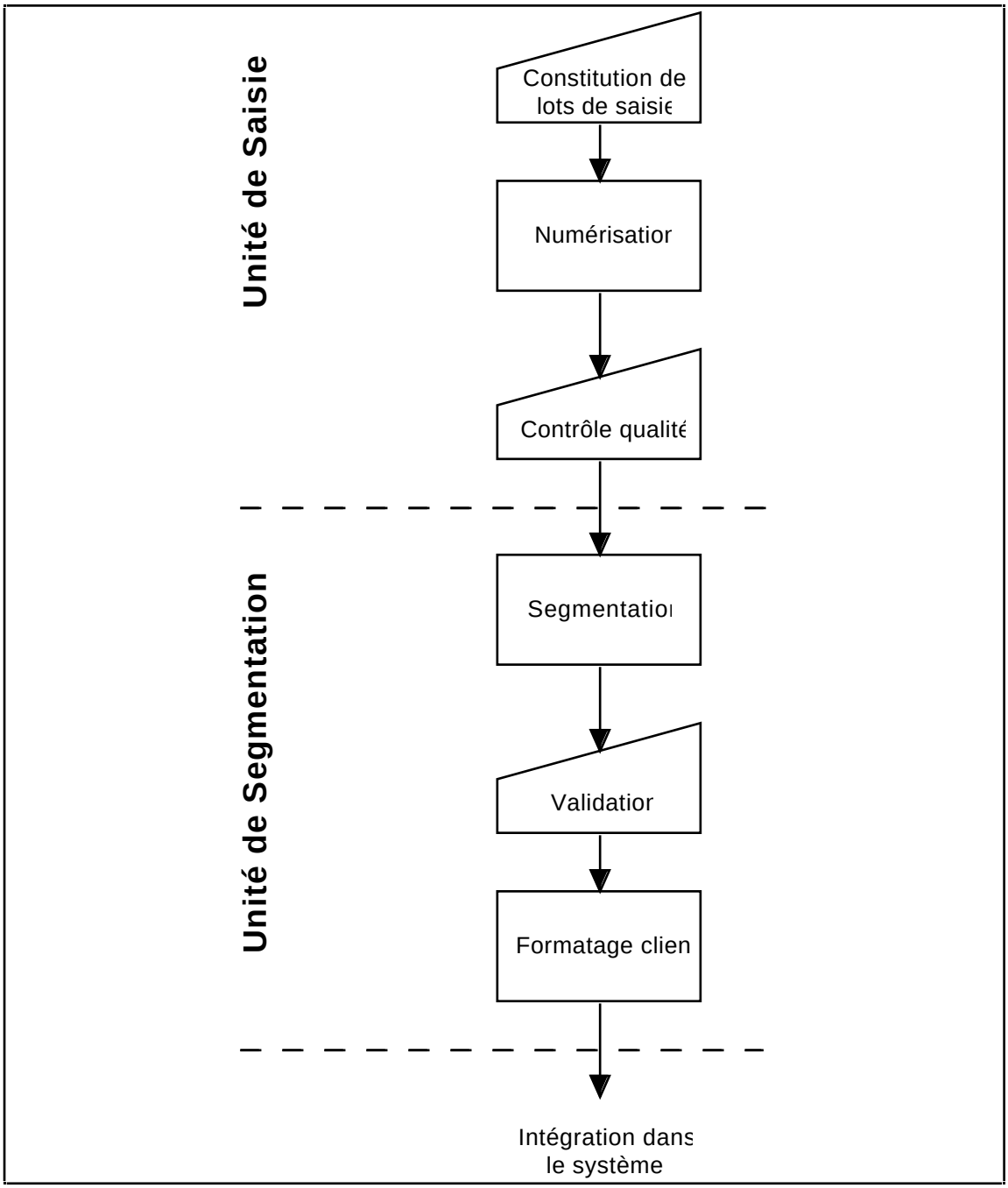

Figure XX. Exemple de traitement intégrant la segmentation.

OBJECTIFS

Soit une entreprise disposant d'une très volumineuse documentation technique devant être consultée par des centaines de personnes dans différentes parties du monde.

Chaque document est précédé d'une page de garde comportant les informations suivantes:

◊ numéro de publication,

◊ organisme émetteur,

◊ auteur,

◊ date d'enregistrement du document,

◊ abstract,

◊ graphique: dessin, formule scientifique, synoptique, etc.

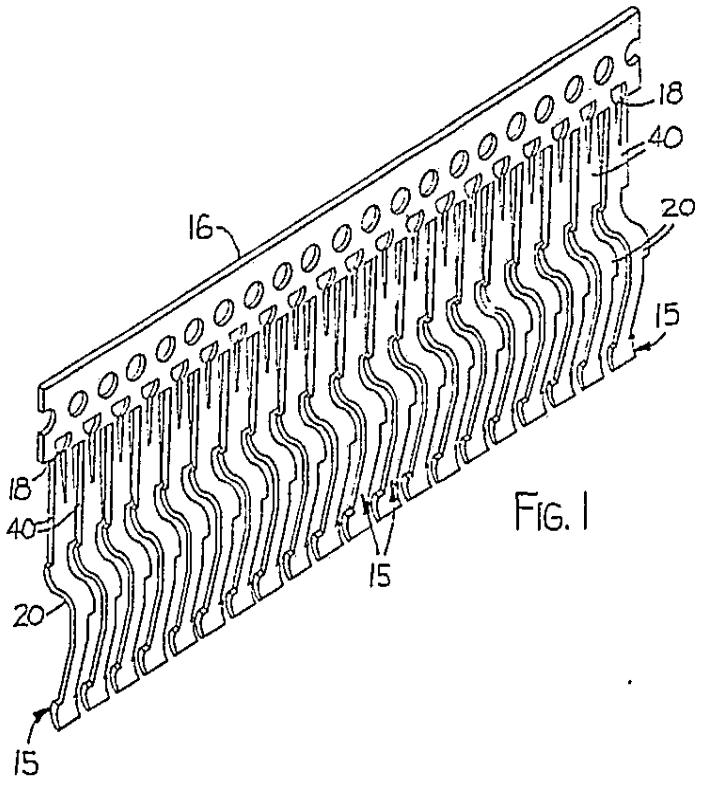

Figure XX. Exemple de dessin.

Afin de rendre accessible cette documentation technique, l'entreprise envisage l'acquisition d'un système de gestion électronique de documents mettant en œuvre un concept de diffusion: après consultation d'une base bibliographique comportant l'ensemble des informations portée sur la page de garde, les utilisateurs commande une copie totale ou partielle du document. L'ensemble de la documentation technique étant stockée dans le système.

La base bibliographique doit donc comporter des éléments textuels et graphiques.

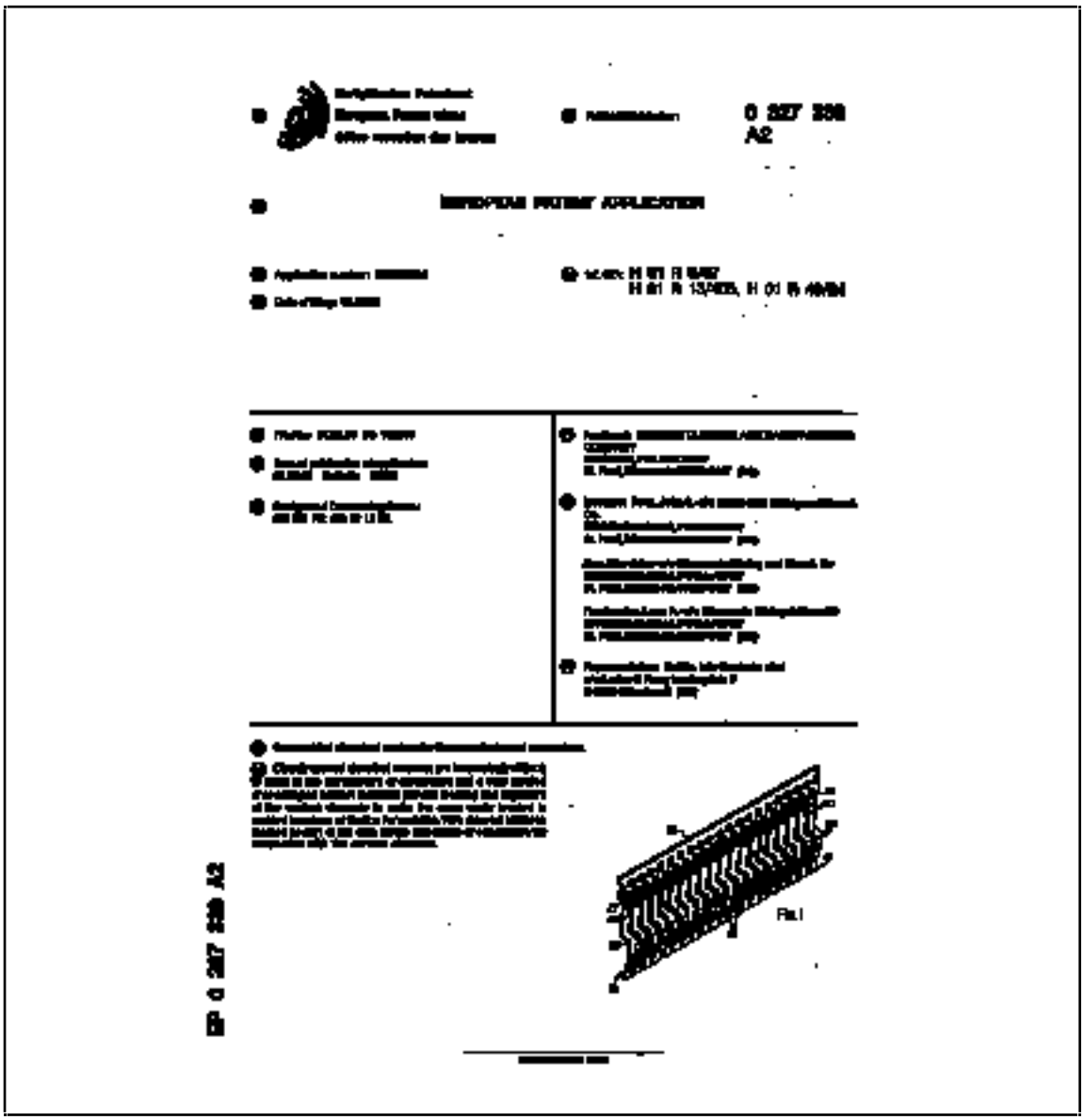

Figure XX. Exemple de page de garde.

ARCHITECTURE

Après une évaluation faite sur un échantillon représentatif de page de garde et une démonstration de la segmentation, la faisabilité technique du projet est démontrée. Le workflow relatif à la saisie et au stockage est établi comme suit:

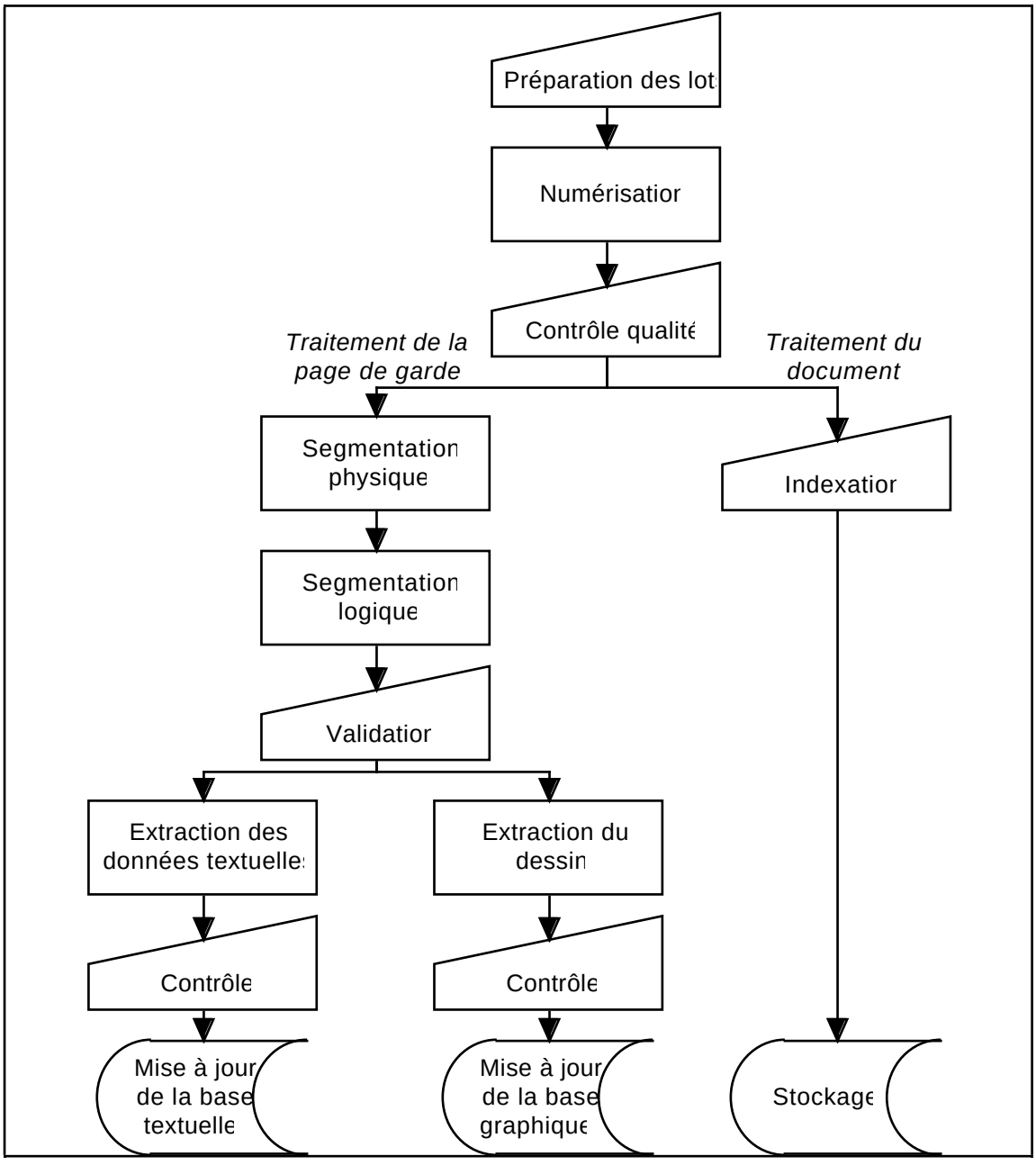

Figure XX. Synoptique Saisie/Segmentation/Stockage.

L'architecture du système étudié se décompose en un ensemble d'unités:

◊ Unité de Saisie par Numérisation (USN),

◊ Unité d'Interprétation de Documents (UID),

◊ Unité de Stockage sur média Optique (USO),

◊ etc.

Dans le cadre de ce document, seule l'Unité d'Interprétation de Documents (UID) remplissant les fonctions de segmentation et de formatage client, est étudiée.

Unité d'Interprétation de Documents

L'Unité d'Interprétation de Documents (UID) est donc le moteur de segmentation qui va permettre d'extraire les éléments textuels et graphiques de la page de garde. Elle se décompose en plusieurs modules:

◊ Module de Segmentation Physique (MSP),

◊ Module de Segmentation Logique (MSL),

◊ Module de Validation de la Segmentation (MVS),

◊ Module d'Extraction des données Bibliographiques (MEB),

◊ Module d'Extraction du Graphique (MEG).

Architecture matérielle

L'Unité d'Interprétation de Documents est abrité sur un calculateur de type IBM PC/AT et d'une architecture parallèle, relié au squelette Ethernet du système:

◊ NEC Powermate 2
- mémoire vive de 640 ko
- disque magnétique de 40 Mo
- lecteur de disquettes 5"1/4 de 1,2 Mo
- clavier • contrôleur Ethernet
- MSDOS
- TCP/IP
- NFS

◊ Ecran haute résolution 19" et controleur 4 Mo

◊ Carte de compression et de décompression

◊ Carte transputeur maître ARCHIPEL VOLVOX
- transputeur T825, 25 MHz
- mémoire DRAM de 8 Mo

◊ Carte transputeur esclave ARCHIPEL VOLVOX
- 4 trasnputers T825, 25 MHz
- 4 mémoires DRAM de 4 Mo

◊ Licence de segmentation

Notes

Codes de diffusion			
Client:			
Interne:			

MC2
4 chemin de Malacher
ZIRST
38240 Meylan
FRANCE

✆ (33) 76 90 22 00