Techniques de Numérisation

Cours licence BDAN – IUT2 Grenoble – 2009-2010

Jean-Michel Mermet
Chargé de Mission Documentation
DSI de Grenoble Universités
Tél : 04 76 63 59 43
Mél : Jean-Michel.Mermet@grenet.fr

A. Introduction

1. Présentation et périmètre du cours

La numérisation est la conversion d'un signal analogique en un signal numérique en fonction de deux paramètres : l'échantillonnage et la quantification. C'est conceptuellement une modélisation mathématique de la réalité.

La numérisation concerne aussi bien les images, les textes, que le son et les vidéos pour couvrir le champ des projets en maturation de nos jours.

Le cours a pour but de vous faire parcourir et réfléchir sur toutes les étapes d'un projet de numérisation, depuis l'idée initiale et l'intérêt d'y recourir jusqu'à la production des données numériques.

La numérisation, en effet, ne s'arrête pas, loin de là, à l'étape technique qui consiste à faire passer par exemple une photo argentique sous forme de fichier image :

- Il y a, en amont toute la réflexion de ce que l'on veut numériser, pourquoi on veut le faire, comment on va le faire, dans quels formats, pour quels publics, etc...

- Il y a, en aval, toute la réflexion sur le mode de mise à disposition des fichiers numériques, leur identification, leur authentification, leur formats de diffusion, leur préservation pour le futur.

Il y a donc bien une véritable chaîne numérique à concevoir, mettre en place, faire fonctionner. Nous verrons que tout projet de numérisation implique une réflexion sérieuse sur le long terme.

a. Périmètre du cours

Ce cours ne couvrira que très rapidement les notions suivantes :

- les aspects juridiques ;
- la gestion des assets numériques par les systèmes de gestion de contenu (CMS) traitées par ailleurs par Jérémie Grépilloux ;
- la théorie des métadonnées traitée par Elizabeth Cherhal ;
- les traitements graphiques, sonores et vidéos à appliquer aux fichiers obtenus ; nous n'aborderons que ce qui est nécessaire pour représenter correctement les documents analogiques sous forme numérique.
- Les aspects d'archivage des données numériques, qui sont traitées dans un nouveau cours.
- La publication des collections sur internet, objet de la suite de ce cours au second semestre

2. L'importance de la numérisation

a. Préambule

« Selon une enquête récente du Réseau canadien d'information sur le patrimoine à propos de l'utilisation de la technologie dans les musées, 94 % des institutions souhaitent faire de la numérisation leur première priorité en matière de technologie pour l'année 2006. La numérisation […] est désormais une activité importante pour les musées. Étant donné que les connections Internet haute vitesse deviennent de plus en plus abordables et accessibles, il est normal que les musées publient de plus en plus sur leur site des images de leurs artefacts. »[1]

Les documents numériques issus d'une numérisation servent à :

- l'impression,
- à la documentation,
- à la recherche et
- à la publication en ligne,
- et le plus souvent pour la gestion des collections, la préparation de catalogues et la promotion d'expositions ou d'autres activités.

Pourquoi la numérisation est-elle devenue si populaire ?[2]

1 http://www.chin.gc.ca/Francais/Contenu_Numerique/lere_de_la_numerisation.html (visité le 27/08/2009)
2 http://www.chin.gc.ca/Francais/Contenu_Numerique/ Guide_Gestionnaires/introduction.html (visité le 24/7/2006)

- Elle permet de rendre les **collections beaucoup plus accessibles** (collections autrefois trop lointaines, objets trop fragiles pour être manipulés ou exposés)

- Elle permet de mettre en valeur **des aspects particuliers** de l'histoire locale ou d'atteindre un public national ou international.

- Elle permet de rassembler, **à des fins de comparaison et de recherche**, des objets ou des collections dispersés en plusieurs lieux

- Elle peut devenir un **puissant outil d'enseignement**.

- La numérisation peut également contribuer à la gestion des collections en procurant à tout le personnel une **meilleure information sur le contenu des collections**.

- Le simple fait de sélectionner des objets à numériser fournit au personnel une occasion supplémentaire d'évaluer et de consigner l'état des objets. La numérisation incite en outre à **améliorer la documentation**, en particulier lorsque des enregistrements d'accompagnement sont destinés à être rendus public en même temps que les images numérisées : le besoin de normes et d'une meilleure information devient rapidement évident.

- Les objets numérisées sont utilisées dans une **vaste gamme de produits de diffusion** comme les sites Internet, le matériel de promotion, de nouveaux articles pour la boutique souvenir d'un musée, etc.

- **La numérisation contribue aux stratégies de conservation** puisque, une fois numérisés, les originaux analogiques fragiles sont davantage à l'abri des manipulations et des aggressions.

- Comme la technologie numérique permet de faire une recherche dans un grand nombre d'enregistrements, de modifier et de traiter des images et des textes et de rassembler des éléments disparates selon des modalités nouvelles, elle peut se révéler un **outil souple et précieux dans tous les secteurs d'un musée**.

- La facilité d'exécution des tâches mentionnées ci-dessus favorise également **une collaboration plus étroite avec d'autres établissements.**

b. Pourquoi numériser ?

- **Plus de détérioration des supports en consultation –** Une fois numérisés, les documents originaux peuvent être conservés dans des conditions optimales, sans le danger que représente leur manipulation. On augmente donc leur sécurité.

- **Reproduction** - il est possible de réaliser des copies de l'information déjà numérisée en utilisant soit le même format de stockage soit un autre format numérique, sans perte de qualité.

- **Automatisation** - sont automatisés non seulement la restitution des éléments demandés par les chercheurs, grâce à l'utilisation de systèmes de stockage robotisés, mais aussi le processus de reproduction. Le document étant représenté par une chaîne de chiffres binaires, il est possible d'automatiser la copie.

- **Recherche** - la numérisation offre la possibilité d'effectuer des recherches dans les catalogues aussi bien locaux que lointains et de créer un tissu de liens ou de pointeurs renvoyant du document consulté par l'utilisateur à des documents connexes de la même collection ou d'autres. Des recherches de texte intégral peuvent en outre être effectuées. Des techniques de recherche analogues sont actuellement élaborées pour les documents sonores et les images (recherche et reconnaissance de phrases musicales, de rythmes, de motifs, de formes, ...)

- **Accessibilité** – « à distance », elle tient à la capacité d'envoyer un signal numérique sur les réseaux de télécommunications sans perte de qualité. « temporelle » la collection numérisée est accessible 24 heures sur 24 avec un minimum de personnel. Finalement, l'accessibilité est grandement améliorée pour des publics déficients visuels ou moteurs. Selon Francis Pisani[3] : « L'intérêt primordial de la numérisation des livres est la mise du savoir à la disposition de populations qui sans cela n'y auraient pas accès, notamment dans le sud ».

- **Rapidité de la copie** - dans le domaine numérique, il est possible de recopier fidèlement ou de transférer des données à une vitesse très élevée. Le futur transfert des collections sur de nouveaux supports sera beaucoup plus rapide que le passage initial à la numérisation.

- **Qualité** - elle tient à la possibilité de numériser un document avec une très forte résolution, selon les besoins. Il est également très facile de réaliser des copies de qualité inférieure à partir de la copie de haute qualité chaque fois que nécessaire.

3 http://pisani.blog.lemonde.fr/pisani/2006/06/plein_de_bonnes.html, (visité le 18/08/2007)

- **Espace nécessaire** - la haute densité du stockage de l'information sur les supports numériques peut se traduire par une réduction majeure de l'espace de rayonnage nécessaire. D'où, également, une réduction de l'espace nécessitant une régulation climatique conforme aux normes archivistiques et, en conséquence, un abaissement des frais de fonctionnement. Par ailleurs, un signal numérique peut être fortement compressé avec ou sans perte d'information. Une compression avec perte d'information peut s'effectuer sans perte d'information « sensible », c'est-à-dire détectable. Une information numérique est facilement stockable sur des supports dont la capacité augmente et dont le coût et la taille diminuent rapidement au fil des années.

- **Futurs transferts de la collection** - si une collection est copiée sur un support analogique, le coût de ses futurs transferts sera identique à celui de cette première reproduction. Si elle est copiée sur un support numérique, on pourra, lors des futurs transferts, tirer parti des possibilités de recopie automatisée inhérentes au numérique. Le transfert de collections de données numériques n'est pas un phénomène nouveau. Les milieux bancaires, par exemple, ont transféré plusieurs fois avec succès leurs banques de données. Leur expérience peut offrir des enseignements utiles pour le transfert de l'information textuelle, visuelle et sonore.

- **Incitation à entreprendre des recherches** – les chercheurs seront d'autant plus enclins à exploiter un fonds qu'il sera numérisé, avec tous les avantages que cela lui procure. La numérisation facilite et rend plus efficace la recherche par les conservateurs, les étudiants, les enseignants, les érudits, les chargés de cours à l'université, les chercheurs et les spécialistes, car elle permet d'étudier des documents disparates dans des contextes nouveaux. Il y a davantage d'exploration de ressources liées aux objets exposés, et l'information à propos du musée et de ses collections importantes s'en trouve améliorée. Les images numérisées peuvent être utiles aux professionnels et chercheurs de musées du monde entier qui explorent les liens, les ressemblances et les différences avec les travaux d'autres établissements.

- **Constitue une copie de sauvegarde/ de remplacement** – Le document numérique constitue une copie de sauvegarde de l'original. La numérisation offre une stratégie de remplacement pour des objets, par exemple des films et des archives sonores, en voie de détérioration.

- **Amélioration de la lisibilité des documents** : le traitement d'images peut améliorer la lisibilité de documents défraîchis, tachés voire carrément illisibles.

- La transmission numérique est plus **résistante** que l'analogique aux **défauts de transmission** puisqu'il suffit de reconnaître, à la réception du message, sa présence et non sa forme pour le reconstituer ; la puissance nécessaire pour la transmission est donc plus faible et les équipements de réception souvent plus petits. On peut également vérifier la bonne transmission par des techniques de chiffrages telles que les checksums.

- Par ailleurs, le numérique permet de **transmettre tous les signaux de la même manière**, quelle que soit la nature de l'information (voix, données, images...); les équipements de transmission sont donc les mêmes pour le téléphone ou la télévision par exemple.

- Le volume ou le débit d'un signal transmis numériquement peut également être fortement réduit par **compression**, sans perte (loseless) ou sans perte sensible de la qualité, en supprimant toutes les informations inutiles (redondantes ou imperceptibles par les sens humains). Dans une transmission vidéo par exemple, au lieu de traiter numériquement 24 images par seconde, on ne traite que la différence entre deux images consécutives; cette différence étant la plupart du temps très faible, plus de 98% de l'information peut être laissée de côté sans perte de qualité des images (format MPEG-4).

Conclusion technique : La numérisation apporte ainsi de nombreux avantages, facilitant le traitement et le stockage des informations et offrant une qualité de transmission incomparable. Ces qualités doublées de son indépendance par rapport à la nature de l'information transmise expliquent la généralisation de l'emploi des technologies numériques aux dépens des analogiques.

Un billet du blog Figoblog[4] détaille un ensemble de raisons pour lesquelles on peut vouloir numériser en bibliothèque. L'auteur distingue plusieurs raisons :

- **la valorisation d'un fonds** : Avantage : c'est joli, ludique, attrayant, ça donne une bonne image de la bibliothèque et ça plaît aussi aux gens qui ne sont pas spécialistes. Inconvénient : ce genre d'interface est inutilisable par des personnes qui s'intéressent au même document à d'autres fins.

- **la numérisation "à la demande" ou spécialisée** : Avantage : le public est déjà ciblé et on répond précisément à ses attentes donc le succès est plus facilement assuré, au moins auprès d'un nombre limité de personnes. Inconvénient : c'est toujours inutilisable par des personnes qui s'intéressent au même document à d'autres fins (typiquement, c'est bien de ne numériser que des enluminures mais celui qui travaille sur le texte du manuscrit se retrouve le bec dans l'eau). Ou alors cela ne couvre qu'un spectre documentaire/thématique très limité.

- **la numérisation de sauvegarde** : Avantage : une grande facilité de consultation par rapport à l'ancien support de substitution, qui n'offrait que des capacités limitées de lecture simultanée et une "expérience de lecture" peu

4 http://www.figoblog.org/document1637.php (visité le 19/07/2007)

optimisée. Inconvénient : pas de public assuré pour consulter cette numérisation, et comme les originaux sont en voie de disparition, il faut qu'elle soit fiable, authentique et pérenne car c'est bientôt (ou déjà) le seul moyen d'accéder à ces documents là.

- **la bibliothèque numérique** : Avantage : c'est un service cohérent avec une politique documentaire, des missions, des services, etc. capable en principe de répondre aux besoins d'un public diversifié. Inconvénient : c'est très compliqué et coûteux à organiser. Même très très compliqué.

c. Inconvénients, limites et risques de la numérisation

I. LE COÛT

C'est un des reproches souvent fait aux projets de numérisation. Les points soulevés ci-après[5] sont parfois judicieux, parfois exagérés. Ils méritent en tout cas qu'on sache y répondre le cas échéant.

- **Investissement initial** - le matériel nécessaire pour effectuer la numérisation peut être coûteux et demande souvent à être utilisé par des opérateurs qualifiés pour donner les meilleurs résultats.

- **Contraintes de rangement** - on croit souvent que les supports utilisés pour stocker l'information numérisée devront être conservés dans un environnement très propre et très stable sur le plan climatique, ce qui accroîtra la consommation d'énergie imputable à la collection. L'aménagement de cet espace de rangement à l'atmosphère stabilisée exigerait en outre un investissement initial.

- **Frais de fonctionnement** - on craint qu'une collection numérisée ne doive elle aussi être fréquemment recopiée, avec les coûts de main-d'oeuvre, d'énergie et d'achat de nouveaux supports que cela implique et que la survie de l'information numérique ne soit pas garantie au-delà de deux à trois ans si elle est stockée sur bande magnétique et de trois à cinq ans si elle est conservée sur disque optique.

- **Coûts de préparation** - un document doit être préparé (parfois de façon destructrice, comme le massicotage des ouvrages avant leur passage par le scanner) pour la saisie numérique. Outre l'éventuelle préparation physique requise, le contenu initial doit en être classé et indexé et les références textuelles introduites dans la base de données par du personnel spécialisé, opérations qui peuvent être coûteuses. On pense que la copie de l'information des supports existants sur de nouveaux supports implique d'importantes contraintes de main-d'oeuvre.

II. LE RISQUE IMPORTANT DE MAUVAIS CHOIX TECHNOLOGIQUES

- **Lors de la numérisation**. Des erreurs notamment lors du choix de l'échantillonnage ou du format de fichier (formats propriétaires) peuvent empêcher une exploitation future des données.

- **Lors de la gestion des documents numériques**. Des erreurs lors du stockage (supports, conditions de stockage) et lors de la préservation des documents peuvent conduire à des pertes sévères de données (cas des données sur les vols lunaires de la NASA). Ne jamais oublier que les débuts de l'ère informatique ont été marqués par la plus grande perte de données de l'histoire de l'humanité.

III. LA PERTE DE REPÈRES PHYSIQUES

Une partie de l'information d'un document est contenue dans le support physique. La numérisation ne « retient » donc pas tout : l'odeur, le poids, la texture ne sont pas reproduits comme le démontre ce compte-rendu d'expérience concernant le « petit Cartulaire »[6] :

> « Le traitement de l'image connaît aussi ses limites, qui sont celles de la source. Comme plusieurs microfilms ont beaucoup vieilli, il serait souhaitable que la numérisation à haute résolution et en couleur soit faite directement sur les manuscrits à l'aide d'un appareil-photo numérique. De plus, le déchiffrement des écritures cursives, bien que facilité par la possibilité d'agrandir la taille de l'image à souhait, n'est guère simplifié. Les changements dans les teintes d'encre demeurent également imperceptibles. Le repérage de mots dans la source est aussi impossible si ceux-ci n'ont pas fait l'objet préalable d'une transcription. Aucune machine à ce jour ne peut donc remplacer l'oeil du spécialiste. »

5 Tiré de http://www.unesco.org/webworld/mdm/administ/fr/MOW_finD.html#3 (visité le 19/07/2007)
6 http://lemo.irht.cnrs.fr/40/mo40-15.htm#P1374_210298 (visité le 19/07/2007)

IV. UNE MOINS GRANDE FACILITÉ D'ASSIMILATION DES CONTENUS

Une moins grande facilité d'assimilation des contenus par rapport aux média analogiques du fait de la non-linéarité de la lecture dans les documents multimédia. Ce point est discuté, mais il faut peut-être le considérer dans la problématique plus générale des limites actuelles des interfaces d'accès à l'information.

d. Quelques grands programmes de numérisation

I. GOOGLE BOOKS

L'extraordinaire projet de Google[7], qui a fait couler beaucoup d'encre ... électronique ! Pour suivre le dossier mouvementé de ce projet, on peut lire avec intérêt le dossier «l'Atelier » de Jean de Chambure[8].

Depuis ce dossier, la partie semble bien mal engagée par Europanea... Départ de Jeaneney, et les Vaudois qui, les premiers européens, rompent le front du refus et signent avec Google. Cf l'article « Les Vaudois vendent leur patrimoine écrit à Google »[9] :

> Quelque 100000 ouvrages, tous libres de droits, du XVIIe au XIXe siècle, seront digitalisés à Lausanne selon un calendrier défini dans un mois. Cette numérisation sera presque entièrement financée par Google, ce qui représente environ 12,5millions de francs. La BCU ne payera que le traitement des fichiers pour la consultation et l'engagement d'une personne qui suivra l'aventure jusqu'au bout. Mais la différence avec les projets publics de bibliothèque sur Internet, c'est que les fichiers de livres numérisés deviennent possession de Google.
>
> La décision de la BCU est inédite: jusqu'ici, toutes les bibliothèques francophones, qu'elles soient de France, de Belgique, de Suisse ou du Canada s'étaient rangées derrière les projets publics concurrents à l'offensive Google, c'est-à-dire la Bibliothèque numérique francophone et Europeana, le noyau embryonnaire de bibliothèque numérique européenne lancée en mars dernier par la France, la Hongrie et le Portugal.
>
> En concluant un partenariat avec Google, la BCU ouvre une brèche et manifeste tout haut l'impatience que d'autres ressentent face à la lenteur des projets de numérisations publics: « Le partenariat avec Google s'annonce autrement plus sérieux que le projet européen », déclare Hubert Villard, directeur de la Bibliothèque cantonale et universitaire vaudoise.

Voir aussi, en juillet 2008, la décision de la ville de Lyon de faire appel à Google pour des projets de numérisation[10]

> L'objectif de la BM de Lyon est de numériser 500000 ouvrages sur les 1 350 000 de son fond ancien, dans un délai de 10 ans. Il s'agit de documents antérieurs au XXè siècle et libres de droits. Google devra les numériser à la fois en mode image et en mode texte, dans leur intégralité. L'internaute pourra alors avoir accès aux informations par le mode classique de recherche par page de Google mais pourra également télécharger les œuvres intégralement. En échange, Google devient propriétaire de ces fichiers numériques et dispose d'une exclusivité commerciale pour leur exploitation durant 25 ans.

Voir enfin les rumeurs (fondées … infondées) de collaboration entre la BNF … et Google. Qui l'eut crû !

II. GALLICA

Serveur de consultation à distance des collections numérisées de la Bibliothèque Nationale de France[11]. Les fonds numérisés constituent une bibliothèque patrimoniale et encyclopédique, avec des ouvrages numérisés en mode image, et en mode texte, des images fixes, des documents sonores (fonds du domaine public). Ces documents sont imprimables et téléchargeables par le lecteur, dans le cadre d'un usage strictement privé.

Gallica offre (au 3/09/2009)[12] :

Documents moissonnés :

- bibliothèques partenaires : 5 834
- partenaires commerciaux : 12 133
- Total : 17 967

Documents de la BnF

- Imprimés

7 http://books.google.com/ (visité le 19/07/2007)
8 http://www.atelier.fr/type/bataille,livre,internet-30041-Dossier.html (visité le 19/07/2007)
9 Le temps.ch du 16 mai 2007
10 http://libelyon.blogs.liberation.fr/info/2008/07/la-bibliothque.html (visité le 25/07/2007)
11 http://gallica.bnf.fr/ (visité le 19/07/2007)
12 http://gallica.bnf.fr/content?lang=fr#stats

    - 124 776 monographies, dont 69 801 consultables en mode texte
    - 3 751 titres de périodiques, représentant 572 380 fascicules dont 238 905 en mode texte
- Documents iconographiques : 38 494 lots, représentant 111 644 images
- Cartes et plans : 5 009 documents
- Documents sonores : 1 056 documents
- Documents manuscrits : 4 164 documents
- Musique notées : 2 127 documents

Le taux mensuel de consultation des documents a passé la barre du million en mars 2006.

III. PROJET NUMDAM DE LA CELLULE MATHDOC À GRENOBLE

Numérisation de Documents Anciens Mathématiques[13] : dans le but de soutenir les revues de mathématiques, le programme NUMDAM met en place un libre accès aux données bibliographiques et au texte des articles qui y sont parus. Pour chaque revue concernée, la totalité des volumes publiés jusqu'en l'an 2000 a été convertie au format numérique, ce qui représente actuellement plus de 560 000 pages numérisées et 27 000 articles mis en ligne. Les articles eux-mêmes sont disponibles pour consultation en ligne à l'issue d'un délai (créneau mobile) pendant lequel ils sont réservés aux seuls abonnés. Il est possible de rechercher directement un article par nom d'auteur, mots du titre ou mots clés présents dans le texte. Il est également possible de feuilleter les sommaires de l'ensemble des volumes. Voir les collections sur la page dédiée du site[14].

IV. INTERNET ARCHIVE

« Internet Archive[15] » est une organisation à but non commercial fondée dans le but d'être la « bibliothèque d'internet ». Son but est de s'adresser aux chercheurs, aux historiens, aux étudiants, aux personnes présentant un handicap et au grand public pour leur offrir des collections historiques qui existent en format numérique. Fondée en 1996, puis refinancée en 1999, cette organisation connu une grande croissance et inclut maintenant des collections conséquentes. Elle propose des textes, de l'audio, des films, des logiciels ainsi que des pages web archivées. Elle travaille à fournir des services plus spécialisés liés à l'enseignement et à la formation et à l'accès aux collections par des personnes présentant un handicap.

V. LE GRAMOPHONE VIRTUEL

Enregistrements historiques canadiens, site Web multimédia en pleine croissance consacré aux débuts de l'enregistrement sonore au Canada. Doté d'une base de données d'images et d'enregistrements audionumériques canadiens, ainsi que de biographies de musiciens et d'un résumé de l'histoire de la musique et de l'enregistrement sonore au Canada, Le Gramophone virtuel offre aux chercheurs et aux mordus de musique un aperçu détaillé de l'époque des 78 tours au Canada.[16]

VI. EXEMPLES D'INTERFACES DE CONSULTATION

- L'extraordinaire « Cité de Dieu » par la Bibliothèque municipale de Nantes[17], à voir pour le réalisme du tourner de pages et pour le zoom très puissant.

- Online Gallery[18], projet de la British Library, ce magnifique exemple pousse le détail très loin : c'est à la souris qu'on tourne les pages de superbes manuscrits.

- Deux exemples d'utilisation d'une loupe magique (retranscrivant le texte issu d'écritures manuscrites) : le Journal de Martha Ballard[19] et Bill of sale for slave named Kate[20].

- Multi-touch screen interface demonstration[21] :

- Fluidbook[22] : une interface spécialisée dans la présentation de catalogues.

13 http://www.numdam.org (visité le 19/07/2007)
14 http://www.numdam.org/spip.php?rubrique4 (visité le 28/8/2009)
15 http://www.archive.org (visité le 19/07/2007)
16 http://www.collectionscanada.ca/gramophone/index-f.html (visité le 19/07/2007)
17 http://arkhenum.picturelan.com/bm_nantes_oeb/ (visité le 19/07/2007)
18 http://www.bl.uk/onlinegallery/ttp/ttpbooks.html (visité le 19/07/2007)
19 http://dohistory.org/diary/exercises/lens/ (visité le 19/07/2007)
20 http://memorialhall.mass.edu/activities/media.jsp?itemid=7797&img=0 (visité le 19/07/2007)
21 http://www.youtube.com/watch?v=89sz8ExZndc (visité le 19/07/2007)
22 http://www.fluidbook.com/demo/fr/ (visité le 28/08/2009)

VII. LISTES DE BIBLIOTHÈQUES NUMÉRIQUES

D'après l'excellent article[23] du blog Figoblog :

- List of digital library projects[24]
- Liste internationale du blog NetBib[25]
- The British Columbia International Digital Library[26], une liste de listes
- A selection of web accessible collections[27] (Harvard University Library)

3. La numérisation ?

a. *Quelques définitions*

I. UN SIGNAL ANALOGIQUE

Un **signal analogique** est un signal qui reproduit à l'analogue (qui transpose) un phénomène physique, tel qu'une onde mécanique (pour le son), une onde électromagnétique (pour l'image). C'est une fonction continue dans le temps ou dans l'espace.

Remarquez que les signaux analogiques ne sont pas forcément perçus par nos sens : on peut numériser par exemple le chant des baleines dont les fréquences ne se situent pas dans l'intervalle de celles perçues par l'homme (20 à 20 kHz environ).

Exemple : analyse d'un signal analogique audio (musique classique).

On analyse le signal audio reproduit par un système analogique (radio FM, platine disque 33 tours, ...) et on représente ici la puissance instantanée du signal en fonction du temps, tout d'abord en vue générale, puis en vue agrandie pour « voir » à quoi ressemble le signal.

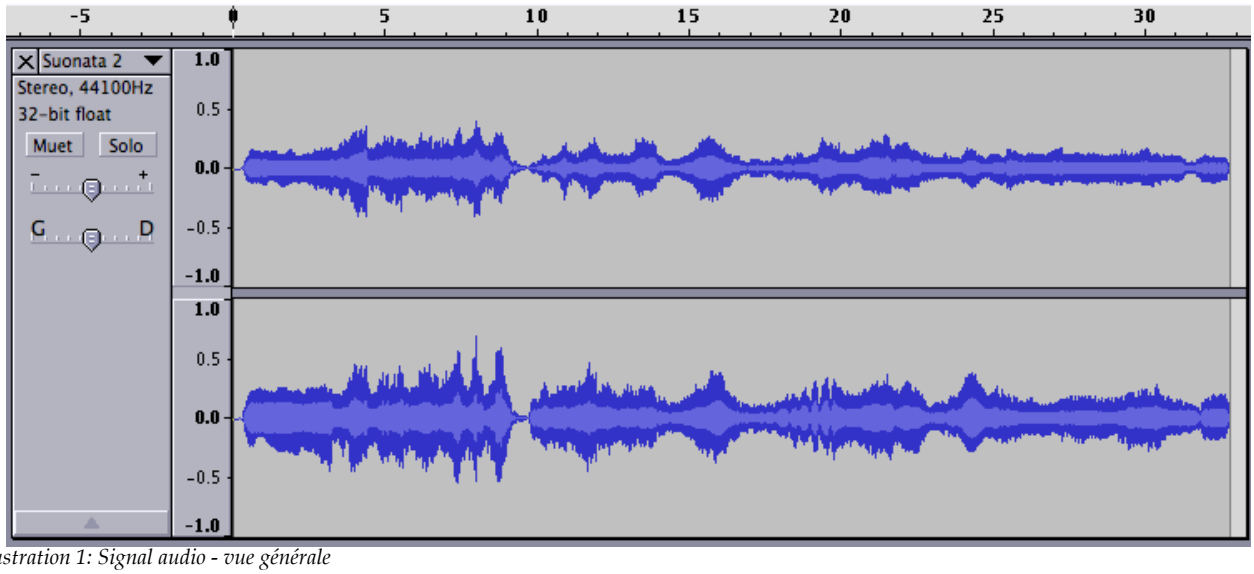


Illustration 1: Signal audio - vue générale

Puis on zomme sur une partie du signal ...

23 http://www.figoblog.org/document933.php
24 http://en.wikipedia.org/wiki/List_of_digital_library_projects (visité le 19/07/2007)
25 http://wiki.netbib.de/coma/DigiMisc (visité le 19/07/2007)
26 http://bcdlib.tc.ca/ (visité le 19/07/2007)
27 http://digitalcollections.harvard.edu/ (visité le 19/07/2007)

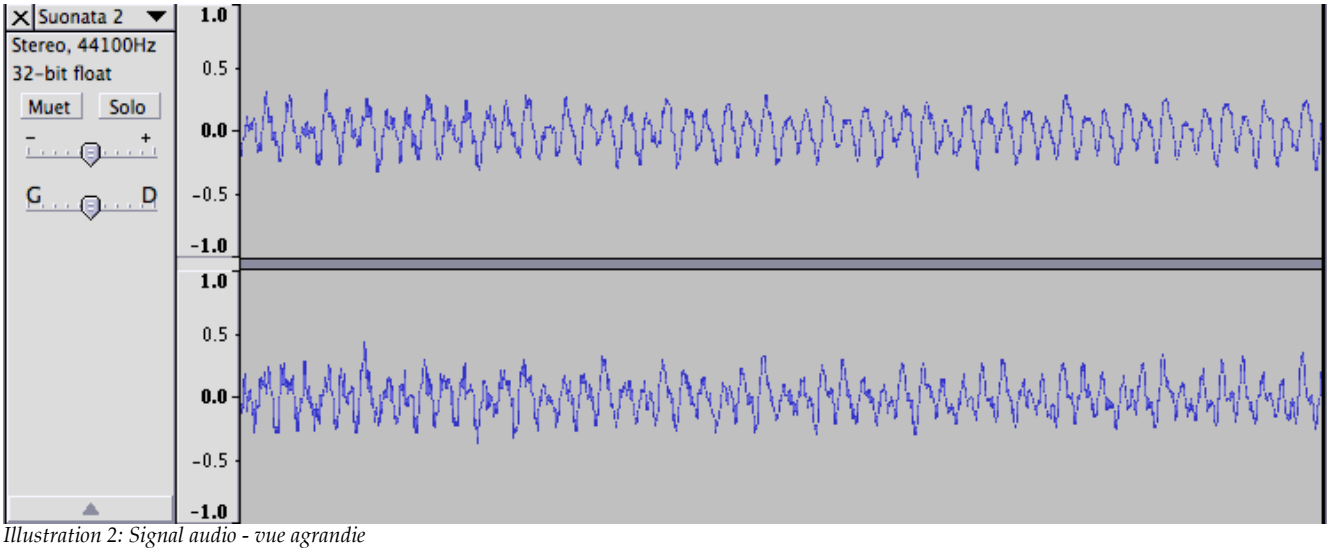


Illustration 2: Signal audio - vue agrandie

II. LA NUMÉRISATION

La **numérisation** est la conversion d'un signal analogique en un signal numérique en fonction de deux paramètres : l'échantillonnage et la quantification[28]. C'est conceptuellement une modélisation mathématique de la réalité.

Le terme anglais est *digitization,* d'où vient le « franglais » *digitalisation* qui est à proscrire. La numérisation d'une image est parfois effectuée à l'aide d'un scanner. On parle alors, toujours improprement d'un « scan », d'un « scannage », etc... A proscrire également !

III. L'ÉCHANTILLONNAGE

L'échantillonnage consiste à remplacer une fonction continue dans le temps ou dans l'espace par la suite des valeurs qu'elle prend en des instants ou des zones discret(e)s périodiques. Ces valeurs suffiront pour reconstituer la fonction dans une étape ultérieure. L'image classique est celle du stroboscope, qui permet de « décomposer » les mouvements (en fait de les échantillonner). La mesure effectuée à un instant/lieu donné est appelée un échantillon.

Prenons un autre exemple : imaginons qu'on veuille numériser l'image d'un parterre de fleurs. On commence par quadriller l'image de façon suffisamment fine, de manière à ce que dans chaque petit carré on ne trouve qu'une couleur. Chaque carré est appelé échantillon.

IV. LA QUANTIFICATION

Il faut maintenant décider de la façon de mémoriser la couleur dans chaque carré. On se définit un ensemble de couleurs précises permises, par exemple : {0➞rouge, 1➞bleu, 2➞vert, 3➞jaune}.

En fonction du lieu de mesure, on relève la couleur. Si cette couleur se rapproche d'un vert, on choisit la valeur « vert », si la couleur est plus proche d'un jaune, on choisit « jaune ». On a le choix ici entre 4 valeurs, pas plus, pas moins. Une fois tous les échantillons analysés et un choix de valeur effectué, on dispose donc d'une numérisation de l'image du parterre de fleurs ... en quatre couleurs. Le résultat sera sans doute médiocre, comme vous pouvez l'imaginer !

Calculons la place nécessaire pour enregistrer cette information. On dispose de 4 valeurs données, et l'information se code en binaire en informatique (base 2, seuls les chiffres 0 et 1 sont autorisés). Il faudra donc un nombre binaire à deux chiffres pour coder le choix :

- 0➞rouge codé 00
- 1➞bleu codé 01
- 2➞vert codé 10
- 3➞jaune codé 11

Ce nombre binaire à deux chiffres, exprimé en *bit*, permet de coder la valeur permise en un échantillon. L'enregistrement de l'information issue de la numérisation de ce parterre de fleurs nécessite donc 2 bits par échantillon. La taille du fichier issu de la numérisation est donc de 2 bits x nombre d'échantillons.

28 On considère ici le processus Pulse-code modulation (PCM). Cf http://en.wikipedia.org/wiki/Pulse-code_modulation pour plus d'explications. D'autres processus de numérisation existent, plus complexes, et ne seront pas examinés dans le cadre de ce cours.

On comprend immédiatement dans cet exemple que plus le nombre de valeurs permises est important, plus fidèle est l'enregistrement du signal. On comprend aussi que plus l'on choisit de valeurs permises, plus il faudra de place pour stocker, dans chaque échantillon, la valeur choisie. Si maintenant on choisit un nombre de valeurs permises plus important (exemple : 1024, codé en 10 bits), la taille finale de la numérisation sera 5 fois plus importante, et le résultat obtenu beaucoup plus fidèle.

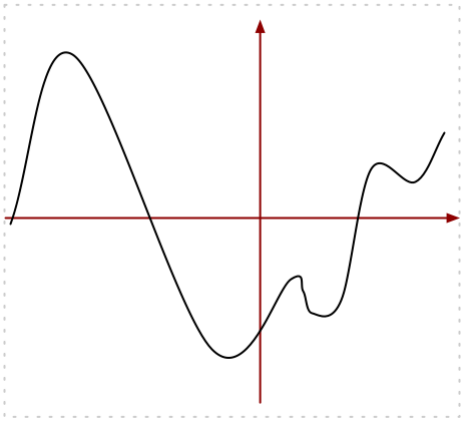
La **quantification** est l'opération par laquelle on examine l'échantillon mesuré et l'on choisit la valeur la plus proche à mémoriser parmi un ensemble prédéfini des valeurs permises.

V. UN DOCUMENT NUMÉRIQUE

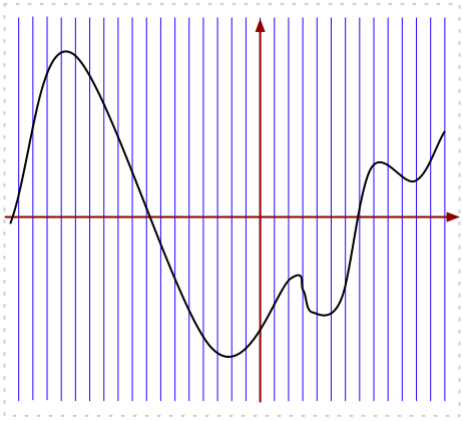
Un **document numérique** est un fichier informatique (et donc représenté à la base par une suite de 0 et de 1) dont le contenu, structuré selon les spécifications d'un format de fichier, représente une information compréhensible par un humain et/ou par un ordinateur.

b. Un Exemple

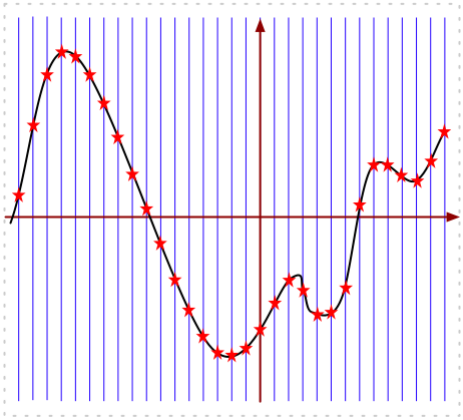
Prenons la numérisation d'un signal audio (agrandi depuis les illustrations précédentes). En ordonnée on indique la puissance instantanée du signal, en abscisse le temps. Le signal se présente ainsi :



on échantillonne ce signal à une fréquence donnée : c'est un découpage temporel.



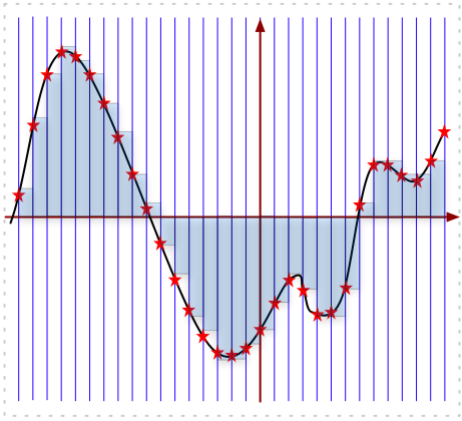
... et on mesure la valeur du signal à chaque découpe.



À chaque intersection, on prend la valeur en ordonnée. On obtient donc une série de valeurs comme dans le tableau suivant (données complètement fictives) dans lequel on mesure à chaque milliseconde une valeur (VM) (par exemple électrique).

T (ms)	0	1	2	3	4	5	6	7	8	9	10	11	12	...
VM	763	783	874	885	910	921	911	917	903	901	902	904	876	...

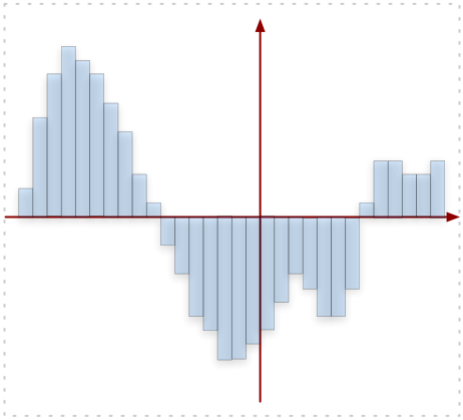
Chaque valeur mesurée est ensuite ramenée à la valeur autorisée la plus proche.



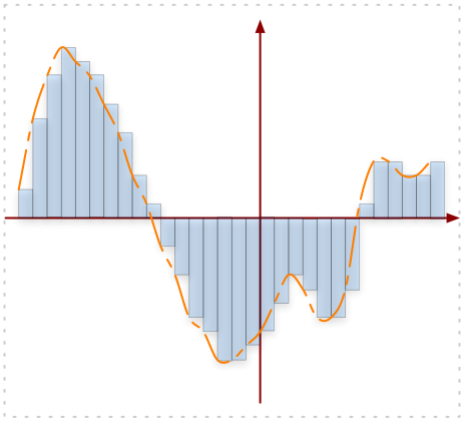
Dans le tableau suivant, (rappel, en données fictives) on compare les valeurs mesurées (VM) avec les valeurs autorisées et on choisit les valeurs autorisées les plus proches (VC) (ici les multiples de 10). En vert sont représentées les valeurs qui ne changent pas (ou peu) par cette opération, en rouge celles qui changent beaucoup.

T (ms)	0	1	2	3	4	5	6	7	8	9	10	11	12	...
VM	763	783	874	885	910	921	911	917	903	901	902	904	876	...
VC	760	780	870	890	910	920	910	920	900	900	900	900	880	**...**
diff.	**3**	**3**	**4**	**5**	**0**	**1**	**1**	**3**	**3**	**1**	**2**	**4**	**4**	

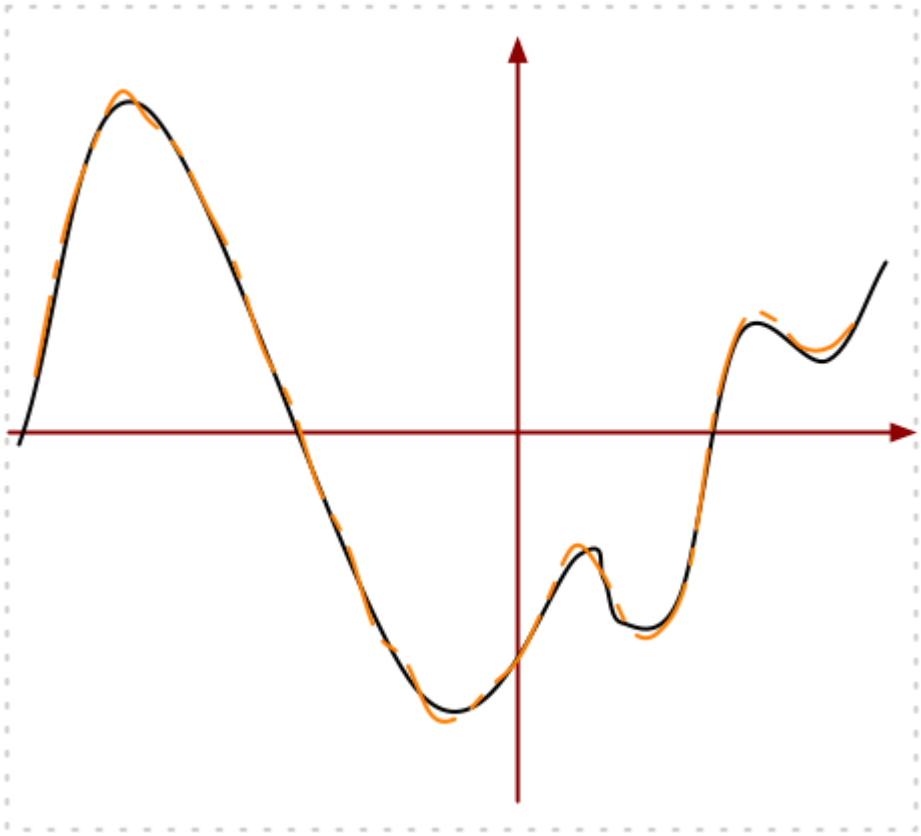
Les valeurs autorisées sont celles qui sont sauvegardées dans le fichier numérique. C'est la modélisation numérique du signal analogique initial. Un fichier numérique en résulte.



Effectuons maintenant l'opération inverse, celle de la reproduction de ce signal numérisé. Comme le dispositif humain de perception des sons est analogique, il faut reconvertir ce signal numérisé en signal analogique. La suite de valeurs numériques est donc convertie en un nouveau signal analogique :



... qu'on peut comparer avec le signal analogique initial pour faire apparaître les approximations et les erreurs dues à la numérisation :



Les facteurs qui influent sur la qualité de la modélisation de ce signal sont de deux ordres : Le fréquence d'échantillonnage et la précision de la quantification.

I. La fréquence d'échantillonnage.

Elle s'exprime en Hertz, l'inverse de la seconde. Plus celle-ci est élevée (plus la période est faible), meilleure est la prise en compte des événements rapides, et meilleure est la fidélité du signal numérique par rapport au signal analogique. Mais comment choisir cette fréquence ?

En pratique, on applique le théorème de Shannon (certains disent qu'il s'agit de celui de Nyquist) qui précise que la fréquence minimale d'échantillonnage d'un signal doit être au moins le double de la fréquence maximale contenue dans le signal à échantillonner.

Soit F_{ech} la fréquence d'échantillonnage, et F_{max_signal} la fréquence maximale du signal à numériser, ce théorème stipule que :

$$F_{ech} \geq 2 * F_{max\_signal}$$

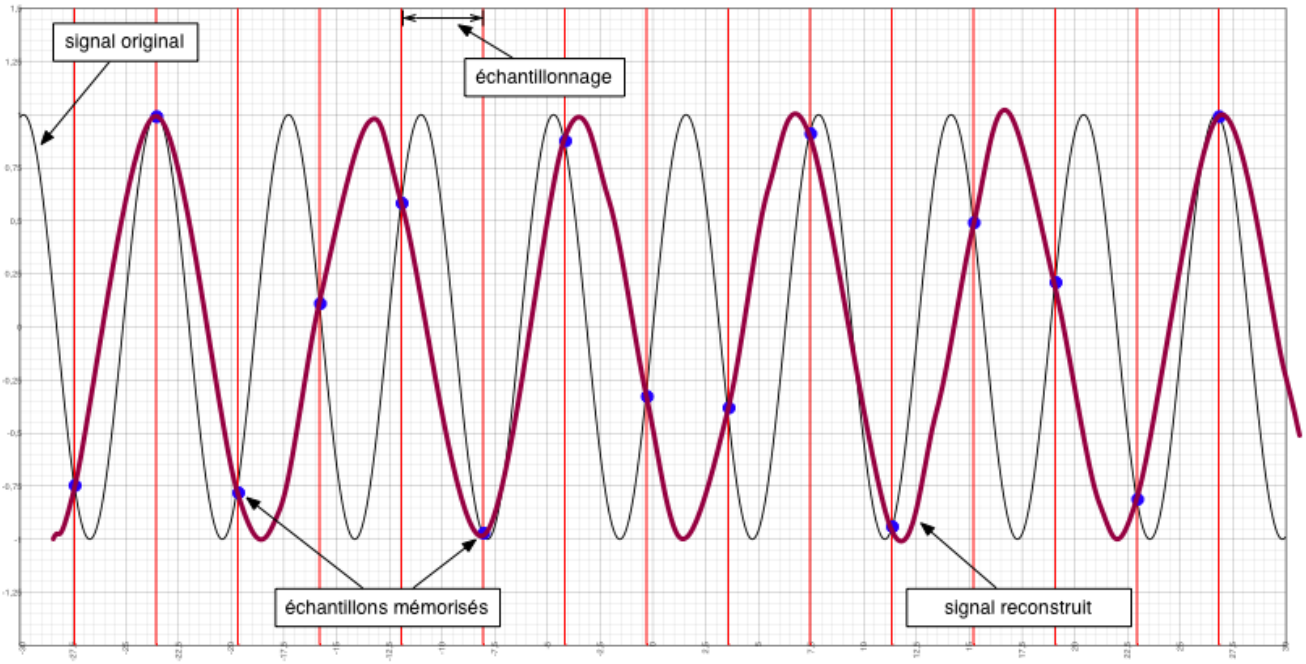
Autre formulation : Pour ne pas perdre d'information dans un signal la distance entre deux échantillons doit être inférieure à la demi-période du signal.

Exemple pour la musique, la fréquence maximale audible est de 20 kHz, en comptant très large. La fréquence d'échantillonnage des CD-audio, de 44,1 kHz, respecte bien ce théorème.

Application à la voix en téléphonie : fréquence maximale : 3700 Hz. Quelle fréquence d'échantillonnage minimale choisir ?

II. Le sous-échantillonnage

On parle de sous-échantillonnage si la fréquence d'échantillonnage n'est pas suffisante pour représenter de façon adéquate le signal. Le résultat issu d'une reconversion en analogique n'a alors plus rien à voir avec le signal de départ. Pour une numérisation audio, cela peut se traduire par de forts échos, des repliements de bandes, etc. très audibles.



Dans le schéma ci-dessus, le signal d'origine est numérisé avec un échantillonnage trop faible. Le signal qui est reconstruit à partir de la numérisation est notoirement différent du signal original. À l'inverse, un sur-échantillonnage consiste à prendre trop d'information, beaucoup plus qu'il n'en faut pour représenter de façon adéquate le signal analogique. Il n'en résulte pas forcément une meilleure qualité sonore détectable, mais en tout cas une quantité de données beaucoup plus importante à traiter.

Ce théorème peut être appliqué aux images : pour ne pas perdre de détails dans une image, la taille des pixels doit être moins de (ou égale à) la moitié du plus petit détail de l'image.

III. La précision de la quantification.

Elle doit être adaptée au signal numérisé, c'est-à-dire que la valeur analogique maximale du signal à numériser doit être codée par la valeur numérique maximale, idem pour les valeurs minimales.

Plus elle comprend de valeurs différentes, plus le codage sera précis, ... mais plus l'information sera volumineuse à stocker.

Exemple : le signal audio d'un CD-audio est codé sur 16 bits sur chaque voie, soit $2\times2^{16}=2\times65536$ valeurs à chaque échantillon stéréophonique.

Exercice : calculer la taille non compressée d'un morceau de musique de 3 minutes codé sur un CD-audio :

- 3x60 = 180 secondes
- Chaque seconde, le signal est codée 44100 fois sur 2 fois 16 bits.
- La taille du morceau est donc de : 180x44100x2x2 soit environ 30 Mo !

Exercice 2 : combien peut-on placer de minutes de musique sur un CD-audio ?