De la diffusion à la conservation des documents numériques

Christian Rossi

De la diffusion à la conservation des documents numériques

Christian Rossi
rossi.ch@free.fr

Ce document à pour thème le problème de la conservation des documents numériques. Il se propose principalement de décrire le problème de la conservation en terme d'intermédiaires matériels et logiciels entre les données et l'utilisateur plutôt qu'en terme de format de fichier.

Il est surtout consacré aux documents textuels et aborde différents aspects du problème : matériel et logiciel, migration, formats de fichiers, typographie, XML, conversion, durée.

Un document numérique n'est pas directement accessible à un utilisateur humain

Une caractéristique importante du numérique est qu'il existe de nombreux intermédiaires entre le support où est stocké l'information numérique et l'utilisateur :

- le support
- un lecteur
- le système d'exploitation
- un logiciel applicatif
- les périphériques

- un mode d'emploi

En premier le support (disque magnétique, cdrom...) où se trouve l'information physique, ensuite un lecteur qui permet de passer de l'information physique à une information binaire, puis le système d'exploitation qui permet de regrouper ces 0 et ces 1 en fichiers et en répertoires. Pour finir un logiciel applicatif qui rend ce fichier accessible à l'être humain par l'intermédiaire d'un périphérique audio, vidéo ou d'une imprimante, et ce pour les cas les plus classiques. Sans oublier, bien sur, une personne qui sache utiliser, si ce n'est réparer, tous ces intermédiaires.

Il s'agit là une différence majeure par rapport au papier où l'accès à l'information est direct.

Pour la conservation, une conséquence importante de l'existence de ces intermédiaires est que ce n'est pas seulement le support qui doit être disponible mais l'ensemble de cette chaîne de consultation, si ce n'est l'originale une équivalente.
La longueur de cette chaîne est en tant que tel un problème mais de plus son espérance de vie, c'est-à-dire celle d'un ordinateur et de ces périphériques est plutôt courte. Actuellement, et en moyenne, il est classique de changer d'ordinateur tout les trois ans, une nouvelle version d'un système d'exploitation standard est disponible tout les ans et des patchs de sécurité doivent être appliquées tous les mois. Et ceci est nécessaire car pour du « vieux » matériels la maintenance officielle s'arrête rapidement ou si elle existe encore se monnaye à un coût exorbitant.

Assurer la conservation de documents numériques c'est donc actuellement choisir entre migrer en permanence l'ensemble des supports et des chaînes matériels et logiciels ou bien

créer des musées vivants de l'informatique où sont conservés en état de marche ces intermédiaires.

En fait on a arrive à devoir migrer non pas à cause du support qui a atteint sa limite de vie physique mais à cause de son obsolescences technico-commercial : le lecteur n'est plus vendu, plus maintenu et les caractéristiques techniques du support sont dépassées. Dans cette optique il est inutile d'utiliser un support dont l'espérance de vie est de cent ans puisque dans trois ans il faudra le migrer. L'espérance de vie technico-commercial est plus faible que l'espérance de vie physique.

Deux autres problèmes. Un qui n'est clairement pas automatisable est de savoir vers quoi il faut migrer. Les erreurs dans ce domaine peuvent créer de nombreuses migrations inutiles et coûteuses. Un autre est tout simplement de garantir que lors des migrations successives l'intégrité des documents est conservés. En effet lors d'une migration il est nécessaire de copier des fichiers d'un support sur un autre. Des outils doivent alors vérifier si au niveau binaire les fichiers sont identiques. Mais si des conversions de format des fichiers sont nécessaires un contrôle automatique semble alors impossible. Comment garantir qu'un fichier Postscript convertie en PDF va donner une sortie graphique identique ? Le fichier n'est plus le même et le logiciel de consultation non plus.

Quand à l'émulation de technologie ancienne elle rallonge et complexifie de manière importante la chaîne entre le document et l'utilisateur, c'est un intermédiaire de plus. Comme problème particulier il y a le fait que les spécifications des technologies à émuler ne sont pas obligatoirement publics. D'autre part un émulateur est un logiciel, qui comme un autre, tourne sur une certaine machine avec une certaine version d'un certain système d'exploitation. Autrement dit l'émulation nécessite des migrations de l'émulateur.

Autonomie et conservation

Quelle serait l'espérance de vie d'un document papier qu'il faudrait, tout les 3 ans, désacidifiés pour le maintenir intact ? Il existe une réponse : le temps pendant lequel il y aurait le personnel, le budget et la volonté de le faire.

La situation est fort semblable avec les documents numériques.

Comme tenu de l'évolution rapide des techniques, de l'espérance de vie du matériel, des versions logiciels, des formats de fichier il est clair que des migrations et des conversions de toutes sortes sont inévitables. Autrement dit un document électronique est très dépendant pour sa conservation d'une intervention humaine. Beaucoup plus dépendant que les documents papiers. On est loin du document historique que personne n'a touché depuis 200 ans et que l'on redécouvre fébrilement. Or d'une manière idéale, à partir d'un moment, un document doit être autonome vis-à-vis de sa conservation.

De plus une contrainte importante est que le coût de conservation soit raisonnable. Ceci n'est pas actuellement réellement compatible avec des migrations logicielles et matérielles fréquentes et avec les interventions humaines assurant ses migrations. Mais au-delà du problème du coût, compter sur des interventions humaines périodiques et fréquentes pour garantir la conservation de documents sur le long terme ne semble pas raisonnable.

Il est possible d'utiliser une image, celle de l 'arbre et de la plante verte. Un arbre dans une foret vit sa vie, les Eaux et Forêts peuvent intervenir, il peut y avoir des catastrophes qui vont abîmer ou détruire l'arbre mais l'arbre est quand même relativement autonome. Par contre une plante verte dans un appartement est totalement dépendante, plus d'arrosage et c'est terminé pour elle. Et il n'y a pas eu de catastrophe pour autant.

Facilité d'usage ou espérance de vie

Si l'on regarde l'évolution à travers le temps des supports des documents il est possible de faire deux constats :

- la facilité d'usage a augmenté
- l'espérance de vie a diminué

En effet s'il est plus facile d'utiliser une feuille de papier qu'une tablette de pierre l'espérance de vie du papier est par contre clairement plus faible. Si l'on part des tablettes argiles, en passant par la parchemin, vers le papier, cette évolution est très nette, évolution qui a d'ailleurs permis une réelle démocratisation de l'accès à l'information, mais évolution qui a aussi entraînée une réelle diminution de l'espérance de vie des documents. Il y a eu un prix a payer : l'espérance de vie.

Il est possible de résumer cette situation en disant que pour les supports classique :

facilité d'usage x espérance de vie = constante

La situation est-elle identique avec les supports numériques ?

Pour se qui est de la facilité d'usage la question ne se pose pas, elle est vraiment très élevé. Le numérique est fantastique pour créer, modifier, stocker, rechercher, diffuser...

Mais d'un autre coté le numérique est en effet complexe, fragile, instable,... et l'on manque de recul. Le nombre élevé d'intermédiaire entre le support et l'utilisateur humain ne facilite pas l'accès à l'information ni sa conservation.

La situation semble donc, pour le moment, identique. Et si nous somme réellement dans un univers où il impossible de n'avoir que des avantages, où se que l'on gagne en facilité d'usage est perdu en espérance de vie, il vaut mieux utiliser le support numérique en connaissance de cause.

Il ne s'agit pas ici d'arrêter l'utilisation des supports numériques, le voudrait-on..., mais être réaliste. Ainsi on ne demande pas à un livre de poche d'être éternel, pour autant personne ne souhaite le voir disparaître. Un livre de poche est simplement utilisé pour ses qualités non pas pour ses défauts. Avec le support numérique avoir le même comportement est raisonnable : l'utiliser pour ses qualités de création/diffusion tout en ayant conscience des ses défauts actuels de conservation.

En numérique le notion d'archive est virtuelle

Il n'y a pas de différence physique entre un document électronique dont l'équivalent papier est une archive et un dont l'équivalent papier est actif. La seul différence peut-être une métadonnée qui indique le statut du document. Autrement dit, sur le plan informatique, le bon fonctionnement d'un site web de type dépôt numérique contenant des archives demande le même travail et donc a le même coût qu'autre contenant des document actifs.

Du très court-terme au long-terme

Si la durée de conservation recherchée est spécifique à chaque type de documents et à l'utilisation prévue essayons quand même de définir une échelle des temps de la conservation des documents numériques :

- très court terme : correspond à l'espérance de vie technico-commercial de la chaîne de consultation, en cas de problème le bon fonctionnement est assuré grâce à un service de maintenance ;
- court terme : espérance de vie physique, ça marche tant que ça marche... ;
- moyen terme : l'accès au données est assuré grâce à la mise en place d'une organisation qui assure des migrations ou autre ;
- long terme : à partir du moment où cette organisation a disparu, on en revient à l'espérance de vie physique de la chaîne de consultation résultant de la dernière migration.

Une fois que l'on a pris conscience que les données numériques ne sont pas autonomes pour leur conservation et qu'il nécessaire de les prendre en charge (stockage hyper-actif) il est possible de garantir la pérennité sur le moyen-terme. Cela nécessite une organisation fiable, des moyens humains et financiers importants. Une telle organisation doit assurer la collecte des données, garantir la conservation proprement dit et gérer leurs accès. Bien sur il pourra y avoir des pertes, des oublis de migration, aucune organisation n'est parfaite.
Mais cela ne donne pas d'indication sur la durée possible de ce moyen-terme ni ne résout le problème pour le long-terme. En fait si la conservation sur le moyen-terme est réaliste l'archivage sur le long-terme pose vraiment problème. Que restera-t-il aux historiens ?

La suite du document se concentre sur la chaîne logicielle.

Des formats sources aux formats visualisables

Si l'on part d'un ficher il faut donc des intermédiaires, notamment une chaîne logicielle pour qu'un document puisse être lu.

Voici différents types de formats utilisés pour les documents textuels :

- format source (LaTeX, RTF, TEI, HTML)
- format visualisable vectoriel (PS, PDF)
- format visualisable bitmap

Et voici les étapes possibles pour passer d'un fichier source à des informations compréhensible à un être humain :

ordinateur : fichier source → (pdflatex / word) fichier vectoriel → (xpdf / acrobat) données bitmap : périphérique : humain

Un fichier source va contenir du texte et des informations exprimées en utilisant un langage donné : informations typographiques (justifier, caractère italique...), des informations de structure (titre, auteur, chapitre...) ou un même un mélange des deux. C'est le type de fichier créé par un auteur et donc orienté vers l'édition du document.

Une fois traité par des outils tel que word ou pdflatex l'on obtient en sortie un fichier visualisable vectoriel (du type PostScript ou PDF). C'est un fichier qui contient du texte et des informations de positionnement : placer la lettre "x" sur la page à tel endroit et la dessiner en utilisant du Times. C'est une description géométrique de la page. Dessine moi un cercle de rayon un au milieu d'une page de format A4. Bien sur pour dessiner une lettre les équations sont plus compliqués que pour un cercle mais le principe est là.

Des outils tel que xpdf, ghostscript, acroread, l'interpréteur postscript d'une imprimantes sont des exemples de logiciel qui peuvent lire ces fichiers PostScript ou PDF et qui savent passer de cette description géométrique à une image sur une écran ou sur une imprimante. C'est-à-dire à une image bitmap identique à celle que l'on peut obtenir avec un scanner. Un fichier visualisable bitmap qui est compris par les périphériques : le pixel est allumé ou non, de l'encre doit être déposée à tel point sur la feuille. L'on est passé d'un fichier à une image qu'un humain peut voir.

Avec les outils wysiwyg comme Word ou OpenOffice ces étapes sont invisibles et hybridées car ils intègrent toutes ces fonctions dans un même logiciels.

Sur le plan logiciel toutes ses étapes n'ont pas le même niveau de complexité.

Afficher un fichier bitmap à l'écran demande un logiciel simple. L'arbitraire entre le contenu du fichier et le résultat final est inexistant : avec un 0 le pixel reste noir, avec un 1 il devient blanc. Tous le complexe travaille de mise en page a déjà été réalisé. Mais il y a un inconvénient : le texte, sous forme d'une codage (ASCII, ISO-8859...) a disparu.

Quant aux logiciels qui traitent des donnés vectorielles ils ont un niveau de complexité intermédiaire. Ils partent par exemple d'un fichier PDF pour obtenir des données bitmap et les envoyer à l'écran.
Comme un format vectoriel est une description mathématique d'une page sur le plan logiciel la liberté d'interprétation est limité : tracer un cercle sur page n'est pas ambigu. Mais elle plus grande qu'avec un fichier bitmap : il faut traiter la pixellisation, les concepteurs ont pu laisser des zone d'ombre dans les spécifications. Les formats vectorielles sont souvent complexes, consulter la description du format PostScript ou PDF suffit pour s'en convaincre, et les erreurs ou les difficultés de programmation plus fréquentes. Comme avantage le texte est souvent présent, le fichier occupe moins de place qu'un fichier bitmap, il est possible de zoomer. Là aussi mise en page a déjà été faite par un autre logiciel.

Un logiciel qui sait traiter un document source est beaucoup plus complexe. Son but est obtenir une description de page en PDF ou autre. C'est avec ces logiciels que le lien entre le contenu du fichier et le résultat final est le plus ténu, ou la composante artistique est la plus grande et la liberté du programmeur la plus élevée. En fin de compte un fichier source ne contient que très peu d'information par rapport un résultat attendu, c'est au logiciel de faire la différence.

En fait le problème vient du fait que pour ces formats les commandes sont de très haut niveau. Que signifie une balise <h1> ? Le lien entre une balise et la sortie graphique est totalement arbitraire. Aussi il n'y a pas deux navigateurs qui donnent le même résultat. Cela se constate sur les navigateurs avec les mises en page complexe, les tables, les CSS...
Que signifie une commande du type « justifier » ? Derrière cette commande il faut bien un logiciel qui implémente un algorithme plus ou moins efficace pour effectuer la justification de paragraphe, les césures. Et un même fichier RTF (ou un autre format source) lu par OpenOffice et par Word ne donnera pas le même document car l'algorithme de justification n'est pas le même.

La majorité du travail de mise en page, travail graphique est apportée par le logiciel et non pas par le fichier source qui ne contient que du texte et des commandes.

Donc trois types de formats différents et trois types de logiciels plus ou moins complexes. Les fichiers aux formats sources se situent au début de la chaîne logicielle et donc demande le traitement le plus complexe. Les fichiers bitmap se situe en fin de cette chaîne. Aussi dans l'optique de la conservation le format bitmap présente un intérêt puisque la chaîne logiciel se limite à un seul logiciel, de plus ce logiciel est simple, simple à écrire ou à récrire. Bien sur par rapport à un fichier au format PDF l'on perd en fonctionnalités possibles (lien hypertexte, zoom, recherche plein-texte...). Mais c'est le prix à payer : en se rapprochant de l'humain on s'éloigne de la machine et de ses potentialités. Et s'il est possible d'imprimer un document numérique textuel il faut savoir que l'existence d'une version analogique n'est que toujours possible ou même utile avec les données numériques (bases de données par exemple). En terme de compromis, entre la complexité et les fonctionnalités du logiciel, PS ou PDF semblent être des formats intéressants.

En généralisant une hypothèse raisonnable est de dire qu'un document numérique a une espérance de vie d'autant plus grande que les logiciels nécessaires à son accès sont simples et que la chaîne logicielle est courte Ceci est également vrai pour les intermédiaires de type matériel.

Structure et typographie

Une tendance actuelle est de limiter un document textuel à sa structure. Ceci peut être vu comme la disparition d'une culture typographique et est souvent associé au développement d'XML. Pourtant dans des domaines comme la physique, les mathématiques ou l'informatique les auteurs peuvent encore apporter une grande attention à la mise en page et au respect des règles typographiques. Ce sont principalement des univers où l'on utilise (La)TeX.
Peut-on limiter un document à son texte et à sa structure exprimée sous forme de commande ? Un document c'est du fond et de la forme, autrement dit c'est aussi du graphisme et respecter l'œuvre d'un auteur c'est respecter ces deux aspects. De même ce que le lecteur souhaite c'est un vrai document lisible, non pas un fichier. Et la typographie existe justement pour améliorer

la lisibilité. Quand à la qualité graphique d'un document elle joue un rôle important dans le plaisir que l'on peut avoir à sa lecture. Se concentrer sur l'aspect structure fait souvent oublier les aspects graphiques et avec, l'importance du logiciel qui contrôle ce graphisme.

Bien sur dans l'optique de la conservation il est légitime de se demander quel est l'apport de chacune de ses composantes. Que faut-il conserver ?

En fait le problème vient du fait que pour la conservation les documents textuels ont un aspect polymorphe (on ne parlera pas ici du contexte du document) :

document = texte + structure + images

Un document textuel c'est du texte représenté en utilisant un codage (ASCII, ISO-8859, Unicode) dans un fichier source et sous forme graphique dans un fichier visualisable. Des images incluses dans le document. C'est une structure qui peut être exprimée dans les fichiers sources sous forme de commandes associées au texte ou directement sous forme graphique dans un document visualisable grâce la mise en page et la typographie.

logiciel
texte + structure + images => graphisme

lecteur
graphisme => texte + structure + images

Prenons un exemple les célèbres RFC (les Request For Comment), les documents qui décrivent les standards de l'internet. Depuis leur création en 1969 ils sont disponibles sous forme de simple fichier ASCII sur le site suivant http://www.rfc-editor.org/. Bien sur avec de l'ASCII les graphiques sont très simples, pas de multilinguisme, pas de math, de molécule, de musique... mais cela marche et cela suffit puisque l'Internet existe.

En fait si l'on ne s'intéresse pas à la typographie et que le document n'a pas de composante graphique un fichier texte suffit pour sauvegarder l'information. Il peut être lu avec un éditeur de texte, logiciel plus simple qu'un traitement de texte.

Il faut savoir que des outils permettent d'extraire le texte d'un document TeX, RTF ou PDF (comme detex, rtf2ascii et pdftotext) Ce sont d'ailleurs ces outils qu'utilisent les indexeurs plein-texte pour traiter les fichiers non HTML.

Inversement si la composante graphique est importante ou si l'on souhaite conserver la mise en page de l'auteur à partir d'un fichier source une chaîne logiciel complexe devient nécessaire pour accéder au document tel que l'auteur l'a créé. Il faut alors utiliser un logiciel compatible avec celui de l'auteur de préférence le même et la même version.

D'une manière générale la question se pose. Faut-il garantir l'intégrité absolu d'un enregistrement numérique ou peut-on le modifier ainsi que sa chaîne de consultation et ne conserver que les aspects considérés comme essentiels ?

XML format pérenne, et le logiciel ?

Que signifie que le format XML est pérenne et est-ce vraiment le cas ?

Avec un format comme HTML les balises sont définies une fois pour toute par un organisme, le W3C, et elles ne sont pas extensibles en fonction des besoins d'un utilisateur. Au contraire XML permet à chacun de créer son propre jeux de balise. OpenOffice, TEI, WordML, XHTML sont des exemple de jeux de balise dans le domaine du document. Il est donc possible de créer des jeux de balises XML décrivant la structure d'un document (comme la TEI, Text Encoding Initiative) mais aussi des balises qui représentent la typographie ou la mise en page (par exemple XHTML, WordML, OpenOffice ou XSL-FO).
En fait XML définit les règles que l'on doit respecter lorsque l'on désire créer ses propres balises (à chaque balise ouvrante doit être associée une balise fermante, pas de balises entremêlée…). A chaque jeu de balise (dont les caractéristiques sont définis par ce que l'on appelle une DTD ou un schéma) il est nécessaire s'associer un logiciel qui sait traiter ses balises.

Donc il n'y a pas vraiment de format XML mais il y a des formats respectant les règles définit par XML, chacun va après utiliser une DTD particulière. En fait parler de format XML est une simplification de langage.

Ceci dit que signifie alors qu'XML est pérenne ? S'il s'agit de dire que la recommandation « Extensible Markup Language (XML) 1.0 » du W3C est pérenne pourquoi pas : les règles que l'on doit respecter lorsque l'on crée ses propres balises pour être conforme avec la version 1.0 de la recommandation XML sont pérennes.

Si par contre cela signifie que quelque soit la DTD utilisée, du moment que ce format est conforme à XML c'est pérenne, là c'est discutable. En effet comme il a été vu précédemment entre le fichier et un résultat accessible à l'utilisateur il y a toujours un logiciel. Et rien ne garantie la pérennité de ce logiciel.

Un format basé sur XML peut ne pas être ni plus ni moins pérenne qu'un autre non basé sur XML, en fait on ne peut vraiment pas parler d'un format pérenne, c'est le logiciel qui est ou non pérenne. Et un logiciel l'est rarement.

Dans 10 ans pourra-t-on lire un document utilisant le format XML d'OpenOffice, de la TEI ?
Oui, si un logiciel existe qui sait traiter se format, en fait le problème est le même que pour RTF ou (La)TeX.

Il est vrai cependant que connaître les spécifications d'un format sont un plus et est même nécessaire. Par exemple les spécifications du format Word ne sont pas connues. Le lecteur potentiel est donc prisonnier ce celui qui connaît le format et sait écrire le logiciel correspondant. Ceci dit une spécification de format ne remplace pas un logiciel qui marche que se soit à cause des spécifications qui peuvent être incomplètes et le nouveau logiciel non totalement compatible avec l'original soit que les ressources nécessaires pour le redévelopper sont trop importante. De plus rien n'oblige le créateur d'un jeu de balise XML à les rendre public. Ou même le schéma ou la DTD peuvent être inexacte par rapport au logiciel disponible.

Il est vrai qu'XML a un avantage certain, si l'on respecte la recommandation sur le fond, c'est de ne pas être un format binaire. Il y est en effet indiqué que les documents XML devraient être lisibles par l'homme et raisonnablement clairs. L'obligation d'utiliser pour le texte un codage standard tel qu'Unicode est une chose importante.

Ceci dit XML est pour la conservation un formalisme parmi d'autre, même si les buts de ses concepteurs sont positifs et louables et qu'il plutôt bien placé. Le problème, là où les choses sont vraiment complexes, c'est au niveau logiciel, et il est peu réaliste de croire que l'on peut résoudre tout les problèmes de pérennité grâce à une question de formalisme.

L'utilisation XML est souvent associé, d'une manière très positive, à la découverte de la séparation possible entre structure et présentation ainsi qu'au problèmes que posent les formats propriétaires. Mais par un excès d'évangélisme on a arrive à faire oublier qu'un fichier, même basé sur XML, a toujours besoin d'une chaîne matérielle et logicielle pour pouvoir être consulter.

Il est même possible de rajouter que cette liberté de créer ses propre balises tout en se protégeant derrière la conformité XML peut poser problèmes. Par exemple dans le domaine du son et des partitions musicales, musique et XML, il existe actuellement pas moins de 16 propositions de balisages différentes : MusicXML, MusiXML, MusicML... (voir par exemple une liste sur http://xml.coverpages.org/xmlMusic.html). On passe de l'espéranto à la tour de Babel.

De LaTeX vers HTML

Lorsque la conversion de format est évoqué des problèmes se posent souvent. Comme des outils permettant de convertir du LaTeX vers de l'HTML existent, voyons la situation. Voici d'un manière simplifiée le fonctionnement des ces outils.

En fait il existe deux cas : il y a ce qui est simple à traiter, le texte et ce qui est compliqué, le reste. Soit il existe une correspondance entre d'une part une commande LaTeX et d'autre part une commande (X)HTML qu'un navigateur sait visualiser (titre, gras) et le logiciel effectue la conversion. Soit il n'existe pas de correspondance. Dans ce cas le programme de conversion lance LaTeX pour générer une image GIF ou PNG qui sera insérée sous forme d'un lien dans le document (X)HTML.

Il y a quelques années les équations mathématiques étaient convertit sous forme d'image mais aujourd'hui les outils récents les convertissent en MathML car les navigateurs commencent à traiter ce format. Mais des extensions de LaTeX savent aussi représenter des molécules de chimie ou des partitions de musique. Et là les convertisseurs vers XHTML génèrent toujours des images.

Ce type de conversion pose deux problèmes. Le premier vient de la richesse des fonctionnalités de LaTeX alors que le format (X)HTML sait faire moins de chose. L'on passe souvent d'une information structurée à des images.

D'autre part la mise en page des fichiers PostScript ou PDF générés par LaTeX est connu pour sa qualité alors que pour (X)HTML elle dépend du navigateur utilisée.

LaTeX n'utilise pas un format source binaire, c'est une bonne chose. Mais l'intérêt de LaTeX ce n'est pas son format. En effet il n'existe pas une grande différence entre `\title{Mon titre}` et `<title>Mon titre</title>`. L'intérêt de LaTeX c'est la richesse du logiciel. Et changer de format c'est aussi changer de logiciel.

Logiciels et développeurs

Les formats PostScript ou PDF sont très utilisés et ce non sans raisons. Cependant il existe en fait très peu de logiciel permettant la consultation de ces fichiers, les plus connus sont :

- pour PostScript : l'interpréteur d'Adobe, ghostscript
- pour PDF : acrobat, ghostscript et xpdf

Et peu de personne participe à la réalisation de ces logiciels. Par exemple pour ghostscript il y a 19 personnes et pour xpdf 1 personnes (avec une quinzaine de contributeurs).

Les savoirs opérationnels sont concentrés dans un nombre de personne très réduit et ce non pas pour de raisons de monopole mais parce que cela n'intéresse personne. Il y a un effet pyramide : beaucoup d'utilisateurs mais en face peu de logiciel et peu de créateurs. Pour aujourd'hui cela ne pose pas de problème, ces logiciels existent, ils marchent et sont maintenus, mais sur le long terme cela peut devenir un véritable problème.

En guise de conclusion

La conservation des documents numériques sur le long terme est un problème ouvert. Le manque de recul est certain. Les formats dont les spécifications ne sont pas connus posent problème néanmoins les questions de formalisme ne vont pas résoudre tout le problèmes. Il fallait rappeler ici l'importance pour la conservation des documents numériques des aspects matériel et logiciel.

Aujourd'hui il ne faut pas trop penser en terme de format pérenne. Les formats ne sont pas pérennes, les logiciels non plus et le matériel encore moins. Mais il faut penser en terme de migration de fichiers, migration du matériel et en terme de conversion de formats. Les difficultés de conservation sont en fait intrinsèque à cette même technique qui permet tant de prodige pour la création ou la diffusion.

Quel est l'avenir ? En fait il est très difficile de deviner les évolutions, les miracles possibles. Certains aspects qui nous semble aujourd'hui inquiétants ne le seront plus demain, non pas que les problèmes auront été résolu mais simplement ne seront plus d'actualités ou bien auront été contournés.

Bibliographie

Bulletin des Archives de France sur l'archivage à long terme des documents électroniques
http://www.archivesdefrance.culture.gouv.fr/fr/publications/DAFbulelectronique.html

Notamment :

Jean-Luc Philip. Le point de vue d'un généalogiste sur la conservation des documents électroniques. Aix-en-Provence (n° 6, juillet 2001)
http://www.archivesdefrance.culture.gouv.fr/fr/publications/dafbuln%B06.html#chapitre9

Claude Huc. La pérennité des documents électroniques - points de vue alarmistes ou réalistes ? Centre National d'Ėtudes Spatiales, Toulouse (n° 7, octobre 2001)
http://www.archivesdefrance.culture.gouv.fr/fr/publications/dafbuln%B07.html#chapitre1

Catherine Dhérent. Les archives électroniques. Manuel pratique. Direction des Archives de France, Paris (2002)
http://www.archivesdefrance.culture.gouv.fr/fr/archivistique/DAFmanuel%20version%207.html

ATICA. Guide de conservation des informations et des documents numériques (2002)
http://www.adae.gouv.fr/spip/article.php3?id_article=7&var_recherche=conservation+documents+num%25E9riques

Stratégie globale pour la conservation à long terme des documents électroniques en Suisse. Association des Archivistes Suisses (2002)
http://www.staluzern.ch/vsa/ag_earchiv/home_f.html

Digital Preservation Testbed. Archives Nationales des Pays-Bas
http://www.digitaleduurzaamheid.nl/index.cfm?paginakeuze=185&categorie=2

Notamment :

Migration: Context and Current Status (2001)
http://www.digitaleduurzaamheid.nl/bibliotheek/docs/Migration.pdf

Emulation: Context and Current Status (2003)
http://www.digitaleduurzaamheid.nl/bibliotheek/docs/white_paper_emulatie_EN.pdf