Chapitre 6. Contenus documentaires

> *Most of the memex contents are purchased on microfilm ready for insertion. Books of all sorts, pictures, current periodicals, newspapers, are thus obtained and dropped into place. Business correspondence takes the same path. And there is provision for direct entry. On the top of the memex is a transparent platen. On this are placed longhand notes, photographs, memoranda, all sorts of things. When one is in place, the depression of a lever causes it to be photographed onto the next blank space in a section of the memex film, dry photography being employed.*
> Vannevar Bush, As we may think, 6.

Dans le chapitre précédent, nous avons mis en place un modèle à base de traces. Dans ce chapitre-ci, nous définirons un premier type de trace que nous appellerons « *contenu documentaire* ».

Le lecteur pourrait s'étonner que l'on consacre un chapitre à un type de trace dont la gestion serait *a priori* plus du domaine de l'ingénierie que de celui de la recherche. Cependant, comme ces contenus documentaires serviront de support aux types de traces que nous verrons par la suite, et qu'aucun outil du commerce, à notre connaissance, n'assure l'intégralité des fonctions proposées[55], il ne nous semble pas superflu d'en faire une présentation détaillée.

[55] Le système Transvision®, déjà cité, bien que proche de ce que l'on souhaite, n'assure qu'une partie des fonctionnalités recherchées [TVML00].

1. Notions

a. Contenu documentaire

« Qu'est ce qu'un document ? » : la question est loin d'être naïve[56]. Prenons l'exemple d'une collection scientifique en ligne (par exemple une revue). Le document se situe-t-il au niveau du paragraphe et de l'illustration ? Du fac-similé de la page ? Du tome ? Du volume ? De la collection complète ? Nous nous abstiendrons ici de faire du document une définition même semi-formelle. Sera « document » ce qu'un individu considèrera comme « document ».

Par conséquent, comme primitive de notre système, nous ne prendrons pas le document, mais tout simplement la « granule » choisie pour le stockage et nous l'appellerons « un *contenu documentaire* ».

Pour être archivé, un contenu documentaire doit être aussi stable que possible. Par conséquent chaque nouvelle version d'un contenu documentaire fera l'objet d'un nouveau contenu documentaire. De même, la clef de référence d'un contenu documentaire ne pourra être modifiée.

b. Objet documentaire

A l'usage, il apparaît très vite que la seule notion de *contenu documentaire* n'est pas suffisante.

Par exemple, lors du projet de numérisation des collections de l'EFA, chaque tome, une fois massicoté, a été placé dans un chargeur pour être numérisé recto-verso. Ainsi, chaque fac-similé de page pouvait être référencé automatiquement par un couple d'entiers : le numéro d'ordre du « codex » numérisé et le rang de la page dans ce codex. Pour référencer et feuilleter convenablement les fac-similés (par la table des matières, des figures, etc.), notre équipe a dû proposer une nomenclature comprenant le nom

[56] Cette question fait d'ailleurs l'objet actuellement d'une rédaction collective au sein du Réseau Thématique Pluridisciplinaire « Documents et contenu : création, indexation, navigation » (CNRS) : <http://archivesic.ccsd.cnrs.fr/documents/archives0/00/00/04/13>.

abrégé de la collection, le numéro de volume, le numéro de tome, le type de page (préliminaire, foliotée, finale, planche) et le folio. Ainsi, comme le montre l'exemple de la figure 6.1, le 4ème fac-similé du 231ème codex numérisé correspondait en fait au 1001ème folio du 3ème tome du 121ème volume de la revue BCH.

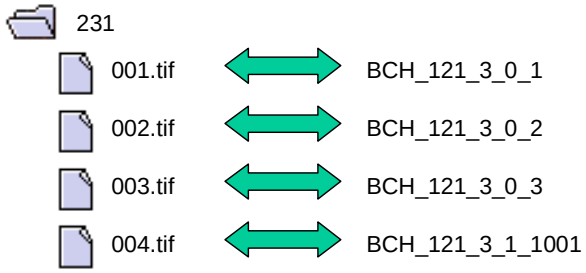

Figure 6.1 : Exemple de correspondance pour un contenu documentaire entre une référence automatique et un nom significatif.

Notons que l'obtention de cette nomenclature finale n'a été possible qu'au prix de l'abandon d'autres nomenclatures et donc au prix du changement (automatique mais long) du nom de tous les fac-similés numérisés. Si l'on refaisait aujourd'hui la numérisation, il serait sans doute préférable de distinguer pour un fac-similé sa référence automatique de son nom.

De manière plus générale, pour manipuler un contenu, il faudra lui donner un nom. Contrairement à la référence automatique, ce nom, résultat d'une interprétation, peut éventuellement être modifié. Nous nous trouvons donc en présence d'un autre niveau que nous appellerons « une *source* ».

Nous avons considéré jusqu'à maintenant les contenus documentaires comme des atomes[57], des éléments amorphes, sans structure. Or, par le seul fait de son inscription sur un support, l'élément documentaire *est* structuré. Dit autrement, l'élément documentaire, par sa structure interne, définit un ensemble de parties virtuellement adressables[58]. Par exemple, une image dans sa représentation matricielle définit virtuellement

[57] Au sens étymologique (indivisible).

[58] Nous reprenons ici la terminologie que nous avons définie (en nous inspirant entre autres de *Xanadu*) au sein du groupe de réflexion de l'ISDN sur les « documents multi-structurés ».

un très grand nombre de zones rectangulaires. La même image, dans une représentation fréquentielle, définit virtuellement un ensemble de version de l'image avec plus ou moins de détails. Lorsqu'une partie virtuellement adressable sera utilisée par un être humain, nous en garderons trace et l'appellerons « un *fragment* ».

Enfin, à la demande des utilisateurs, nous avons été amenés à définir les *notes de lecture* comme des éléments dynamiques qui à la différence des sources peuvent être modifiées sans créer de nouvelles versions. Notons que le caractère dynamique de leur contenu nous empêche de définir dessus des fragments[59].

Nous définissons la notion d'*objet documentaire* comme la généralisation des notions de « source », de « fragment » et de « note » (cf. Figure 6.2). Cet objet documentaire est caractérisé par le couple formé :

- d'un espace de nom (le numéro IP de son serveur de correspondance),
- et d'un nom, aussi significatif que possible (dans l'exemple de la Figure 6.1: « BCH_121_3_1_1001 »).

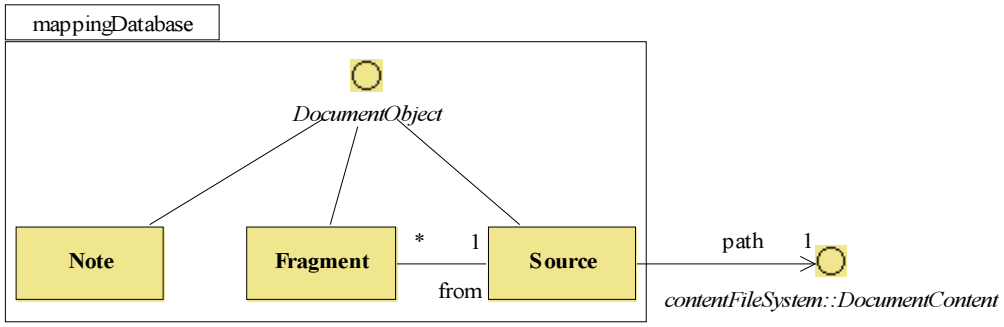

Figure 6.2 : Contenus documentaires (diagramme de classe UML)

Ce couple de valeur appelé « *localisation d'objet documentaire* » sera utilisé comme référence par les autres types de trace. On considèrera que deux traces font référence au même objet documentaire si et seulement si leur localisation d'objet documentaire est identique.

[59] En effet une étude portant sur le « balisage » de textes évolutifs mériterait sans doute une thèse à elle toute seule.

2. Traitements

a. Sur un objet documentaire isolé

Pour obtenir une objet documentaire dont on connaît la localisation (par exemple : « 134.214.105.147/BCH_121_3_1_1001 »), il faut s'adresser au serveur de correspondance de la source (« 134.214.105.147 ») en lui fournissant le nom de la source (« BCH_121_3_1_1001 »). Celui-ci nous renvoie un objet de la classe abstraite *DocumentObject*, instancié en fonction du type de serveur de contenu (ici, la version 2003 du serveur de contenu de *Porphyre*). Cet objet comporte un certain nombre de méthodes permettant entre autres d'obtenir l'URL de visualisation (en fonction d'une largeur maximale donnée) et celle de sa vignette.

La mention d'une largeur maximale permet pour des contenus documentaires de type image, archivés à très haute définition[60], d'obtenir des vues redimensionnées à la baisse en fonction des besoins[61].

Si notre objet documentaire est une image source, nous obtiendrons des URL du type :

- « http://contentserver.porphyry.org/Image/getThumbnail?file=231/4 » pour sa vignette (cf. Figure 6.3a),
- « http://contentserver.porphyry.org/Image/getSource?file=231/4&max=640 » pour la vue réduite à *640* pixels de largeur maximum (cf. Figure 6.3b).

S'il s'agit d'un fragment d'image, nous obtiendrons une URL du type :

- « http://contentserver.porphyry.org/Image/getFragment?file=231/4&coord=1000+1100+700+400&max=600 » pour la vue obtenue par extraction de la zone ayant pour coin supérieur gauche, le point de coordonnée cartésienne (*1000,1100*), pour largeur *700* et pour hauteur *400*. La vue après extraction est réduite à *640* pixels de largeur maximum (cf. Figure 6.3c).

[60] Et souvent compressés sans pertes (par exemple en TIFF).

[61] Et compressées avec pertes – par exemple en JPEG – pour plus de fluidité sur le réseau.

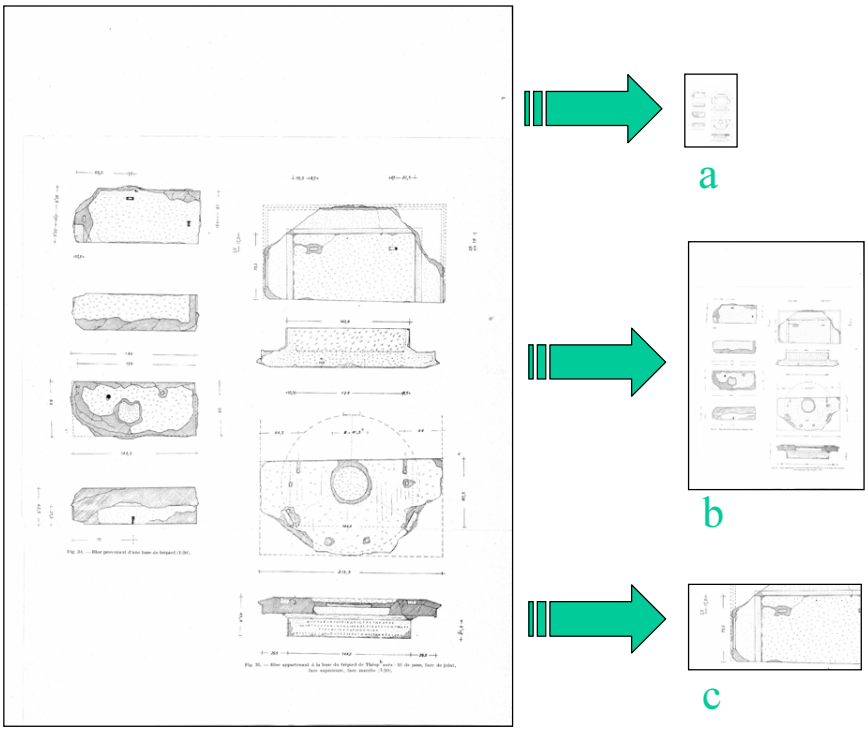

Figure 6.3 : A partir d'un même contenu documentaire : a. une vignette, b. une vue de la source, c. une vue d'un fragment.

Ces URL, et d'autres permettant de gérer le texte intégral, font appel à des scripts PHP du serveur de contenu de *Porphyre*. Nous invitons le lecteur intéressé par l'implémentation optimisée qui en a été faite de se reporter au mémoire CNAM en cours de rédaction de Régine Tribollet [Tribollet03].

b. Sur un contexte de lecture

Dans l'approche intertextuelle qui est la nôtre, l'objet documentaire ne peut se comprendre que dans un ensemble. Nous appellerons cet ensemble « un *contexte de lecture* ». Or, il serait illusoire de penser que gérer un tel contexte se ramène à mettre bout à bout plusieurs objets documentaires. Il s'agit au contraire de trouver des métaphores formelles et visuelles à la « sélection » de sens qui s'effectue selon François Rastier entre deux textes[62] lus en vis-à-vis.

[62] Au sens large (cf. chapitre 3) : texte intégral, photographie, diagramme…

La requête au serveur de correspondance ne porte donc plus sur un objet documentaire isolé mais sur un contexte de lecture. Les URL construites pour chaque objet peuvent alors tenir compte de ce contexte de lecture.

Dans la version 2003 du système *Porphyre*, un premier traitement du contexte de lecture a été mis en place. Il vise à matérialiser dans une source la relation qu'elle entretient avec ses fragments quand ils sont lus en contexte[63]. Ceci est valable aussi bien pour des contenus textuels que graphiques (cf. Figure 6.4).

L'URL du document source, avec encadrement des zones appartenant à ses fragments est alors de la forme (cf. [Tribollet03]) :

- « http://contentserver.porphyry.org/Image/getSource?file=231/10&coord=600+450+150+100;760+400+200+100&max=640 ».

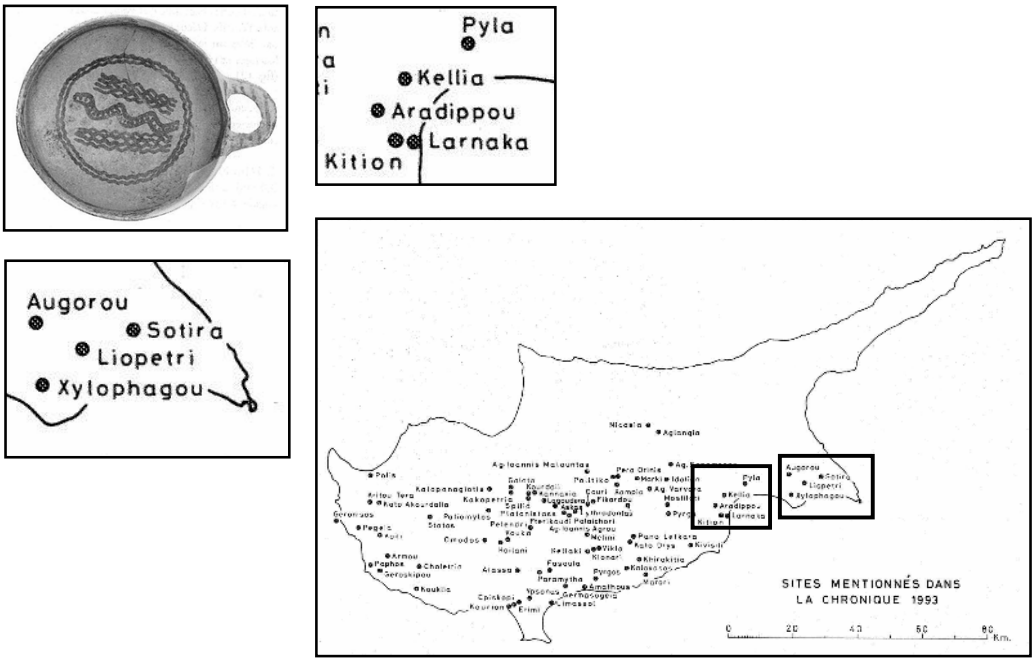

Figure 6.4 : Fonction d'encadrement automatique dans un contexte de lecture comprenant des fragments et leur source.

Nous envisageons de mettre en place par la suite d'autres traitements portant sur les contextes de lecture. Quand un de ces derniers comporte des objets textuels, il pour-

[63] Nous nous inspirons ici de la visualisation dans *Xanadu*® des liens de citation [Nelson99].

rait être intéressant, par exemple, de distinguer graphiquement les termes propres à un objet de ceux que l'on retrouve dans plusieurs. Dans le même ordre d'idée mais avec un aspect plus statistique, l'utilisation de l'incontournable « *tf.idf* »[64] permettrait de faire ressortir les termes à la fois fréquents dans un objet documentaire et rares dans le contexte de lecture.

[64] *tf.idf* (de l'anglais : « term frequency, inverse document frequency ») : Variable statistique couramment utilisée en recherche d'information pour extraire des termes présents dans le texte intégral d'un document de telle sorte qu'ils soient les plus discriminants possible par rapport au corpus.