

SCI6135
HIVER 2016
INDEXATION DE COLLECTIONS NUMÉRIQUES

CONTEXTE HISTORIQUE, TECHNOLOGIQUE ET SOCIAL

1

PLAN DU COURS

- Aspects historiques,
- aspects technologiques,
- aspects sociaux,
- autres aspects

... qui ont modelé la création de collections numériques et les besoins ainsi que les possibilités pour leur organisation ou leur description

ASPECTS HISTORIQUES

- Développement de la technologie informatique
 - coûts décroissants du matériel
 - sophistication du logiciel
 - progrès technologique effréné
 - avènement du web
 - auto-publication
 - aisance
 - création de nouveaux types de collections numériques
- Point charnière : la démocratisation du web (début/milieu des années 1990)
 - auparavant : certaines collections développées et distribuées sur format de stockage (bandes, disquettes, CD-ROM)

ASPECTS HISTORIQUES (SUITE)

- Influences qui donneront naissance aux bibliothèques numériques
 - Bush (1945)
 - proposition de solution pour alléger la surcharge informationnelle
 - machine imaginaire (Memex) pour stocker les documents d'un utilisateur et les consulter rapidement
 - Engelbart – première implémentation de ce qui deviendra l'hypertexte (que Nelson baptisera au milieu des années 1960)
 - Berners-Lee : projet hypertexte global (WWW)
- Donc : avec les progrès de l'informatique, premiers essais dans l'organisation et la manipulation de grandes quantités d'information

ASPECTS HISTORIQUES (SUITE)

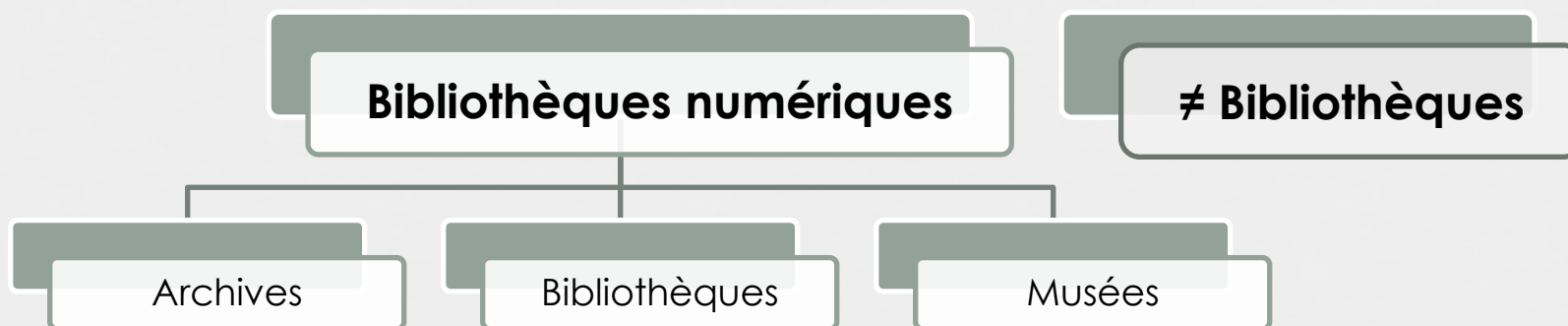
- Collections numériques : pré-WWW
 - documentation technique
 - ex. secteur de l'aviation, appareils électriques et électroniques
 - articles scientifiques
 - ex. IBM, domaine médical,...
 - corpus de textes
 - ex. le corpus Brown (1965)
 - intranets des organisations
 - données de recherche : ICPSR (depuis 1962)
[<http://www.icpsr.umich.edu/>]
 - données avec documentation

ASPECTS HISTORIQUES (SUITE)

- Collections numériques : pré-WWW (*suite*)
 - projet Perseus (1985) [<http://www.perseus.tufts.edu/hopper/>]
 - textes, images, dictionnaires et cartes hyper-reliés
 - période grecque classique, puis au-delà
 - répertoire de logiciels : Netlib (1985) [<http://www.netlib.org/>]
 - logiciels mathématiques « de qualité »; aujourd'hui : + articles, etc.
 - toujours un leader dans le stockage et l'organisation de l'information numérique
 - « archives de prépublications » : Los Alamos E-print Archive (1991) [<http://arxiv.org/>]
 - prépublication d'articles en physique (puis autres sciences connexes)
 - en 1996 (Arms, 2000:28) : 35 000 utilisateurs de 70 pays, 70 000 transactions quotidiennes
 - collection de 1 344 162 documents à ce jour (7 janvier 2018)
 - statistiques d'utilisation : <http://arxiv.org/stats/today>

ASPECTS HISTORIQUES (SUITE)

- Collections numériques : depuis le WWW
 - le web comme collection
 - périodiques numériques
 - dépôts institutionnels
 - bibliothèques numériques



ASPECTS HISTORIQUES (SUITE)

- Collections numériques : depuis le WWW (suite)
 - bibliothèques numériques (suite)
 - ressources

Livres

- Arms (2000), Papy (2005)

Revue

- D-Lib Magazine [<http://www.dlib.org/>]
- International Journal on Digital Libraries

Organismes

- Digital Library Federation [<https://www.diglib.org/>]
- SIG-DL [<http://www.asis.org/SIG/sigdl/>]

Congrès

- ACM/IEEE Joint Conference on Digital Libraries (JCDL) [<http://www.jcdl.org/>]
- International Conference on Theory and Practice of Digital Libraries (TPDL) [<http://www.tpd.eu/>] (ex-ECDL)

ASPECTS TECHNOLOGIQUES

- Évolution de la micro-informatique
 - portabilité croissante des ordinateurs
- Logiciels, notamment interfaces : convivialité, simplicité
- Croissance des réseaux
 - configuration, vitesse
- Technologie propre à la gestion de textes
 - technologies de numérisation
 - représentation des jeux de caractères
 - langages de description de pages : TeX, PostScript, PDF
 - langages de description de fichiers texte
 - pour décrire structure et apparence des textes (à l'écran ou à l'écrit)
 - langages de balisage propriétaires
 - SGML, HTML, XML → DTD (spécification d'un type de fichier)

MICRO-INFORMATIQUE



INTERFACES

```

Command Prompt
C:\>dir /?
Displays a list of files and subdirectories in a directory.

DIR [drive:][path][filename] [/A[:attributes]] [/B] [/C] [/D] [/L]
[/O[:sortorder]] [/P] [/Q] [/S] [/T[:timefield]] [/U] [/X]

[drive:][path][filename]
    Specifies drive, directory, and/or files to list.

/A
    attributes      Displays files with specified attributes.
                    D Directories
                    H Hidden files
                    S System files

/C
    Uses bare format (no heading)
                    Displays the thousand separator
                    default. Use /C to disable.

/D
    Same as /L but files are sorted by date.

/L
    Uses lowercase.

/H
    More long list format where the first column is the
                    file name.

/O
    List by files in sorted order.

sortorder
    N By name (alphabetic)
    E By extension (alphabetic)
    G Group directories first

/P
    Pauses after each screenful.
    
```

Home Gopher server: gopherproject.org

THE GOPHER PROJECT

Welcome to GOPHER! Gopher is a slim, powerful, and fast way to present information in a hierarchical catalog of Gopher actually predates the Web -- although most web browsers make excellent gopher browsers too.

Good places to start are the "Why Gopher?" and "Using Gopher" areas!

- [12] *** GOPHER TURNS 10 / GOPHER 3.0 (FurryTerror) RELEASED ***
- [13] *** GOPHER TURNS 10 ..R 3.0 (FurryTerror) RELEASED *** [n
- [14] A Brief Introduction to Gopherspace
- [15] Clients, Servers, and Downloads/
- [16] Home Gopher at LMN (a good place to browse)/
- [17] Home Gopher at LMN [alternate]/
- [18] Mailing List
- [19] Mailing List Archives/
- [20] Major Gopher Servers/
- [21] Screenshots/

The top screenshot shows the NCSA Mosaic browser window displaying the 'NCSA Mosaic Home Page'. The browser interface includes a menu bar (File, Edit, Options, Navigate, History, Annotate) and a toolbar with icons for back, forward, home, and search. The document title is 'NCSA Mosaic Home Page' and the URL is 'http://www.ncaa.uiuc.edu/SDG/Software/Mosaic/NCSA/'. Below the browser window, an arXiv.org e-Print archive page is visible, showing a 'Physics' section with a list of articles.

SET MENU

- MEMORY
- AUTO SETUP
- MANUAL SETUP
- SOUND SET MENU
- SOUND OUT MENU
- INPUT MENU
- DISPLAY MENU

[▲] / [▼]
[ENTER]

The screenshot shows the Slackware Linux Setup menu. At the top, it says 'Welcome to Slackware Linux Setup. Select an option below using the UP/DOWN keys and SPACE or ENTER. Alternate keys may also be used: ←, →, and TAB.' The menu items are:

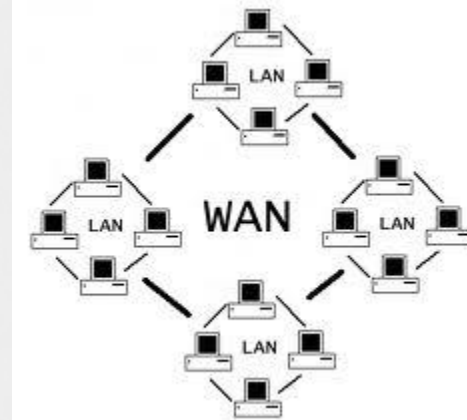
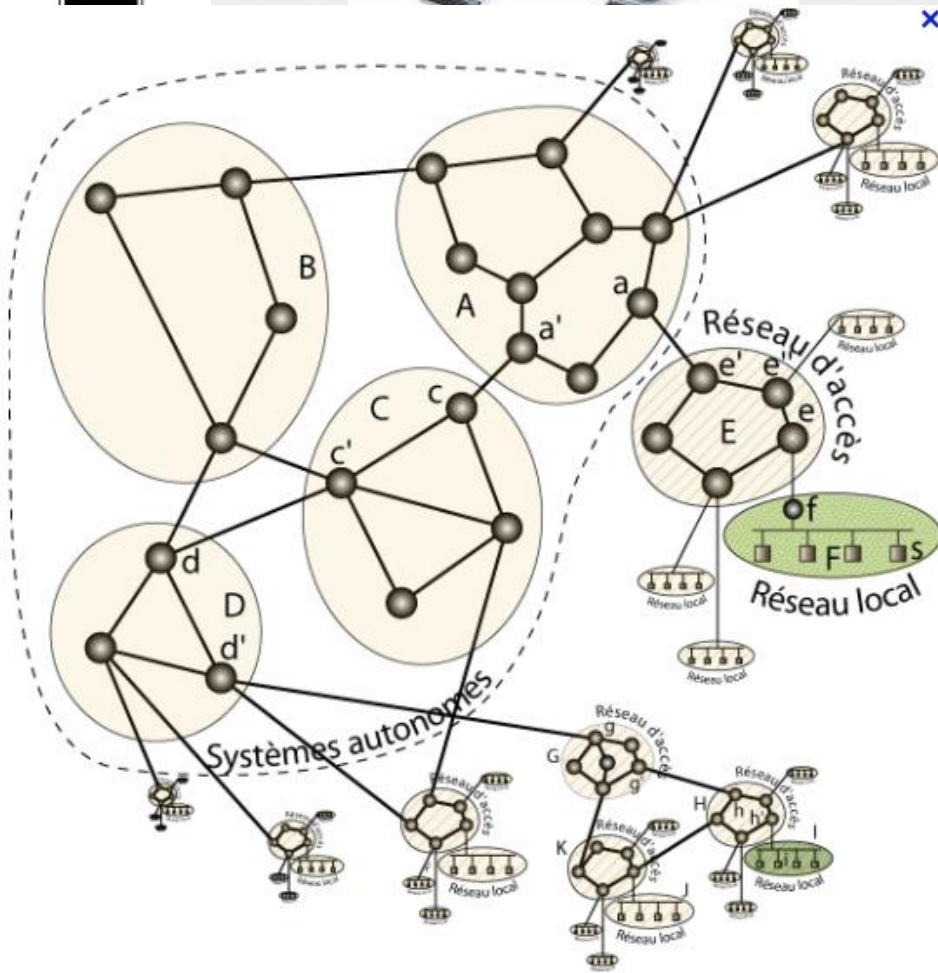
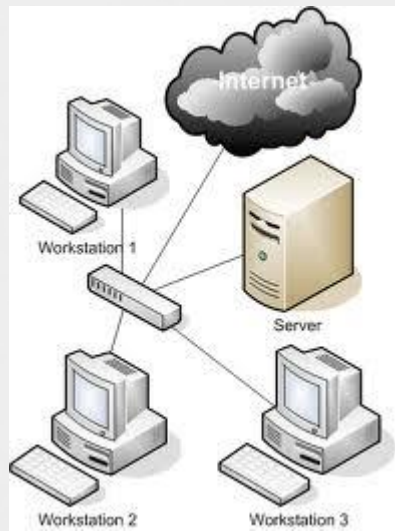
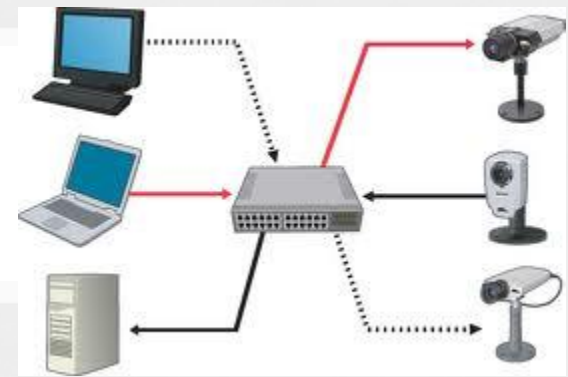
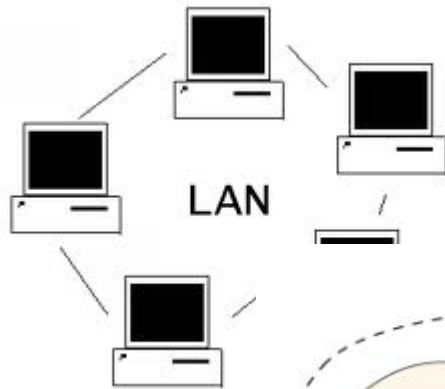
- HELP: Read the Slackware Setup HELP files
- KEYMAP: Remap your keyboard if you're not using a US one
- AWAYOUT: Set up your swap partition(s)
- TARGET: Set up your target partitions
- SOURCE: Select source media
- SOFTSET: Select categories of software to install
- INSTALL: Install selected software
- CONFIGURE: Reconfigure your Linux system
- EXIT: Exit Slackware Linux Setup

 Navigation options at the bottom include '< OR >' and '<Cancel>'.

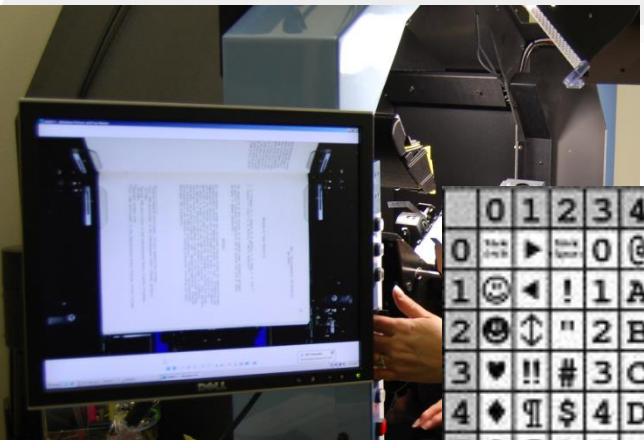
The screenshot shows a Windows Start menu with a purple background. The date is '22 Monday'. The taskbar includes icons for Mail, Photos, Music, Video, Internet Explorer, Firefox, Google Chrome, and the Windows Start button. The Start menu is open, displaying various live tiles for applications like Mail, Photos, Music, Video, Internet Explorer, Firefox, Google Chrome, and Windows Store. The bottom right corner shows the 'Acer' logo.

The screenshot shows a Windows desktop environment. The taskbar includes icons for Internet Explorer, Firefox, Google Chrome, and the Windows Start button. The desktop background is a blue gradient with a clock and calendar widget. A 'Larc-Breaking News' window is open, displaying a news article. The system tray shows the date and time as '11:42 AM'.

TECHNOLOGIE RÉSEAU

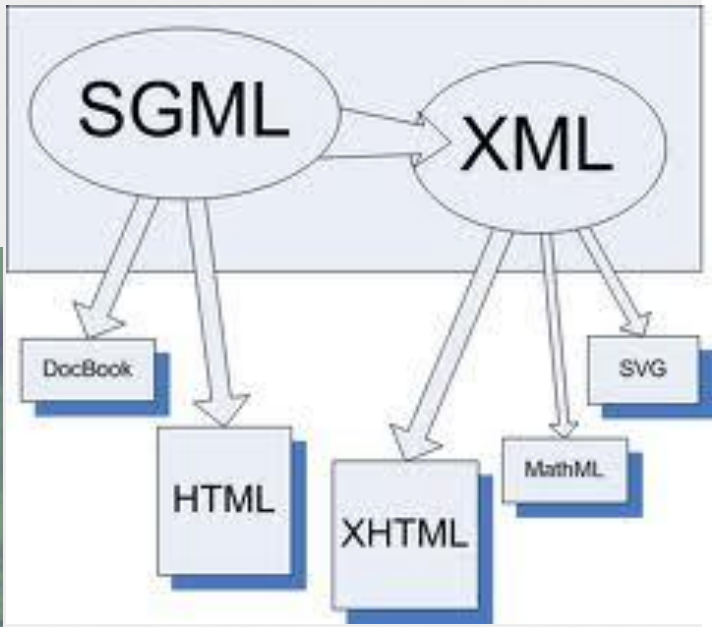
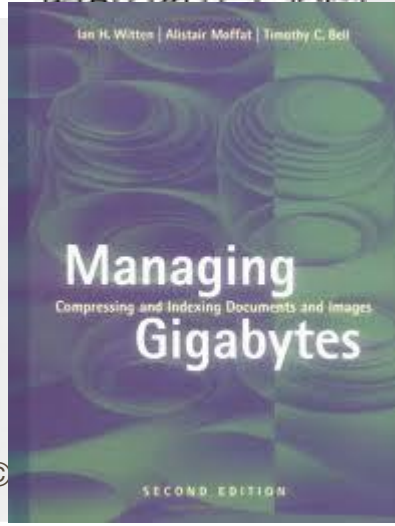
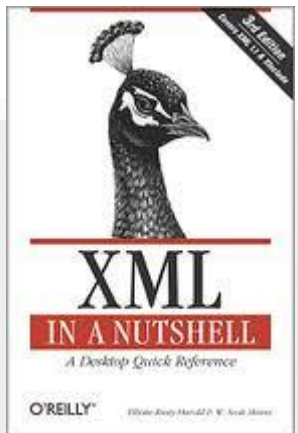


TECHNOLOGIES TEXTUELLES



0	1	2	3	4	5	6	7
0	☺	▶	◀	0	@	P	' p
1	☺	◀	!	1	A	Q	a q
2	☺	↕	"	2	B	R	b r
3	♥	!!	#	3	C	S	c s
4	♦	¶	\$	4	D	T	d t
5	+	§	&	5	E	U	e u
6	♣	-	&	6	F	V	f v
7	•	↕	`	7	G	W	g w
8	•	↑	(8	H	X	h x
9	○	↓)	9	I	Y	i y
A	○	→	*	:	J	Z	j z
B	σ	←	+	;	K	[k (
C	○	!	~	~	~	~	~
D	○	!	~	~	~	~	~
E	○	!	~	~	~	~	~
F	○	!	~	~	~	~	~

8	9	A	B	C	D	E	F
0	Ç	É	á	☐	α	=	
1	ü	æ	í	☐	β	±	
2	é	Æ	ó	☐	Γ	≥	
3	â	ô	ú	☐	π	≤	
4	ä	ö	ñ	☐	Σ	Γ	
5	à	ò	Ñ	☐	σ	J	
6	â	û	ä	☐	μ	÷	
7	ç	ù	ó	☐	τ	≈	
8	ê	ÿ	ç	☐	Φ	φ	
9	ë	Ö	γ	☐	Θ	•	
A	è	Ü	γ	☐	Ω	•	
B	ï	ç	¼	☐	δ	√	
C	î	£	¼	☐	∞	π	
D	ì	¥	ì	☐	∅	z	
E	Ä	Þ	«	☐	€	■	
F	Å	f	»	☐	™	■	



ASPECTS TECHNOLOGIQUES (SUITE)

- En parlant d'XML... Deux DTD importantes
 1. TEI (1987) [www.tei-c.org] : pour représenter tout type de texte
 - support : papier, parchemin, papyrus, roche
 - technique : imprimerie, écriture manuscrite, dactylo et plus
 - ajouts : avec annotations, suppressions et parties abimées
 - genres : prose, poésie, théâtre, dictionnaires, fichiers de données...
 - spécialement pour chercheurs en sciences humaines
 - familles de DTD
 - « Chicago pizza model » : une DTD de base, une spécialisation, des ajouts éventuels
 2. EAD (1993) [<http://www.loc.gov/ead/>] : pour encoder versions numériques d'outils de recherche archivistiques
 - développée par et pour les archivistes (sur la base de l'expérience de TEI)
 - documentation imposante

ASPECTS TECHNOLOGIQUES *(SUITE)*

- Métadonnées
 - i.e. clés de description
 - évolution parallèle à la tradition en sciences de l'information
- Organisation
 - aspect collaboratif de la description et de l'indexation

ASPECTS SOCIAUX

- Agents impliqués dans les bibliothèques numériques
 - informaticiens
 - littéraires
 - linguistes(-informaticiens)
 - puis, plus récemment : bibliothécaires et archivistes
- Accès aux collections par les utilisateurs
 - direct (sans intermédiaire, grâce aux logiciels de RI)
 - contributions de documents + informations sur eux
- Création de communautés virtuelles distribuées géographiquement (ex. médecins, juristes, physiciens,...)
- Multilinguisme croissant sur le Web
- Édition numérique (notamment les périodiques)

Autonomie et exigences croissantes des utilisateurs

AUTRES ASPECTS

- Aspects juridiques
 - flou concernant le droit d'auteur (diffusion)
- Aspect économique
 - pour les bibliothèques numériques : cohabitation (Arms, 2000:3)
 - d'accès libre (issus des scientifiques à l'origine de certaines technologies)
 - d'accès payant (issus d'éditeurs et bibliothécaires)

AUTRES ASPECTS *(SUITE)*

- Quelques cas spéciaux
 - Google Books
 - répercussions sociales et économiques importantes
 - données ouvertes
 - mouvement mondial de mise à la disponibilité de données de types variés
 - données de recherche
 - de grands chantiers en cours et à prévoir
 - images et vidéos en nombre croissant
 - auto-publications

CONCLUSION PROVISOIRE

- Facteurs précités : ont modelé la création de collections numériques
- Préoccupations
 - préservation de l'information numérique
 - interopérabilité des systèmes
 - modes d'organisation et de description des collections
 - importance de la définition des métadonnées

BIBLIOGRAPHIE

- Arms, William. 2000. *Digital Libraries*. Londres; Cambridge (Mass.) : The MIT Press.
- Chowdhury, G.; Chowdhury, S. 2003. *Introduction to digital libraries*. Facet Publishing, UK.
- **Jacquesson, Alain; Rivier, Alexis. 2005. *Bibliothèques et documents numériques. Concepts, composantes, techniques et enjeux*. Paris : Éditions du Cercle de la Librairie. (chap. 3 “Fondements historiques des documents numériques”, pp. 47-69)
- Papy, Fabrice. 2005 . *Les bibliothèques numériques*. Paris : Lavoisier.
- Tedd, Lucy A.; Large, J.A. 2005. *Digital libraries: principles and practice in a global environment*. München : K.G. Saur.