Data for the Web

INSA de Lyon

Computer Science and Information Technology Department

3rd Year

Part 1: Introduction, Documents

Előd EGYED-ZSIGMOND

INSTITUT NATIONAL DES SCIENCES APPLIQUÉES LYON

Plan

- Introduction
  - Databases overview
  - Documents
- XML Core
- XML Galaxy
- NOSQL
- Conclusion

Databases overview

- Database definition
  - Data and the way they are structured
  - More or less related data collection

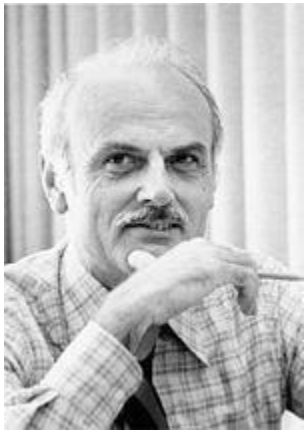
- Features
  - Usually a representation of the real world
  - Collection specifying
    - Data Signification (name, type, ….)
    - Intra and inter data constraints

Databases overview

- DBMS database management system
  - (SGBD Système de Gestion de BD)
  - Set of programs enabling:
    - Database content description (*dictionary*)
      - structure
      - data types
      - constraints
    - data storage (*the database*)
    - Data handling / manipulation (*langages*)
      - data manipulation language (DML):
        - query (search info)
        - creating / updating data
      - data description language (DDL)
  - database design and programing

Short history

- 1960 navigational models (hierarchical and CODASYL)
- 1970 relational model (Codd)
- 1980 object-relational
- 2009 NoSQL



Edgar Frank Codd
(source: Wikipedia)

DBMS history

- 2nd generation
  - data model: relational

see
Relational database course

table Name

name-col 1	**name-neck 2**	**-------**	**No name-neck**
Val (1.1)	**Val (2.2)**		**Val (2, n)**
Val (j, 1)	**Val (j, 2)**		**Val (j, n)**

Schema: Table-name (col_name_1, col_name_2.... col_name_n)

DBMS history

- 2nd generation
  - data model: relational
  - DDL: interpreted with dynamic link creation
  - DML: ensemblist
    - independently usable (query language)
    - or in a tier language (C, JAVA, COBOL, ...)
  - Standardization (widely adopted):
    - first standards in 1986 (SQL1, revised in 1989)
    - second standard in 1992 (SQL2)
    - SQL3 standard in 1998 *(Introduction of object concepts).*
  - Oracle 1 (1978)
  - Genealogy on:
    - http://fadace.developpez.com/sgbdcmp/story/ (accessed 01.20.2019) or on wikipedia : https://en.wikipedia.org/wiki/Database

DBMS history

- 2nd generation
- advantages:
  - more conceptual data and well defined mathematical model
  - dynamic structuring of the physical space (more separation between DDL and DML)
  - wide spectrum of users (end-users)
  - single query language for the database and the dictionary (SQL)
  - powerful development environment
    - states generators,
    - transactions generators,
    - development tools...
  - distributed versions
  - existing versions on all hardware types

DBMS history

- 2nd generation
- weaknesses
    - low modeling power with respect to new applications (model with a single hierarchical level of description)
    - integrity constraints difficult to express in a declarative manner
    - less efficient data access (compensation by the increased power of computers in the 90s)
    - constraints poorly adapted to distributed data

DBMS history

- 3rd generation
  - use richer data models enabling:
    - more complex and user-defined data structures
      - complex objects,
      - semistructured data
      - distributed data

      (hyper) Documents

      WEB
    - **multimedia** data management (Images, videos, sound, ...)
    - Object oriented aspects
    - DBMS level Automatic persistent management

DBMS history

- NoSQL (N only SQL) used for the first time in 1998
- June 11, 2009 NoSQL meetup San Francisco
- Too many distributed data / too many constraints
  - in the big data and real-time Web
- Types
  - Column
  - Graph
  - Document
  - Key - value

Atomicity
Consistency
Isolation
Durability

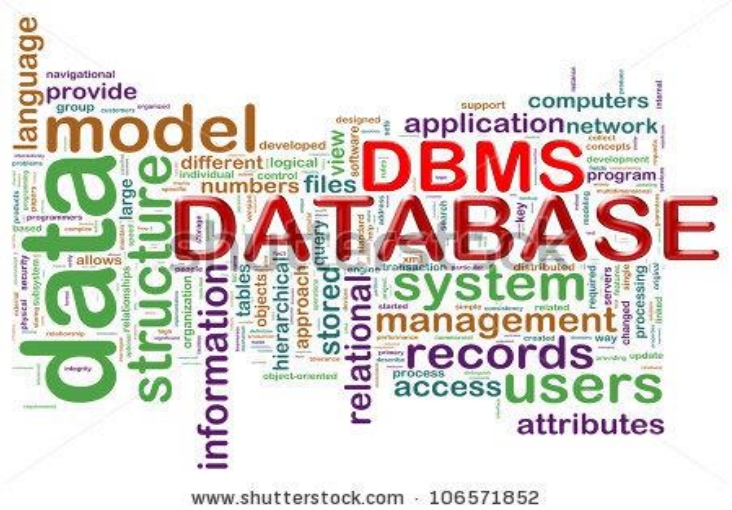
DBMS history - Conclusion

Several DBMS Generations

The evolution is Not over

SQL survives!

Other technologies
born and
co-exist



NOSQL

Plan

- Introduction
  - Reminders of BD
  - The documents
    - Introduction
    - Document modeling
    - Hyperdocuments
    - Document types (text, image, …)
- XML Core
- XML galaxy
- NOSQL
- Conclusion

Document

Docere (lat.): teach

The 3 meanings of "document":

1. What the author wants to express *(Intentio auctoris)*
2. The "proper" meaning of the docment *(Intentio operas)*
3. The sense understood by the consumer *(Intentio lectoris)*

Umberto Eco

The media

In communication, the media (or channel) is the support on which a message is based, through which it is distributed.

- information Diffusion : printing press, radioTV...
- Description, type of information: text, sound, image, video
- Perception of the information: hearing, sight, ...
- Storage of the information: paper, tape, DVD, ...

Plan

- Introduction
  - Reminders of BD
  - The documents
    - Introduction
    - Document modeling
    - Hyperdocuments
    - Document types (text, image, …)
- Core XML
- XML galaxy
- NOSQL
- Conclusion

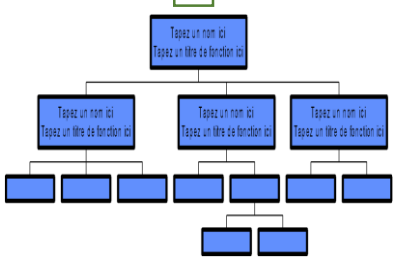
Introduction to documents

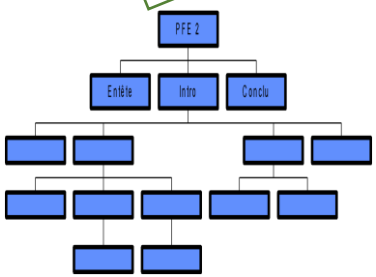
Logical structure — **Document models** — **Physical structure**

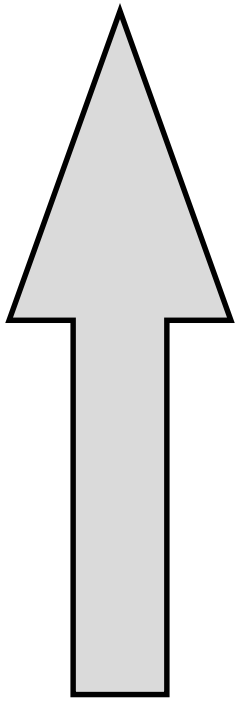
Master thesis

Project

Title
Athor
Date
Introduction
State-of the art
Contribution
Experiments
Conclusion
References







INSA LYON year

author

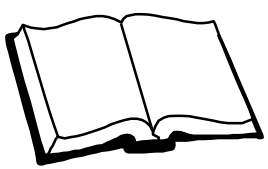
"Title"

business

p1

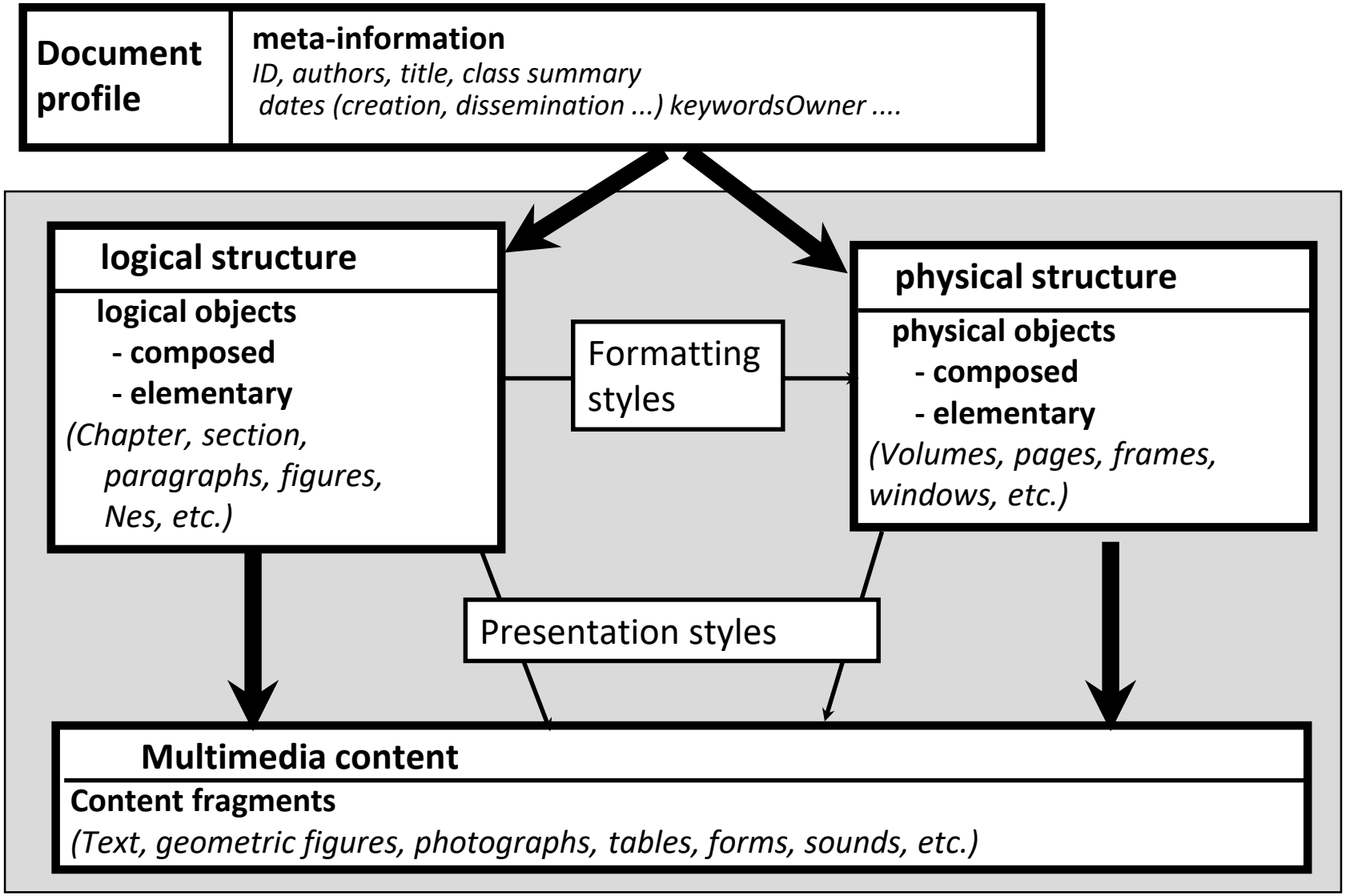
p2

A4 format



Logical structure — **specific Documents** — **Physical structure**

Document modeling



Specific document

INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES LYON

the NHI 01/10/2018

Management

Paul Haddock

topic : Media

Dear Colleague

aaaaaaaaaaaaaaaaaaaaaaaaaaa
aaaaaaaaaaaaaaaaaaaaaaaaaaaaaa
aaaaaaaaaaaaaaaaaaaaaaaaaaaaaa
aaaaaaaaaaaaaaaaaaaaaaaaaaaaaa
aaaaaaaaaaaaaaaaaaaaaaaaa.

bbbbbbbbbbbbbbbbbbbbbbbbbb
bbbbbbbbbbbbbbbbbbbbbbbbbbbb
bbbbbbbbbbbbbbbbbbbbbbbbbbbb

ref. JD / PRF / NHI-C / 1 *P 1/2*

bbbbbbbbbbbbbbbbbbbbbbbbbbb
bbbbbbbbbbbbbbbbbbbbbbbbbbb
bbbbbbbbbbbbbb.

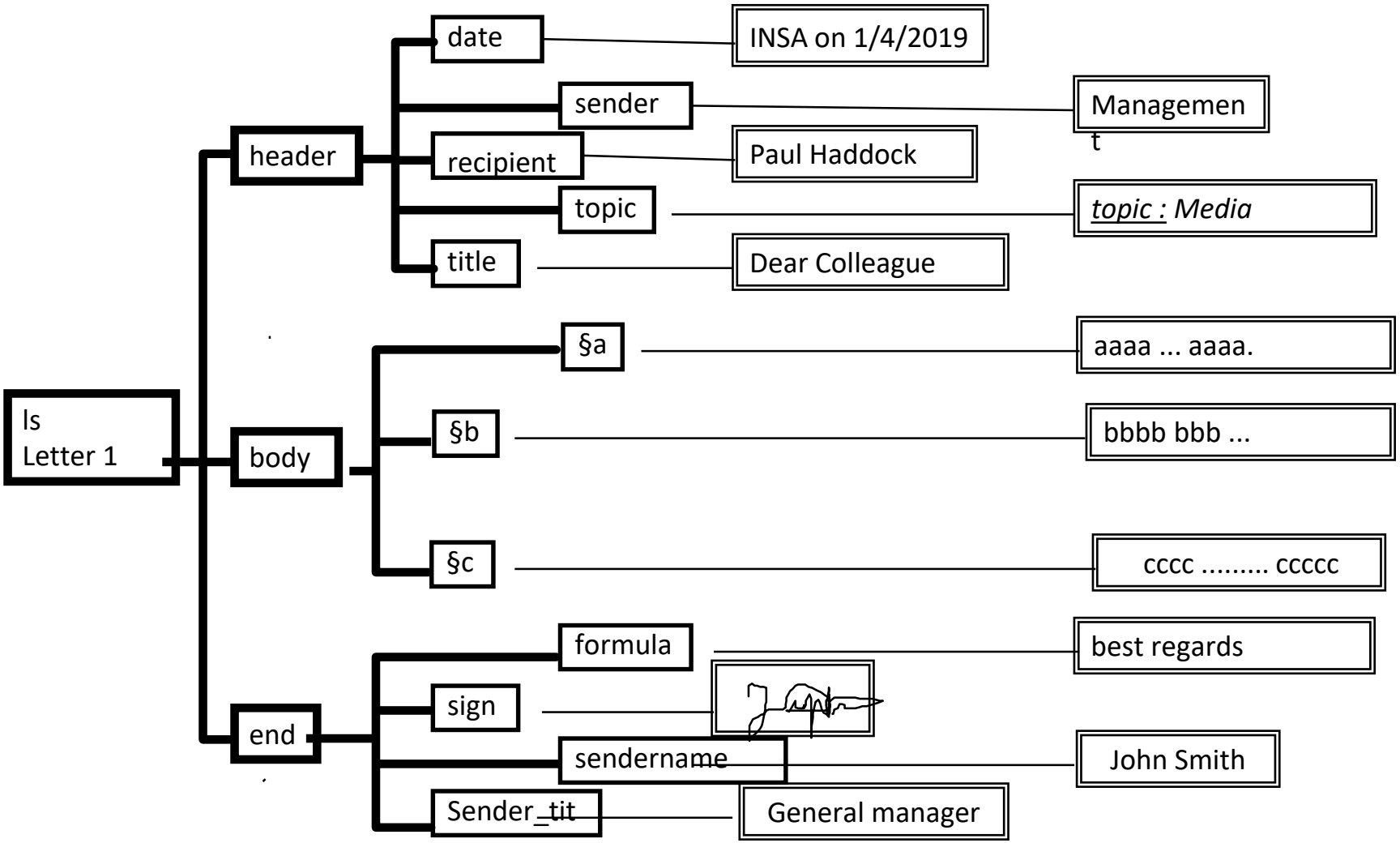
cccccccccccccccccccccccccccc
cccccccccccccccccccccccccccccc
cccccccccccccccccccccccccccccc
cccccccccccccccccccccccccccccc
cccccccccccccccccccccccccccccc

best regards

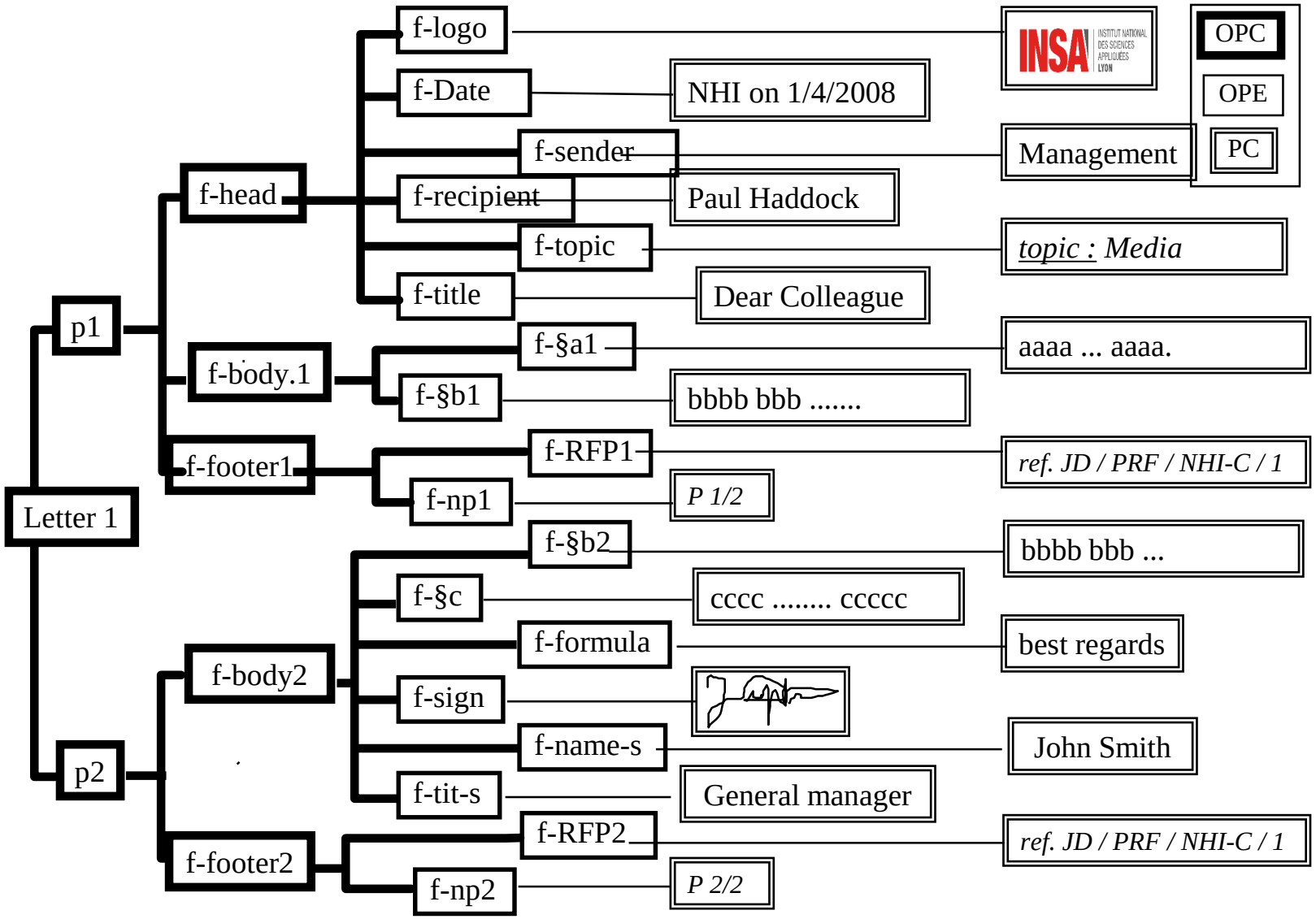
John Smith
General manager

ref. JD / PRF / NHI-C / 1 *P 2/2*

Logical structure



specific document



Document Class

letter 1

INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES LYON

the NHI 01/10/2018

Management

Paul Haddock

topic : Media

Dear
Colleague

aaaaaaaaaaaaaaaaaaaaaaaaaaaa
aaaaaaaaaaaaaaaaaaaaaaaaaaaaaa
aaaaaaaaaaaaaaaaaaaaaaaaaaaaaa
aaaaaaaaaaaaaaaaaaaaaaaaaaaaaa
aaaaaaaaaaaaaaaaaaaaaaaaa.

bbbbbbbbbbbbbbbbbbbbbbbbbb
bbbbbbbbbbbbbbbbbbbbbbbbbbb
bbbbbbbbbbbbbbbbbbbbbbbbbbb

ref. JD / PRF / NHI-C / 1 *P 1/2*

bbbbbbbbbbbbbbbbbbbbbbbbbbb
bbbbbbbbbbbbbbbbbbbbbbbbbbb
bbbbbbbbbbbbbb.

cccccccccccccccccccccccccccc
cccccccccccccccccccccccccccccc
cccccccccccccccccccccccccccccc
cccccccccccccccccccccccccccccc
cccccccccccccccccccccccccccccc

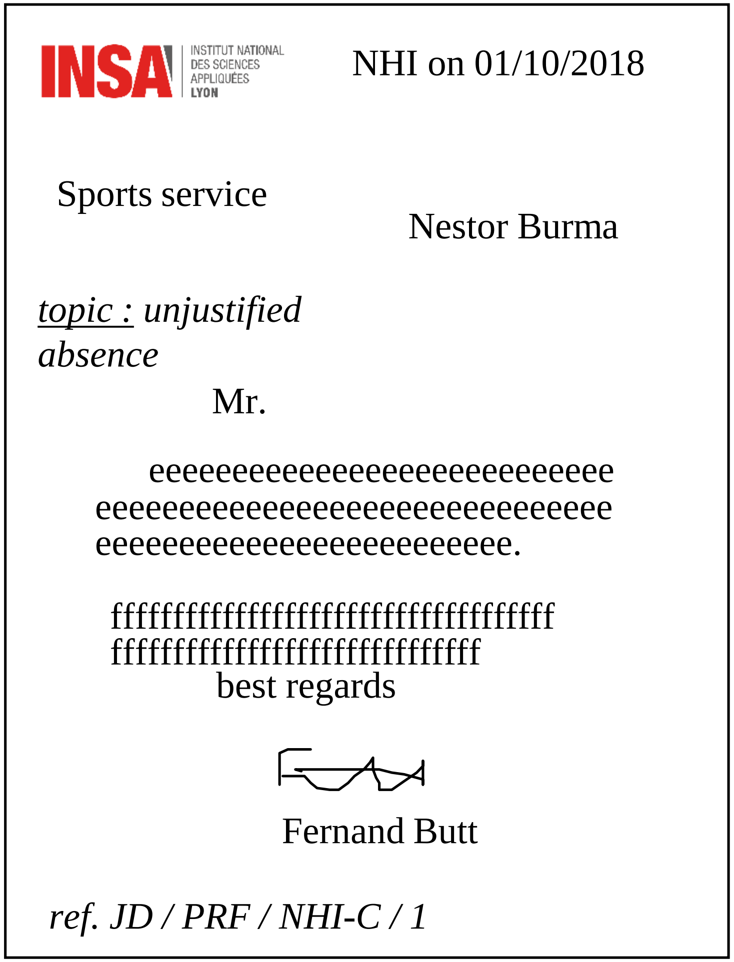
best regards

John Smith
General manager

ref. JD / PRF / NHI-C / 1 *P 2/2*

Document Class

letter 2



NHI on 01/10/2018

Sports service

Nestor Burma

<u>topic :</u> unjustified absence

Mr.

eeeeeeeeeeeeeeeeeeeeeeeeeeee eeeeeeeeeeeeeeeeeeeeeeeeeeeeeee eeeeeeeeeeeeeeeeeeeeeeeee.

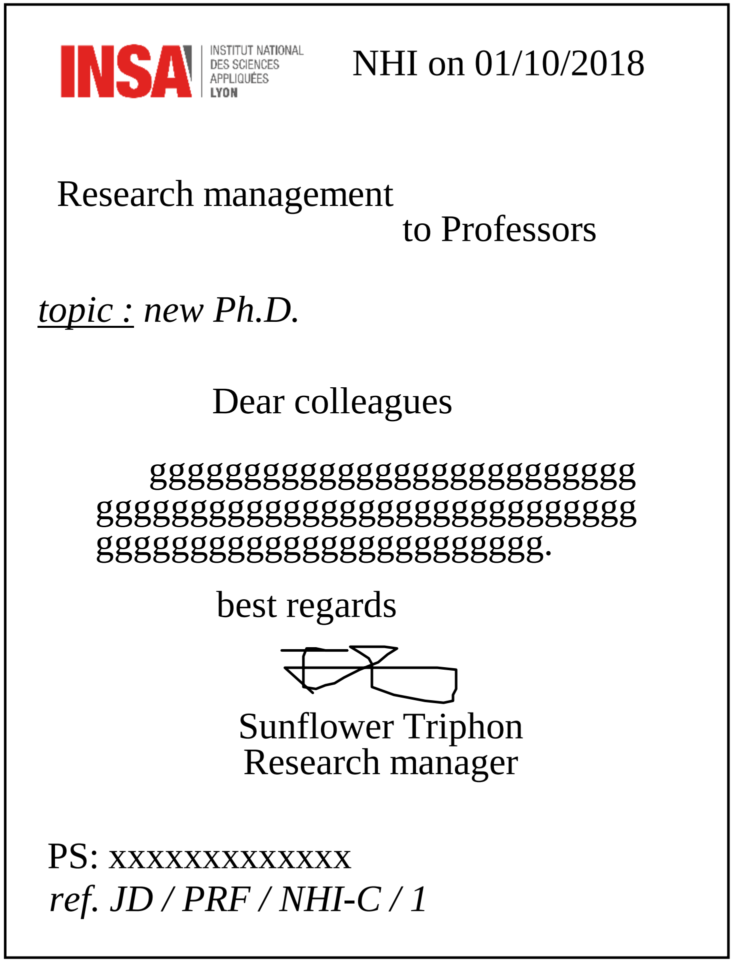
ffffffffffffffffffffffffffffffffffff ffffffffffffffffffffffffffffff

best regards

Fernand Butt

ref. JD / PRF / NHI-C / 1

letter 3



NHI on 01/10/2018

Research management

to Professors

<u>topic :</u> new Ph.D.

Dear colleagues

gggggggggggggggggggggggggg ggggggggggggggggggggggggggggg gggggggggggggggggggggggg.

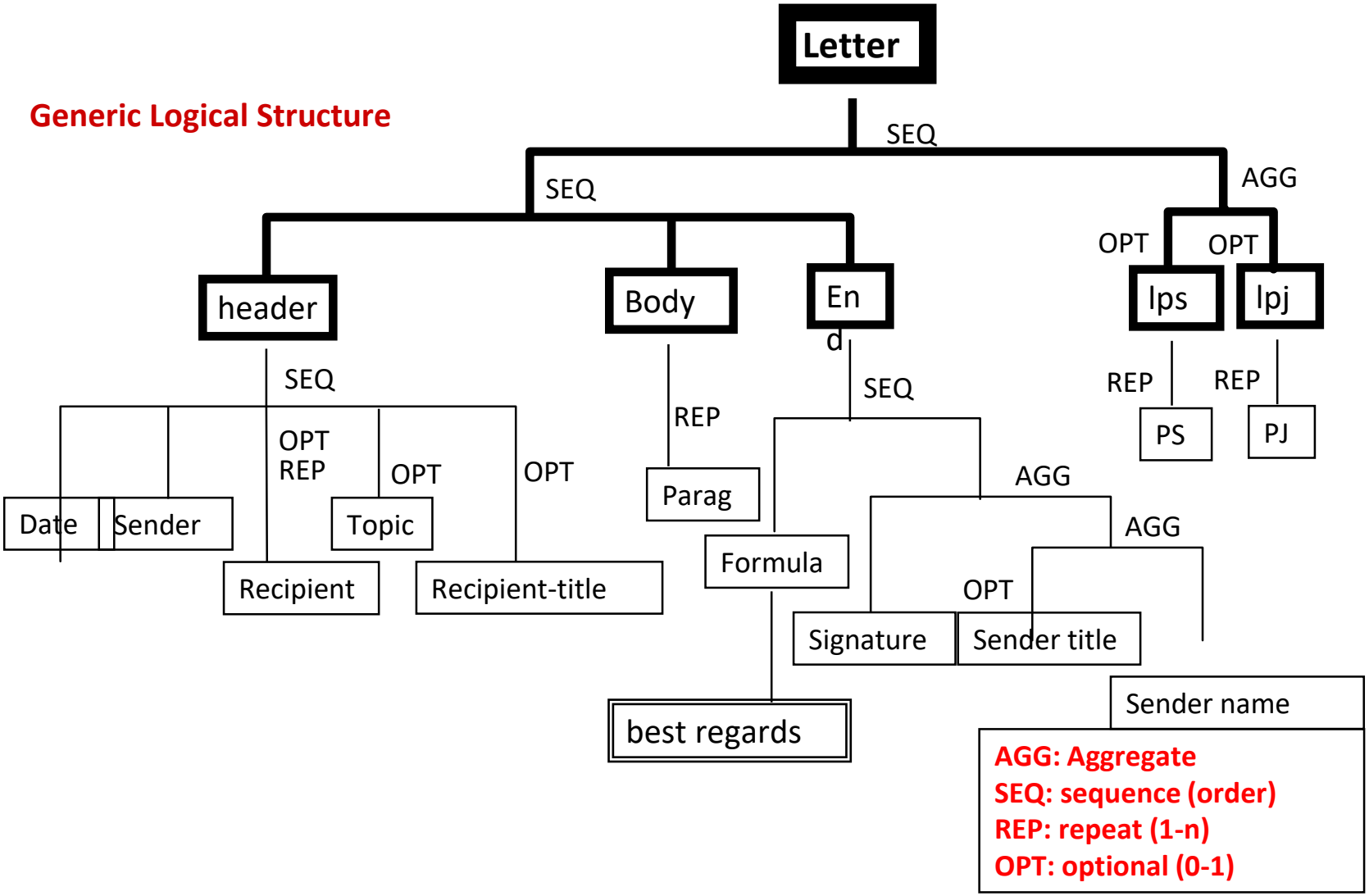
best regards

Sunflower Triphon
Research manager

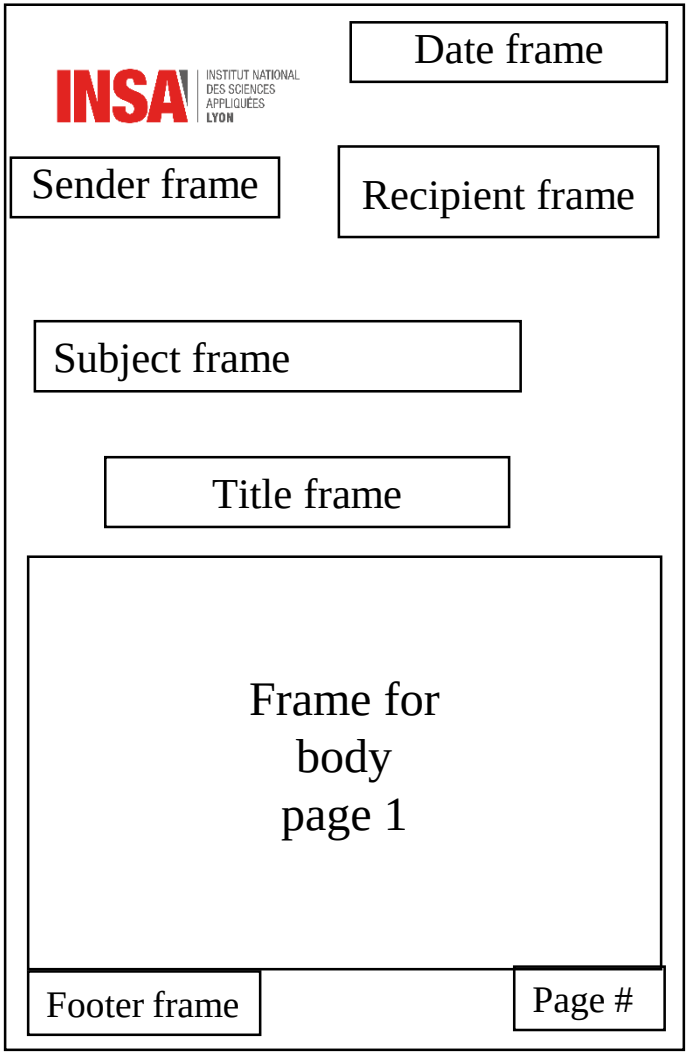
PS: xxxxxxxxxxxxx

ref. JD / PRF / NHI-C / 1

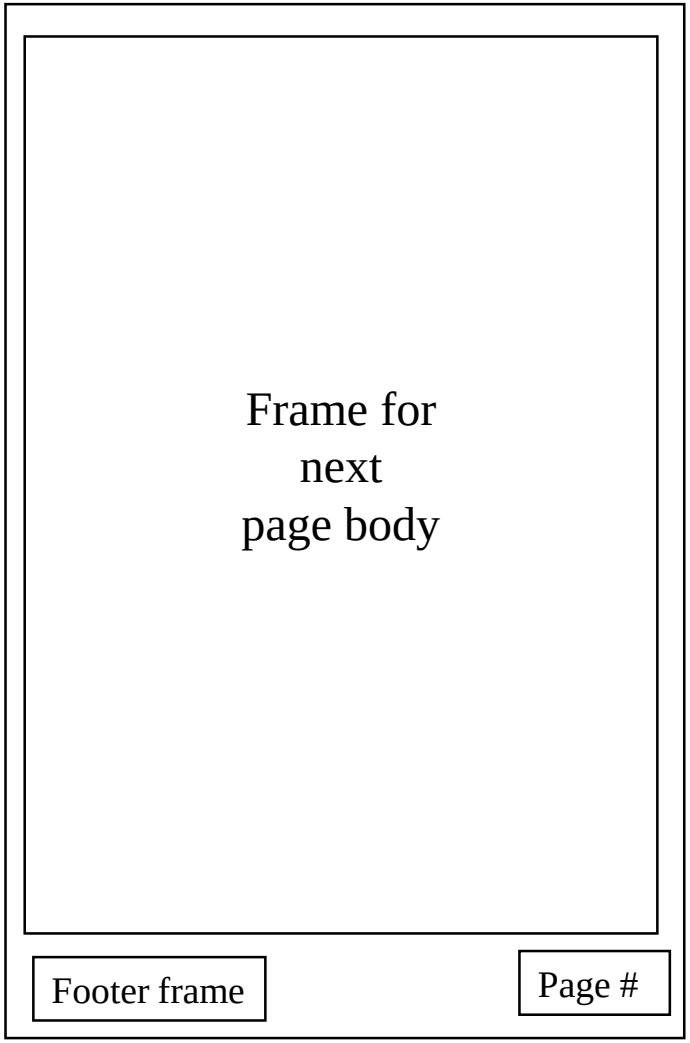
Document Class



Document Class



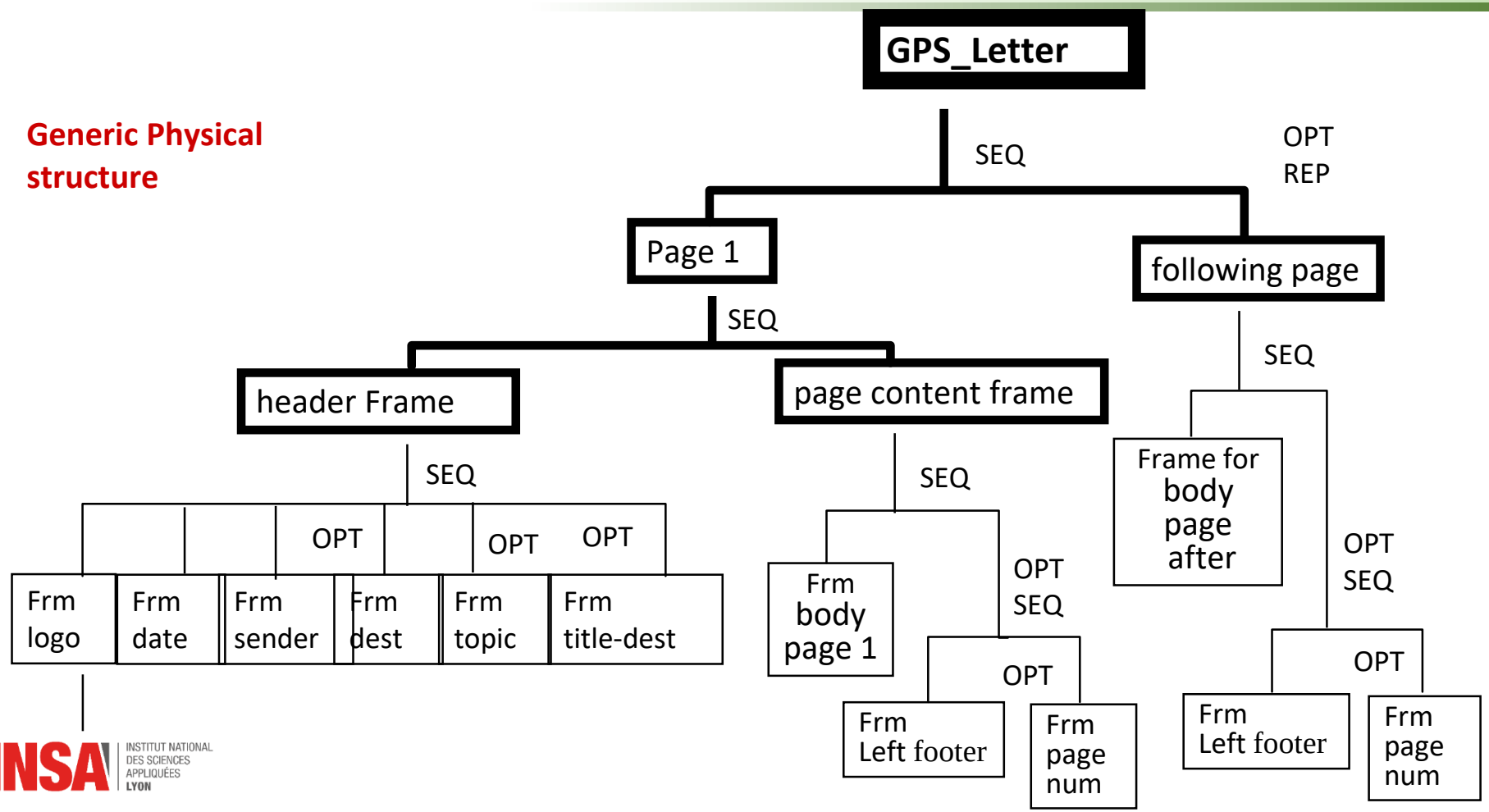
a. First page



b. Second page

Document Class

Generic Physical structure



SEQ: sequence (order)
REP: repeat (1-n)
OPT: optional (0-1)

Modeling documents

Document profile

physical and logical structure

contents

Plan

- Introduction
  - Reminders of BD
  - The documents
    - Introduction
    - Document modeling
    - hyperdocuments
    - Document types (text, image, …)
- Core XML
- XML galaxy
- NOSQL
- Conclusion

Hyperdocuments

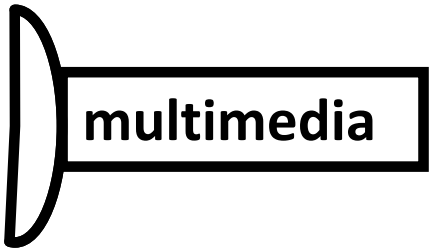
- **The first hypertext was created by Vannevar Bush** (Roosevelt's advisor) **in 1945.**
  - complex notes (paper ...) network management system
  - extending the memory capacity *(Called MEMEX)*
  - which didn't use computer tools.

- **IT has been introduced to manage the network in 1960** (Engelbart, Nelson, ...)
  - it was at this time that were invented
    - the word "hypertext"
    - the mouse.

☞ **WARNING : distinguish :**
  - **hypertext system** (software)
  - **textual hyperdocument** (Information managed by a system. hypertext)
    likewise
  - **hypermedia system**(software)
  - **multimedia hyperdocument** (inform. managed by asystem. hypermedia)

Hyperdocument specification

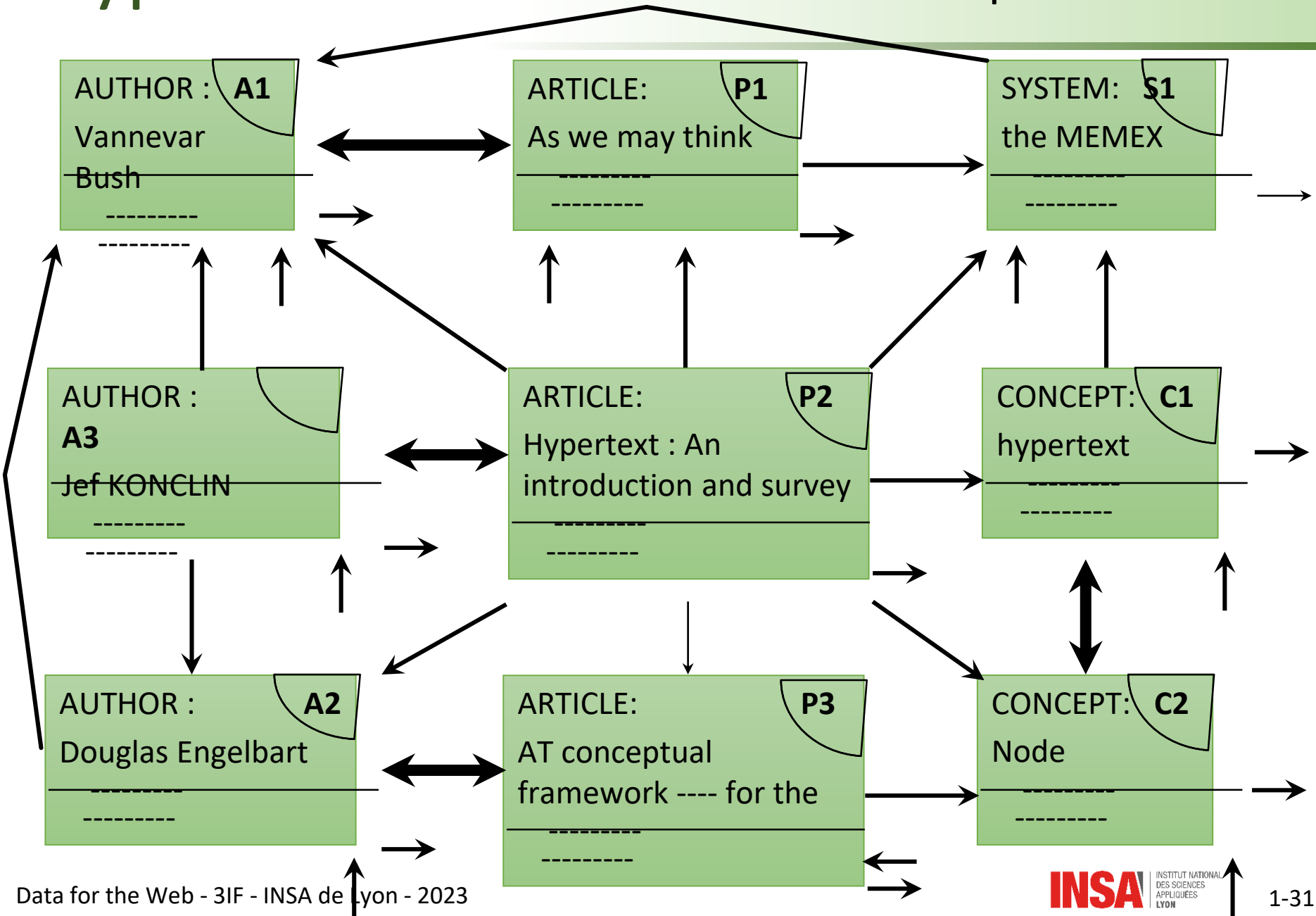
- Structured set of nodes and links
- The nodes are associated with content
  - text *(With or without options)*
  - graphics *(Geometric or photographic)*
  - sound or video
  - Node -> Semantic unit



- The links (typed) between nodes define the structure
  - hypergraph
  - graph
  - Tree (= document)
- Several possible views
  - paths to guide the reader are required
  - designing a hypermedia is different than designing a document

Hyperdocuments

example



Hyperdocuments

example

AUTHOR : Vannevar Bush	**A1**
Vannevar Bush was advisor President Roosevelt. He is credited with the first work on **hypertext**. In 1945, he published an article**« As we may think"** wherein he presents his system **MEMEX** ----------	

ITEM: As we may think	**P1**
author : **Vannevar Bush** Date: July 1945 Editor: Atlantic Monthly pages: 101-108 Summary: V. Bush presents a graph management system of handwritten Nes, named the **MEMEX** (Memory Extension) ---	
ace we may think *Vannevar Bush* ---------- ---------- ------ ---------- ------ ----------	

CONCEPT: Hypertext	**C1**
The hypertext concept is simple: windows in a screen are associated with objects called **nodes.** These nodes are connected by **connections,** --------- ---------	

Hypermedia, Internet and Web

- hypermedia = hyperdocument+ multimedia
  - hyperdocument : Generalization of the document
    - logical structure
      - a document: a tree
      - a hypermedia : A hypergraph
  - multimedia: content nodes or hyper-nodes may be
      - text, table,
      - Image (pattern, photography, drawing)
      - formula (mathematical, chemical, ...)
      - sound (speech, music, sound, ...), video, ...
    - *Example of multimedia hyperdocument (hypermedia)*
      - *electronic encyclopedia*
      - *an electronic game*
      - *etc.*

Hypermedia, Internet and Web

(W3) *"World Wide Web"*

- multimedia information system for exchange on the INTERNET;
- designed in 1989 by Tim Berners-lee *(CERN)* to allow researchers and visitors at CERN to exchange scientific information (articles, reports) after they leave;
- based on the extension of the "hyperdocument" concept to International Networks (distributed hypermedia);
- improving the existing *(Not a revolution)* :
  - allows more user-friendly access to existing servers *(WAIS, Gopher, FTP, ... existed before but required specific "clients")*

Hypermedia, Internet and Web

HTML: HyperText Markup Language:

- ❑ hyperdocument representation m**odel**
- ❑ used by **WEB** servers and clients
- ❑ based on the **SGML** standard *(This is an SGML document template - Document Type Definition)*
  - defines both: - the logical structure of a node
    - its physical structure and presentation
- ❑ constantly **evolving**:
  - **HTML-1 (1989)** : text, some styles, hyperlinks
  - **HTML-2 (1994):** HTML-1 images, interactive forms
  - **HTML-3 (1996)** HTML-2 + vector graphics, sound, applets
  - **HTML 4, DHTML (1998):** HTML-3 + video + CSS + tools for Intranet,
  - **XHTML (2000)** HTML reformulated as XML *(I.e. a DTD)*
  - **HTML-5 (2014)** *:* multimedia, semantics
  - *HTML-5.2 (2017)*

Hypermedia, Internet and Web

- HTTP a Communication protocol
  - Classical "Client - Server" model.
    - **WEB server :** program *"running"* on a computer whose only purpose is to *reply* to client requests.
      - file transfer request
      - ***execution result of a*** program on the server
        the originality of the web that allows its interfacing with virtually any software
    - **WEB client :** Program that enables a user:
      - to ***submit requests to*** a Web server, to ***display*** results and ***navigate*** in a **HTML** document;
      - To ***communicate*** with other types of servers *(FTP, mail, ...);*
      - possibly other services *(Custom type)*

Hypermedia, Internet and Web

- XML = EXtensible Markup Language
  - if you know HTML: *extensible HTML form that defines its own tags*
  - XML
    - **was designed for INTERNET and INTRANET**
    - **must support a variety of applications**

- XML is developed and managed by the W3C

  http://www.w3.org/

Plan

- Introduction
  - Reminders of BD
  - The documents
    - Introduction
    - Document modeling
    - hyperdocuments
    - Document types (text, image, …)
- Core XML
- XML galaxy
- NOSQL
- Conclusion

Multimedia

- *Multus* and *medium* adjective / name

- digital product incorporating several media on the same support(text, images, sounds)

- New creation and dissemination and consultation domain

- Realize multimedia is: create, process, store, organize, annotate, link and synchronize digital files.

Different media: size concepts

1 Bit	Zero or 1	2 ^ 1	Zero or 1
1 Byte	8 Bit	2 ^ 8 (bits)	Value from 0 to 255, or a character
2 Bytes	16 Bits	2 ^ 16 (bits)	Value from -32768 to 32767 or character of any writing system in the world
8 Bytes	64 Bit	2 ^ 64 (bits)	Floating point value representing +/- 16 digits of precision (scientific number)
1 kilobytes	1024 Bytes	2 ^ 10 bytes	An average page of text or a standard color icon
1 megabytes	1024 kilobytes	2 ^ 20 bytes	1,000 pages of text, graphic screen 1 full page, 6 seconds of sound CD quality.
1 Gigabytes	1024 megabytes	2 ^ 30 bytes	1 million pages of text, 1 hour and a half of his CD quality, 50 seconds of uncompressed video.
1 Terabytes	1024 Gigabytes	2 ^ 40 bytes	The library of congress full text form (approximately) 62 continuous days of music, 14 hours of uncompressed video
1 Peta-octets	1024 Terabytes	2 ^ 50 bytes	Probably more text than anything that has been produced in the history of humanity (for all the known languages), 170 years of music, video 19 months.
1 Exa-octets	1024 Peta-octets	2 ^ 60 bytes	overall monthly data traffic in 2004

Plan

- Introduction
  - Reminders of BD
  - The documents
    - Introduction
    - Document modeling
    - hyperdocuments
    - Document types (text, image, …)
- Core XML
- XML galaxy
- NOSQL
- Conclusion

The text: encoding

- **ASCII** (American Standard Code for Information interchange 1963), 8 bit
- **16 bit Unicode** : Over 65 000 characters covering 100 scripts
- **UTF-8** : Variable length (1-4 bytes) used by 82.5% of websites in February 2015[1]

1. http://w3techs.com/technologies/overview/character_encoding/all

ASCII table

Dec	Hx	Oct	Char		Dec	Hx	Oct	Html	Chr	Dec	Hx	Oct	Html	Chr	Dec	Hx	Oct	Html	Chr
0	0	000	NUL	(null)	32	20	040		Space	64	40	100	@	@	96	60	140	`	`
1	1	001	SOH	(start of heading)	33	21	041	!	!	65	41	101	A	A	97	61	141	a	a
2	2	002	STX	(start of text)	34	22	042	"	"	66	42	102	B	B	98	62	142	b	b
3	3	003	ETX	(end of text)	35	23	043	#	#	67	43	103	C	C	99	63	143	c	c
4	4	004	EOT	(end of transmission)	36	24	044	$	$	68	44	104	D	D	100	64	144	d	d
5	5	005	ENQ	(enquiry)	37	25	045	%	%	69	45	105	E	E	101	65	145	e	e
6	6	006	ACK	(acknowledge)	38	26	046	&	&	70	46	106	F	F	102	66	146	f	f
7	7	007	BEL	(bell)	39	27	047	'	'	71	47	107	G	G	103	67	147	g	g
8	8	010	BS	(backspace)	40	28	050	(	(	72	48	110	H	H	104	68	150	h	h
9	9	011	TAB	(horizontal tab)	41	29	051	)	)	73	49	111	I	I	105	69	151	i	i
10	A	012	LF	(NL line feed, new line)	42	2A	052	*	*	74	4A	112	J	J	106	6A	152	j	j
11	B	013	VT	(vertical tab)	43	2B	053	+	+	75	4B	113	K	K	107	6B	153	k	k
12	C	014	FF	(NP form feed, new page)	44	2C	054	,	,	76	4C	114	L	L	108	6C	154	l	l
13	D	015	CR	(carriage return)	45	2D	055	-	-	77	4D	115	M	M	109	6D	155	m	m
14	E	016	SO	(shift out)	46	2E	056	.	.	78	4E	116	N	N	110	6E	156	n	n
15	F	017	SI	(shift in)	47	2F	057	/	/	79	4F	117	O	O	111	6F	157	o	o
16	10	020	DLE	(data link escape)	48	30	060	0	0	80	50	120	P	P	112	70	160	p	p
17	11	021	DC1	(device control 1)	49	31	061	1	1	81	51	121	Q	Q	113	71	161	q	q
18	12	022	DC2	(device control 2)	50	32	062	2	2	82	52	122	R	R	114	72	162	r	r
19	13	023	DC3	(device control 3)	51	33	063	3	3	83	53	123	S	S	115	73	163	s	s
20	14	024	DC4	(device control 4)	52	34	064	4	4	84	54	124	T	T	116	74	164	t	t
21	15	025	NAK	(negative acknowledge)	53	35	065	5	5	85	55	125	U	U	117	75	165	u	u
22	16	026	SYN	(synchronous idle)	54	36	066	6	6	86	56	126	V	V	118	76	166	v	v
23	17	027	ETB	(end of trans. block)	55	37	067	7	7	87	57	127	W	W	119	77	167	w	w
24	18	030	CAN	(cancel)	56	38	070	8	8	88	58	130	X	X	120	78	170	x	x
25	19	031	EM	(end of medium)	57	39	071	9	9	89	59	131	Y	Y	121	79	171	y	y
26	1A	032	SUB	(substitute)	58	3A	072	:	:	90	5A	132	Z	Z	122	7A	172	z	z
27	1B	033	ESC	(escape)	59	3B	073	;	;	91	5B	133	[	[	123	7B	173	{	{
28	1C	034	FS	(file separator)	60	3C	074	<	<	92	5C	134	\	\	124	7C	174			|
29	1D	035	GS	(group separator)	61	3D	075	=	=	93	5D	135	]	]	125	7D	175	}	}
30	1E	036	RS	(record separator)	62	3E	076	>	>	94	5E	136	^	^	126	7E	176	~	~
31	1F	037	US	(unit separator)	63	3F	077	?	?	95	5F	137	_	_	127	7F	177	&#127;	DEL

Source: www.LookupTables.com

ASCII table (continued)

128	Ç	144	É	160	á	176	░	192	└	208	╨	224	α	240	≡
129	ü	145	æ	161	í	177	▒	193	┴	209	╤	225	ß	241	±
130	é	146	Æ	162	ó	178	▓	194	┬	210	╥	226	Γ	242	≥
131	â	147	ô	163	ú	179	│	195	├	211	╙	227	π	243	≤
132	ä	148	ö	164	ñ	180	┤	196	─	212	╘	228	Σ	244	⌠
133	à	149	ò	165	Ñ	181	╡	197	┼	213	╒	229	σ	245	⌡
134	å	150	û	166	ª	182	╢	198	╞	214	╓	230	µ	246	÷
135	ç	151	ù	167	º	183	╖	199	╟	215	╫	231	τ	247	≈
136	ê	152	ÿ	168	¿	184	╕	200	╚	216	╪	232	Φ	248	°
137	ë	153	Ö	169	⌐	185	╣	201	╔	217	┘	233	Θ	249	∙
138	è	154	Ü	170	¬	186	║	202	╩	218	┌	234	Ω	250	·
139	ï	155	¢	171	½	187	╗	203	╦	219	█	235	δ	251	√
140	î	156	£	172	¼	188	╝	204	╠	220	▄	236	∞	252	ⁿ
141	ì	157	¥	173	¡	189	╜	205	═	221	▌	237	φ	253	²
142	Ä	158	₧	174	«	190	╛	206	╬	222	▐	238	ε	254	■
143	Å	159	ƒ	175	»	191	┐	207	╧	223	▀	239	∩	255	

Source: www.LookupTables.com

Windows-1252 encoding

- Windows-1252 (Misnamed ANSI) or CP1252 is a character set, historically used by default on the operating system Microsoft Windows in English and in the main languages of Western Europe, including French.

```
................
................
 !"#$%&'()*+,-./
0123456789:;<=>?
@ABCDEFGHIJKLMNO
PQRSTUVWXYZ[\]^_
`abcdefghijklmno
pqrstuvwxyz{|}~.
€.‚ƒ„…†‡ˆ‰Š‹Œ.Ž.
.‘’“”•–—˜™š›œ.žŸ
 ¡¢£¤¥¦§¨©ª«¬-®¯
°±²³´µ¶·¸¹º»¼½¾¿
ÀÁÂÃÄÅÆÇÈÉÊËÌÍÎÏ
ÐÑÒÓÔÕÖ×ØÙÚÛÜÝÞß
àáâãäåæçèéêëìíîï
ðñòóôõö÷øùúûüýþÿ
```

The text

- formats
  - Combining physical and logical structure
    - DocX, HTML, RTF, ...
  - Separation between physical and logical structure
    - XML, Latex...;

The text

Verlaine....

Fichier Edition Format Affichage ?

Chanson d'automne

Les sanglots longs
Des violons
De l'automne
Blessent mon coeur
D'une langueur
Monotone.

Tout suffocant
Et blême, quand
Sonne l'heure,
Je me souviens
Des jours anciens
Et je pleure

Et je m'en vais
Au vent mauvais
Qui m'emporte
Deçà, delà,
Pareil à la
Feuille morte.

Verlaine.docx [Mode de compatibilité] - Word

FICHIER ACCUEIL INSERTIO CRÉATIO MISE EN P RÉFÉRENC PUBLIPOS RÉVISION AFFICHAG EGYED-ZS...

Coller Times New Roman 24 Styles Modification

Presse-papiers Police Paragraphe Style

Chanson d'autom

Les sanglots longs
Des violons
De l'automne
Blessent mon coeur
D'une langueur
Monotone.

Tout suffocant
Et blême, quand
Sonne l'heure,
Je me souviens
Des jours anciens
Et je pleure

Et je m'en vais
Au vent mauvais
Qui m'emporte
Deçà, delà,
Pareil à la
Feuille morte.

PAGE 1 SUR 1 47 MOTS FRANÇAIS (FRANCE)

Verlaine

file:///D:/Elod/2014-2015/IF/3IF/xml/exer

Save to Mendeley

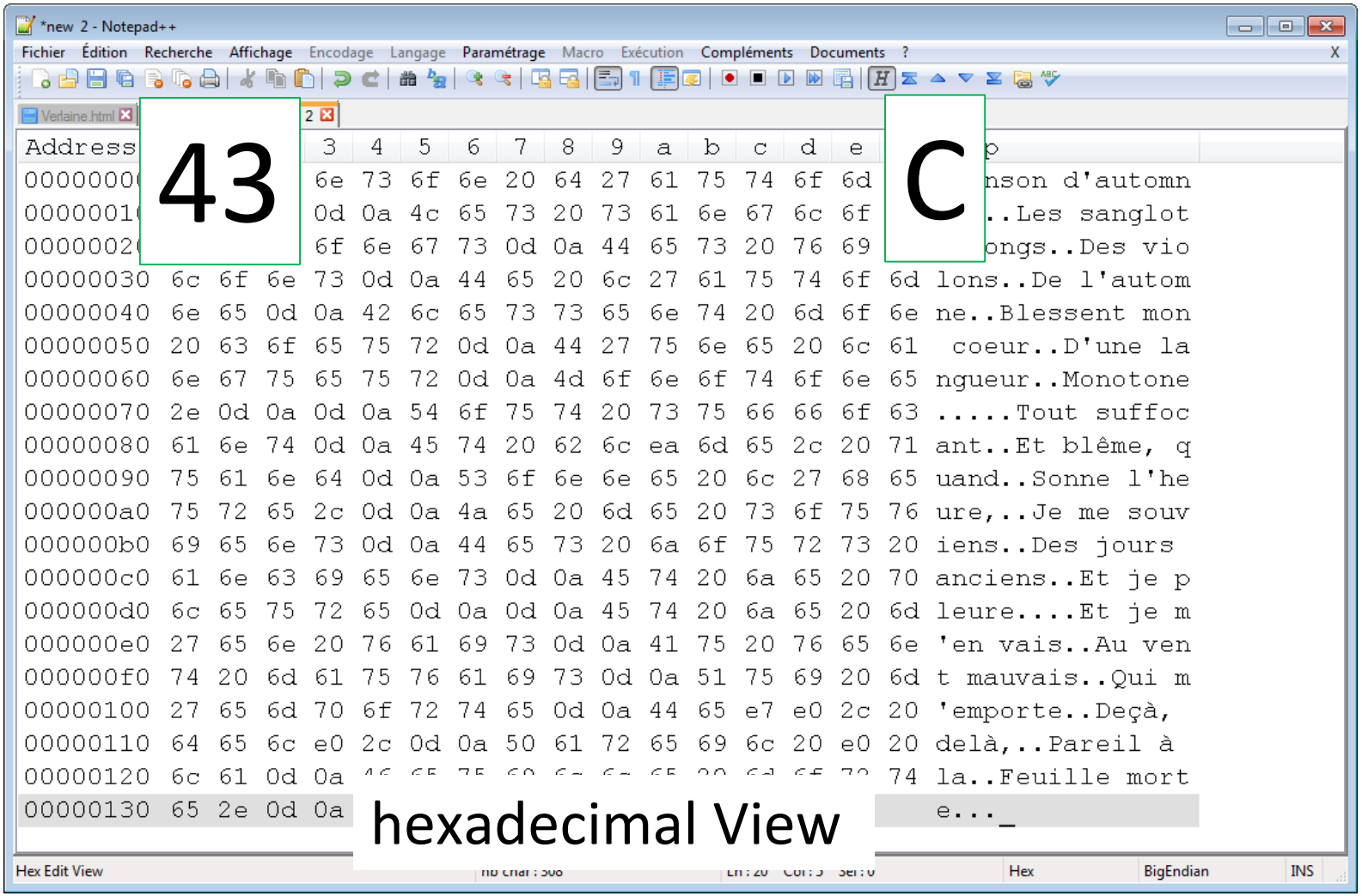
Chanson d'automne

Les sanglots longs
Des violons
De l'automne
Blessent mon coeur
D'une langueur
Monotone.

Tout suffocant
Et blême, quand
Sonne l'heure,
Je me souviens
Des jours anciens
Et je pleure

Et je m'en vais
Au vent mauvais
Qui m'emporte
Deçà, delà,
Pareil à la
Feuille morte.

The text

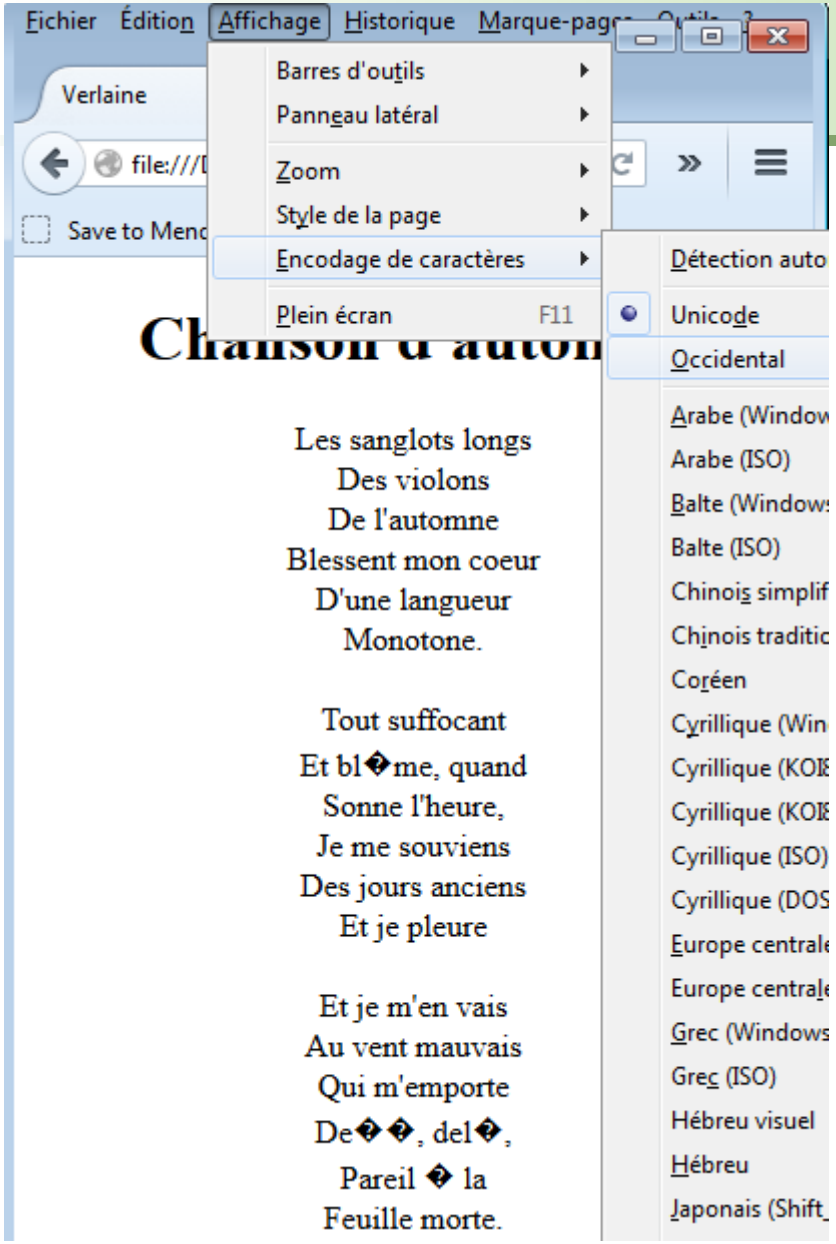

*new 2 - Notepad++
Fichier Édition Recherche Affichage Encodage Langage Paramétrage Macro Exécution Compléments Documents ?
Verlaine.html
43
C
Address 3 4 5 6 7 8 9 a b c d e
00000000 6e 73 6f 6e 20 64 27 61 75 74 6f 6d nson d'automn
00000010 0d 0a 4c 65 73 20 73 61 6e 67 6c 6f ..Les sanglot
00000020 6f 6e 67 73 0d 0a 44 65 73 20 76 69 ongs..Des vio
00000030 6c 6f 6e 73 0d 0a 44 65 20 6c 27 61 75 74 6f 6d lons..De l'autom
00000040 6e 65 0d 0a 42 6c 65 73 73 65 6e 74 20 6d 6f 6e ne..Blessent mon
00000050 20 63 6f 65 75 72 0d 0a 44 27 75 6e 65 20 6c 61 coeur..D'une la
00000060 6e 67 75 65 75 72 0d 0a 4d 6f 6e 6f 74 6f 6e 65 ngueur..Monotone
00000070 2e 0d 0a 0d 0a 54 6f 75 74 20 73 75 66 66 6f 63 .....Tout suffoc
00000080 61 6e 74 0d 0a 45 74 20 62 6c ea 6d 65 2c 20 71 ant..Et blême, q
00000090 75 61 6e 64 0d 0a 53 6f 6e 6e 65 20 6c 27 68 65 uand..Sonne l'he
000000a0 75 72 65 2c 0d 0a 4a 65 20 6d 65 20 73 6f 75 76 ure,..Je me souv
000000b0 69 65 6e 73 0d 0a 44 65 73 20 6a 6f 75 72 73 20 iens..Des jours
000000c0 61 6e 63 69 65 6e 73 0d 0a 45 74 20 6a 65 20 70 anciens..Et je p
000000d0 6c 65 75 72 65 0d 0a 0d 0a 45 74 20 6a 65 20 6d leure....Et je m
000000e0 27 65 6e 20 76 61 69 73 0d 0a 41 75 20 76 65 6e 'en vais..Au ven
000000f0 74 20 6d 61 75 76 61 69 73 0d 0a 51 75 69 20 6d t mauvais..Qui m
00000100 27 65 6d 70 6f 72 74 65 0d 0a 44 65 e7 e0 2c 20 'emporte..Deçà,
00000110 64 65 6c e0 2c 0d 0a 50 61 72 65 69 6c 20 e0 20 delà,..Pareil à
00000120 6c 61 0d 0a la..Feuille mort
00000130 65 2e 0d 0a e...
hexadecimal View
Hex Edit View
Hex
BigEndian
INS

The text

```html
<!DOCTYPE html PUBLIC "-//W3C//DTD HTML 4.01//EN" "http://www.w3.org/TR/html4/strict.dtd">
<html>
<head>
<meta content="text/html; charset=ANSI" http-equiv="content-type">
<title>Verlaine</title>
</head>
<body>
<h1 align="center">Chanson d'automne</h1>
<p align="center" class="last">Les sanglots longs<br/>
  Des violons<br/>
  De l'automne<br/>
  Blessent mon coeur<br/>
  D'une langueur<br/>
  Monotone.<br/>
  <br/>
  Tout suffocant<br/>
  Et blême, quand<br/>
  Sonne l'heure,<br/>
  Je me souviens<br/>
  Des jours anciens<br/>
  Et je pleure<br/>
  <br/>
  Et je m'en vais<br/>
  Au vent mauvais<br/>
  Qui m'emporte<br/>
  Deçà, delà,<br/>
  Pareil à la<br/>
  Feuille morte.<br/>
</p>
</body>
</html>
```

HTML source

The text


Fichier
Édition
Affichage
Historique
Marque-pages
Verlaine
file:///
Save to Menc
Barres d'outils
Panneau latéral
Zoom
Style de la page
Encodage de caractères
Plein écran F11
Détection auto
Unicode
Occidental
Arabe (Window
Arabe (ISO)
Balte (Windows
Balte (ISO)
Chinois simplif
Chinois traditic
Coréen
Cyrillique (Win
Cyrillique (KOI8
Cyrillique (KOI8
Cyrillique (ISO)
Cyrillique (DOS
Europe centrale
Europe centrale
Grec (Windows
Grec (ISO)
Hébreu visuel
Hébreu
Japonais (Shift_
Les sanglots longs
Des violons
De l'automne
Blessent mon coeur
D'une langueur
Monotone.
Tout suffocant
Et bl►me, quand
Sonne l'heure,
Je me souviens
Des jours anciens
Et je pleure
Et je m'en vais
Au vent mauvais
Qui m'emporte
De►►, del►,
Pareil ► la
Feuille morte.

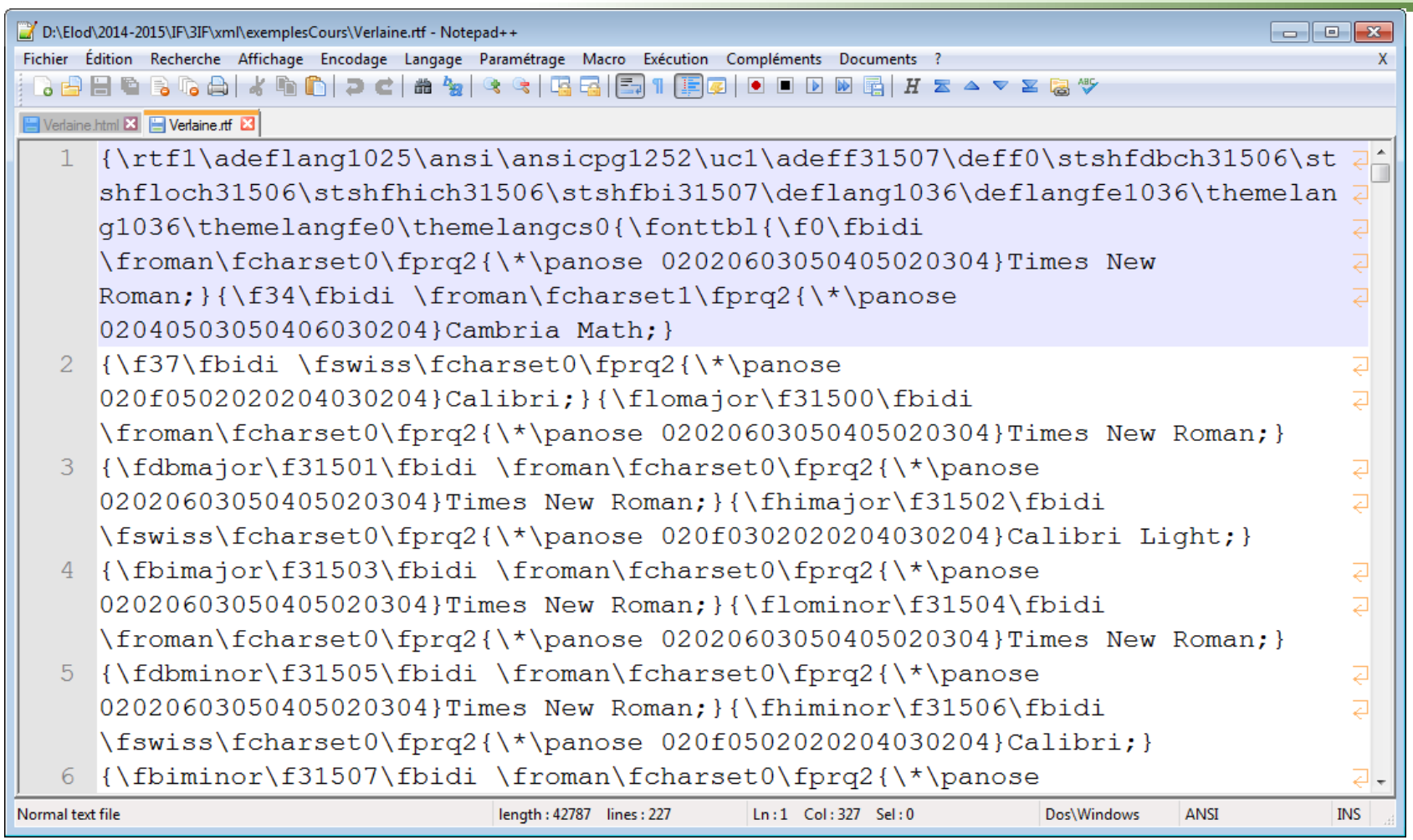
The text

Verlaine.xml

```xml
<!DOCTYPE poeme SYSTEM "poeme.dtd">
<poeme>
  <titre>Chanson d'automne</titre>
  <auteur>Paul Verleine</auteur>
  <strophe>
    <vers>Les sanglots longs</vers>
    <vers>Des violons</vers>
    <vers>De l'automne</vers>
    <vers>Blessent mon coeur</vers>
    <vers>D'une langueur</vers>
    <vers>Monotone.</vers>
  </strophe>
  <strophe>
    <vers>Tout suffocant</vers>
    <vers>Et blême, quand</vers>
    <vers>Sonne l'heure,</vers>
    <vers>Je me souviens</vers>
    <vers>Des jours anciens</vers>
    <vers>Et je pleure</vers>
  </strophe>
  <strophe>
    <vers>Et je m'en vais</vers>
    <vers>Au vent mauvais</vers>
    <vers>Qui m'emporte</vers>
    <vers>Deçà, delà,</vers>
    <vers>Pareil à la</vers>
    <vers>Feuille morte.</vers>
  </strophe>
</poeme>
```

XML

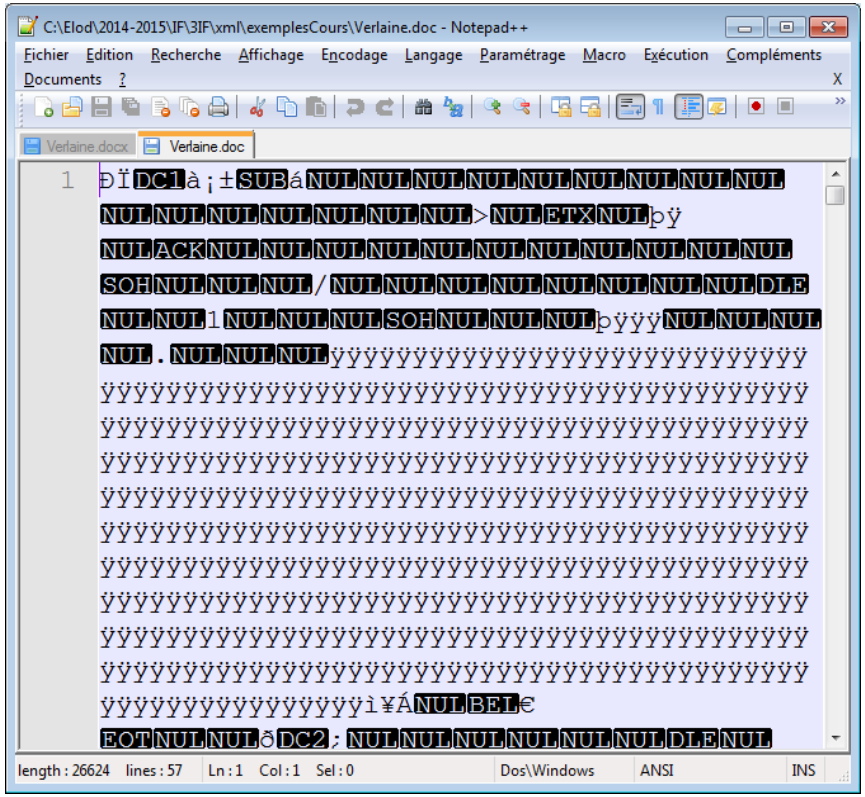
The text

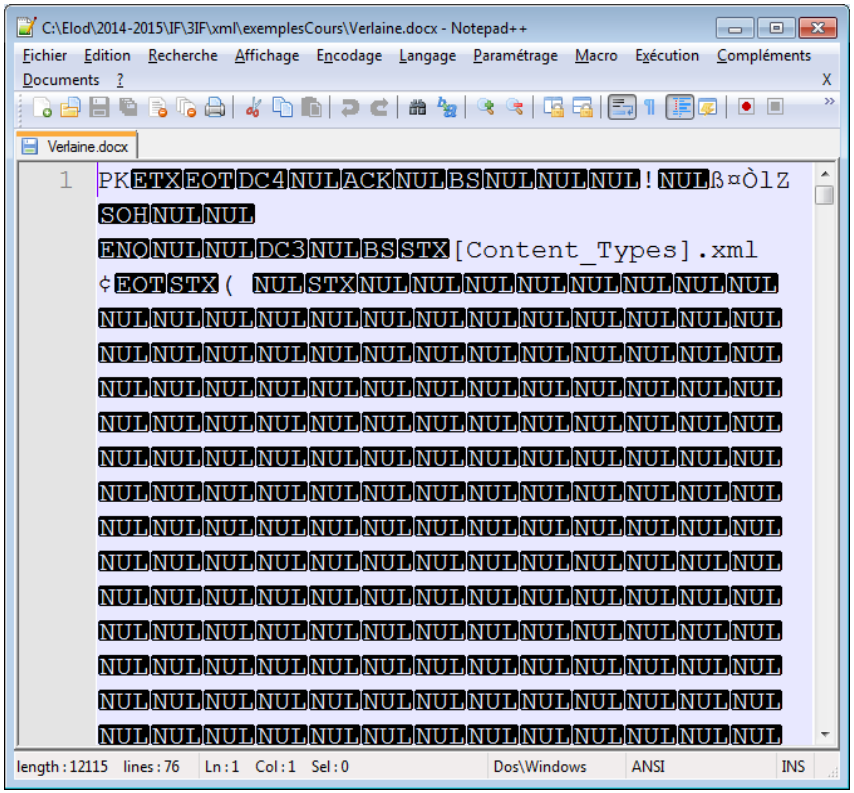

D:\Elod\2014-2015\IF\3IF\xml\exemplesCours\Verlaine.rtf - Notepad++
Fichier Édition Recherche Affichage Encodage Langage Paramétrage Macro Exécution Compléments Documents ?
Verlaine.html
Verlaine.rtf
1 {\rtf1\adeflang1025\ansi\ansicpg1252\uc1\adeff31507\deff0\stshfdbch31506\stshfloch31506\stshfhich31506\stshfbi31507\deflang1036\deflangfe1036\themelang1036\themelangfe0\themelangcs0{\fonttbl{\f0\fbidi \froman\fcharset0\fprq2{*\panose 02020603050405020304}Times New Roman;}{\f34\fbidi \froman\fcharset1\fprq2{*\panose 02040503050406030204}Cambria Math;}
2 {\f37\fbidi \fswiss\fcharset0\fprq2{*\panose 020f0502020204030204}Calibri;}{\flomajor\f31500\fbidi \froman\fcharset0\fprq2{*\panose 02020603050405020304}Times New Roman;}
3 {\fdbmajor\f31501\fbidi \froman\fcharset0\fprq2{*\panose 02020603050405020304}Times New Roman;}{\fhimajor\f31502\fbidi \fswiss\fcharset0\fprq2{*\panose 020f0302020204030204}Calibri Light;}
4 {\fbimajor\f31503\fbidi \froman\fcharset0\fprq2{*\panose 02020603050405020304}Times New Roman;}{\flominor\f31504\fbidi \froman\fcharset0\fprq2{*\panose 02020603050405020304}Times New Roman;}
5 {\fdbminor\f31505\fbidi \froman\fcharset0\fprq2{*\panose 02020603050405020304}Times New Roman;}{\fhiminor\f31506\fbidi \fswiss\fcharset0\fprq2{*\panose 020f0502020204030204}Calibri;}
6 {\fbiminor\f31507\fbidi \froman\fcharset0\fprq2{*\panose
Normal text file
length : 42787 lines : 227
Ln : 1 Col : 327 Sel : 0
Dos\Windows
ANSI
INS

RTF

The text

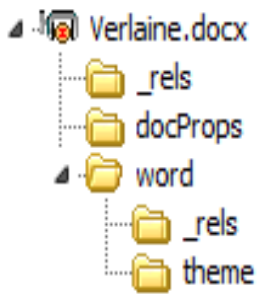
Format .doc et .docx





The text

Format .docx

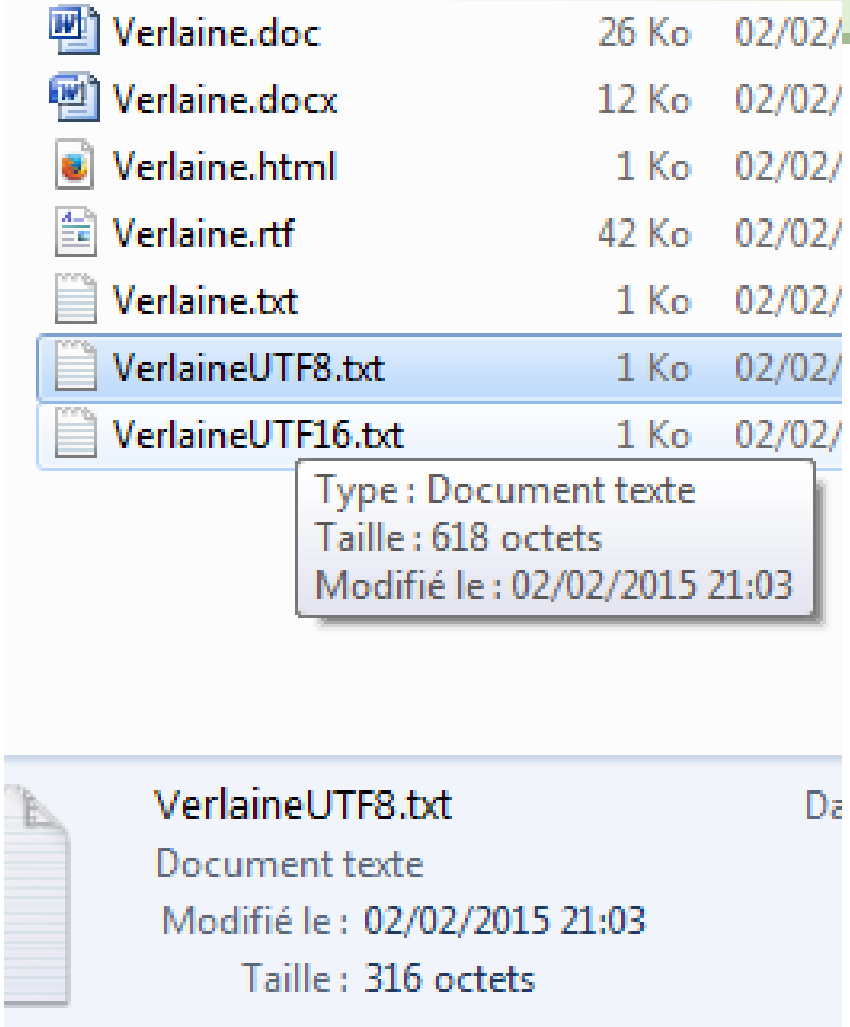


Nom	Type	Date de m...	Taille	Ratio
theme	Folder			
_rels	Folder			
document.xml	Document XML	01/01/1980	5,957	85%
fontTable.xml	Document XML	01/01/1980	1,261	64%
settings.xml	Document XML	01/01/1980	2,120	58%
styles.xml	Document XML	01/01/1980	30,192	90%
webSettings.xml	Document XML	01/01/1980	913	58%

Fragment of document.xml

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<w:document mc:Ignorable="w14 w15 w16se wp14"
xmlns:wpc="http://schemas.microsoft.com/office/word/2010/
......
        <w:body>
                <w:p w:rsidR="00123C86" w:rsidRPr="00123C86"
w:rsidRDefault="00123C86" w:rsidP="0089176C">
        <w:r w:rsidRPr="00123C86">
                                <w:rPr>
                                        <w:rFonts w:ascii="Times New
Roman" w:eastAsia="Times New Roman" w:hAnsi="Times New Roman"
w:cs="Times New Roman"/>
                                        <w:b/>
......
                                </w:rPr>
                                <w:t>Chanson d'automne</w:t>
                        </w:r>
```

The text



File sizes

Plan

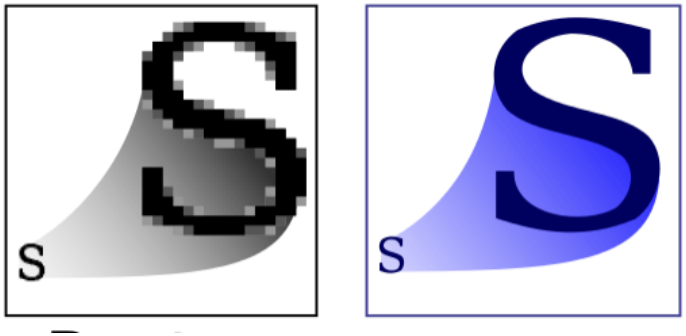
- Introduction
  - Reminders of BD
  - The documents
    - Introduction
    - Document modeling
    - hyperdocuments
    - Document types (text, image, …)
- Core XML
- XML galaxy
- NOSQL
- Conclusion

Different media: image

- 2D representation of the world (photo) or the imagination of a person (drawing).

- physical representation (raw data) 2 formats:
  - pixel array
  - Vector representation (set of graphics primitives)

- Logical representation:
  - The external characteristics (author ...)
  - The internal features (objects, color ...)



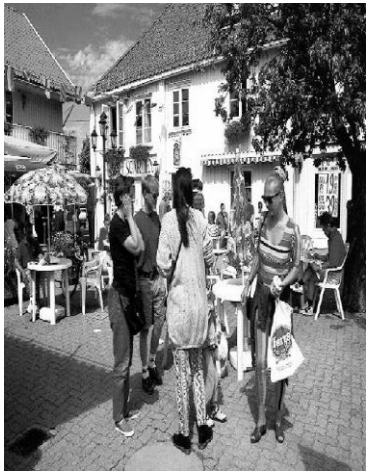
Bitmap

Vector

raster (bitmap)

The data of each pixel are stored in a space of p dimensions, coded on M values.

p = 1, M = 2



p = 1, M = 256

source: (F. Lebourgeois)

raster (bitmap)

The data of each pixel are stored in a space of p dimensions, coded on M values.

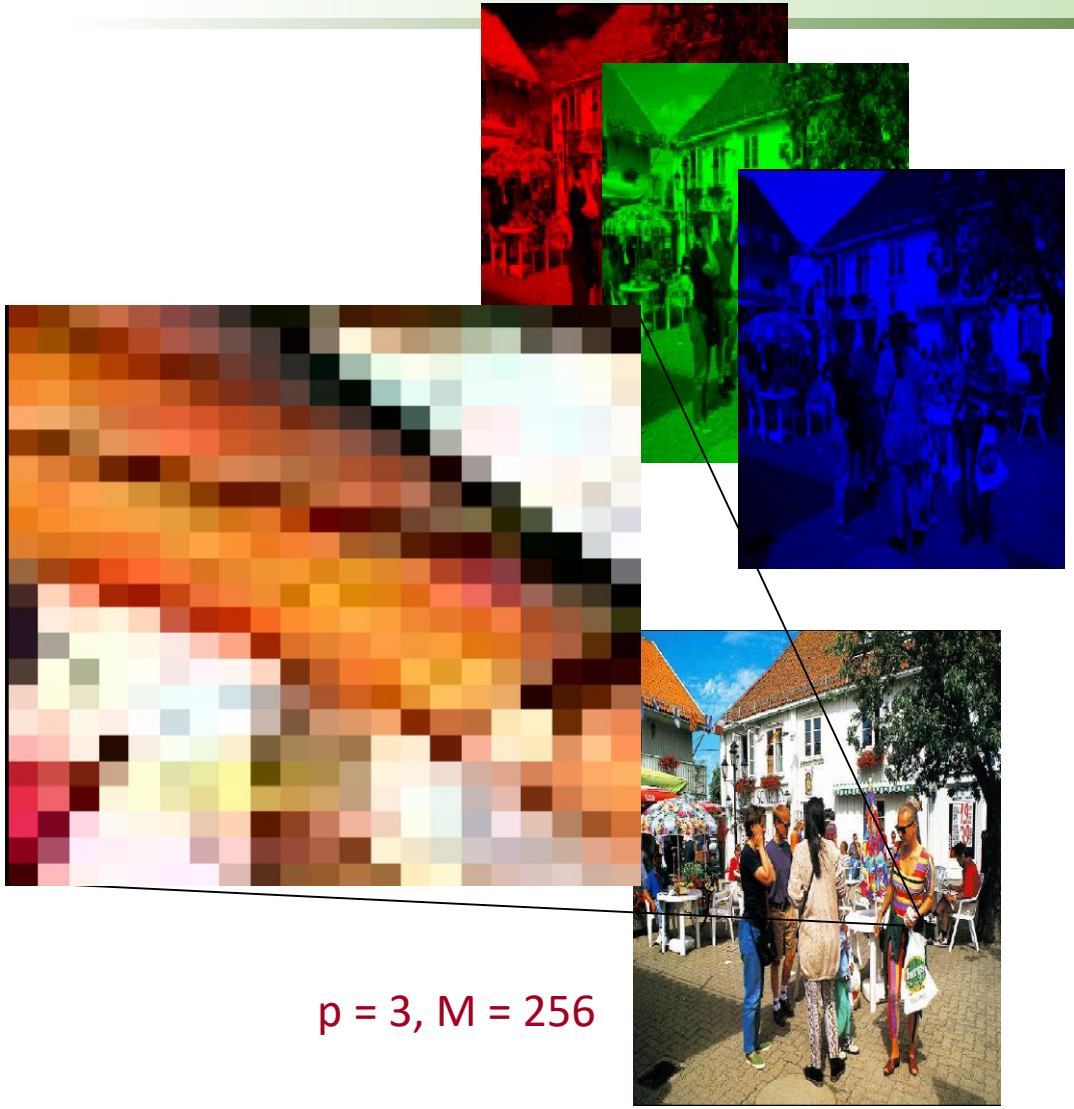


Image credit : F. Lebourgeois

p = 3, M = 256

Colors

The additive synthesis of light, or RGB:

The image is obtained by superimposing three light radiation: **red** (**R**), the **green** (**G**) and the **blue** (**B**). In the case of a cathodic screen, these three radiations are obtained by bombarding the photosensitive phosphor screen.

RGB mode:

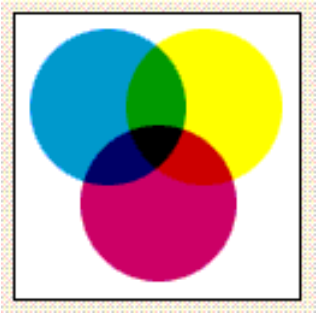
A RGB picture consists in the sum of three light rays red, green and blue whose beams are superimposed. At maximum intensity they produce a white light beam.

Colors

Subtractive color synthesis, or CMYK

In color printing, the usual primary colors are Cyan, Magenta and Yellow (CMY). Cyan is the complement of red, meaning that the cyan serves as a filter that absorbs red. Magenta is the complement of green, and yellow the complement of blue. Combinations of different amounts of the three can produce a wide range of colors with good saturation.

The inks deposited on the paper act as filters that absorb light. Their superimposition should theoretically produce a total black: no more light which is not the case in practice. In inkjet color printing and typical mass production photomechanical printing processes, a black ink K (Key) component is included



Colors

Conversion in the spectral domain and removing high frequencies (details)

Jpeg file original size 113kb

Colors

Conversion in the spectral domain and removing high frequencies (details)

Jpeg 22% coefficients, size 35ko

Colors

Conversion in the spectral domain and removing high frequencies (details)

Jpeg 3% coefficients Size 7kb

Vector image

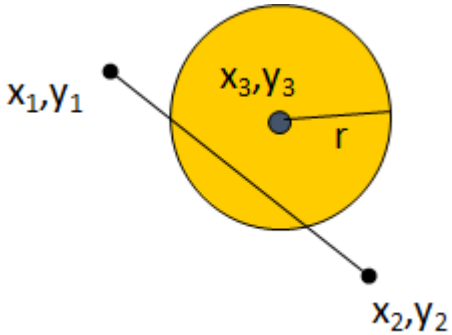
- In a **vector image** data are represented by simple geometric shapes that are described from a mathematical point of view.

- *for example : a circle is described by information like (position of the center, radius).*

  These images are mainly used to make drawings or plans.
  Industrial design software works on this principle;

  Word processing or desktop publishing (desktop publishing) also offer such tools.

- *(Ps, pdf, CorelDraw, Adobe Illustrator, SVG...)*



- *These images show 3 advantages :*
  - they need little space in memory and
  - they can be resized without information loss.
  - permit independent manipulation of different parts

Conclusion

Document modeling is important

difference between logical/physical structure

The digital representation of documents has big effect on document based databases.