Data Interoperability and Semantics

< Part 2. Data Formats >

ICM – Toolbox Engineering and Interoperability of Software Systems – Course unit on Data Interoperability and Semantics
M1 Cyber Physical and Social Systems – Course unit on Data Interoperability and Semantics
Maxime Lefrançois https://maxime-lefrancois.info
Course unit URL: https://ci.mines-stetienne.fr/cps2/course/data

Data Interoperability and Semantics
Outline

- < Part 2. Data formats >
  - Part 2.1 Generalities
  - Part 2.2. Delimiter-separated values
  - Part 2.3. Extensible Markup Language (XML)
  - Part 2.4. JavaScript Object Notation (JSON)
  - Part 2.5. Configuration file formats
  - Part 2.6. YAML Ain’t Markup Language (YAML)
  - Part 2.7. Lightweight markup languages
  - Part 2.8. Compressed formats
  - Part 2.9 Multimedia formats
  - Part 2.10 3D models

Data Interoperability and Semantics

Part 2. Data Formats

Part 2.1 Generalities

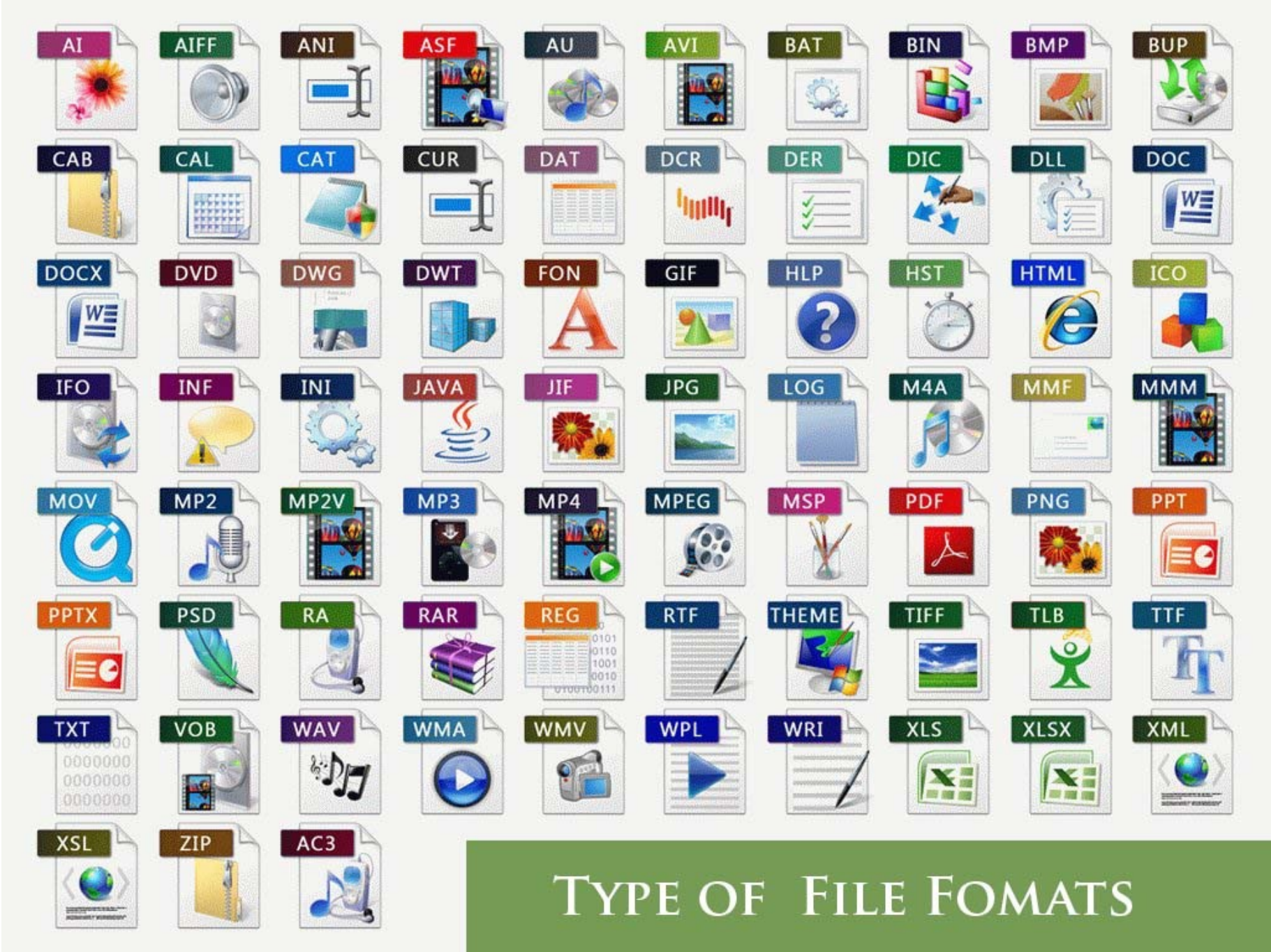
ICM – Toolbox Engineering and Interoperability of Software Systems – Course unit on Data Interoperability and Semantics
M1 Cyber Physical and Social Systems – Course unit on Data Interoperability and Semantics
Maxime Lefrançois https://maxime-lefrancois.info
Course unit URL: https://ci.mines-stetienne.fr/cps2/course/data

File format

standard way that information is encoded for transfer or storage in a computer file

A format is what enables an application to interpret the raw data contained in a file. It is the mode of representation of these data

File formats are marked in the extension of the file name



Contents [hide]

ToC of https://en.wikipedia.org/wiki/List_of_file_formats

Proprietary file formats vs free file formats

Proprietary file format

Data encoding, s.t. one needs the company's software to decode and interpret it.
Either the data encoding specification is secret, or restricted through license.

Examples of proprietary file formats

- **mp3**: open standard, but subject to patents in some countries
- **dwg**: non documented, AutoCAD
- **psd**: documented, Adobe Photoshop's native image format
- **rar**: partially documented, archive and compression file format owned by Alexander L. Roshal
- **zip** (newest versions have patented features)
- **gif**: CompuServe's Graphics Interchange Format, **patent expired** in 2004
- **pdf**: open since 2008 ISO 320000-1. Still some features proprietary by Adobe (forms, scripts)
- **doc**, **xls**, **ppt**: formerly closed/undocumented, now Microsoft Open Specification Promise

https://en.wikipedia.org/wiki/Proprietary_format

Open vs unpublished file formats

Open file format

published specification usually maintained by a standards organization.

Linux Information Project: *“any format that is published for anyone to read and study but which may or may not be encumbered by patents, copyrights or other restrictions on use”*

US government: *“An open format is one that is platform independent, machine readable, and made available to the public without restrictions that would impede the re-use of that information”*

FR government: loi n° 2004-575 du 21 juin 2004 pour « la confiance dans l’économie numérique »:
« On entend par standard ouvert tout protocole de communication, d’interconnexion ou d’échange et tout format de données interopérable et dont les spécifications techniques sont publiques et sans restriction d’accès ni de mise en œuvre. »

Open formats, what for?

So that users might read my documents unhindered

Users exchanging reports.

So that your documents might be read more easily by other people, without you having to worry about which software they use, choose open formats.

CLOSED FORMAT, **IDENTICAL** SOFTWARE

Alice uses the software program "Carcera[1]." She records her report in a closed format (one that does not permit interoperability), then sends it to Bob, who has the same software program. He can read the document, modify it and send it back to Alice.

PROGRAMS WITH **CLOSED FORMATS**, **DIFFERENT** SOFTWARE

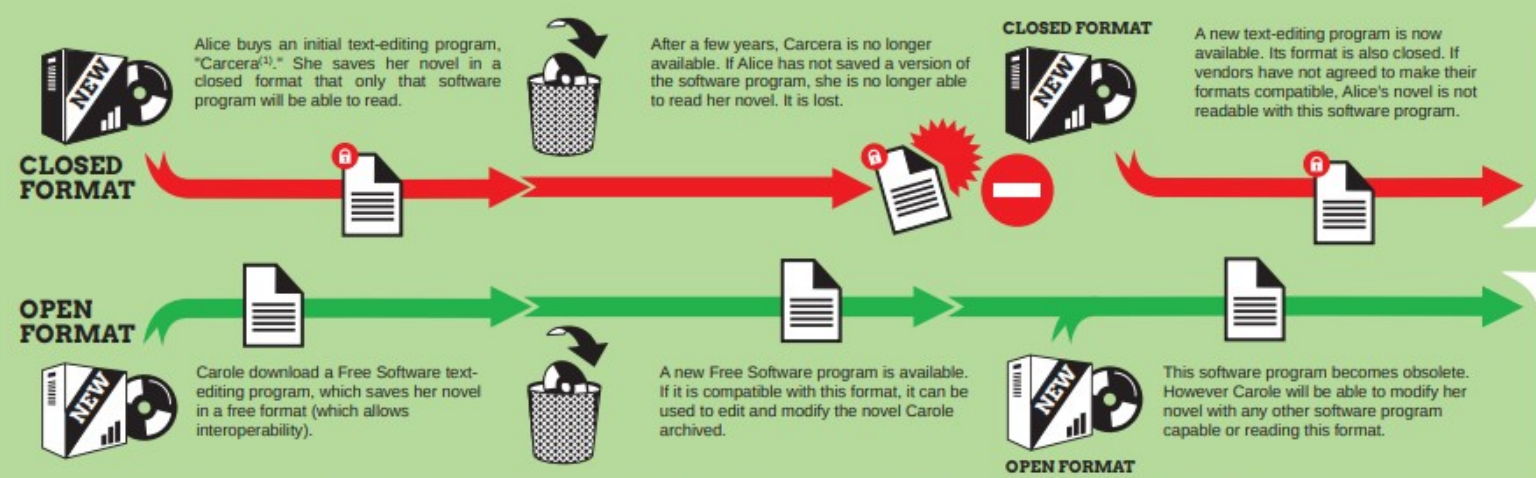
The following day, Alice sends her report to Albert. He doesn't have the same software program, which refuses to open the document. Albert has no other choice than to acquire the Carcera software used by Alice, with the hope it is compatible with its computer.

PROGRAMS WITH **OPEN FORMATS**, **DIFFERENT** SOFTWARE

Carole, another user, choses to record her report in open format (allowing for interoperability) and sends it to David. David can read the document, modify and record it, either by using the same open format software or by using another interoperable software.

To ensure the longevity of my documents

Alice and Carole use text editors to write, save, and preserve a novel.



The availability and longevity of your documents, saved in a closed format, depends on the decisions of software vendors.

In your interest, choose software programs that save your documents in open formats. They are not dependent on any particular software program.

To learn more, go to www.april.org. Documented created by April using Free Software. Design: Antoine Bardelli. License: Free Art License 1.3 or later / Creative Commons Attribution-ShareAlike 2.0 or later / GNU Free Documentation License 1.3 or later.

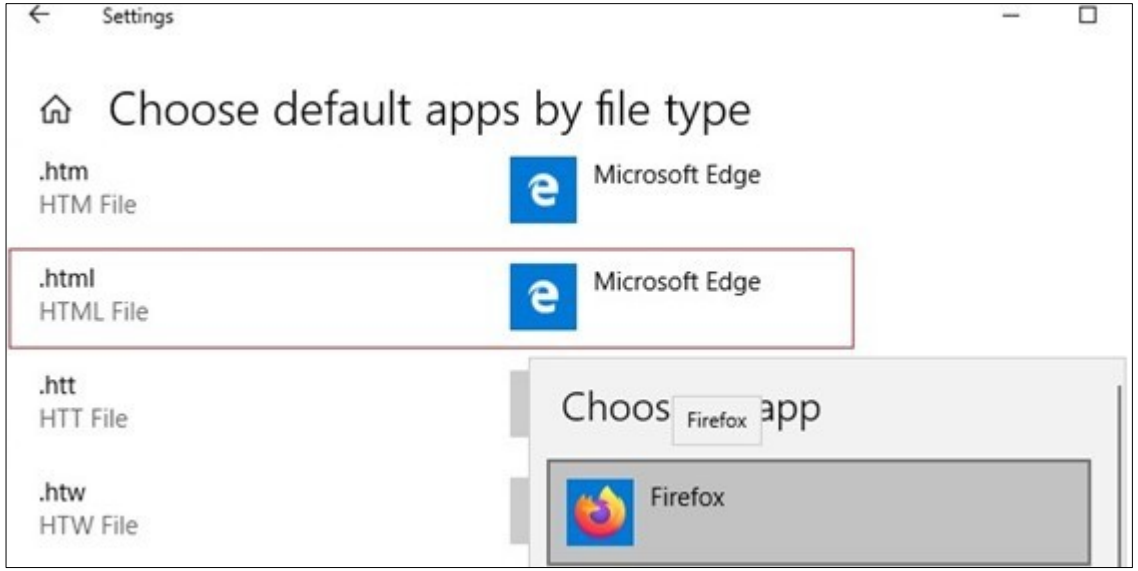
(1) Ficticious proprietary software program name, for illustration purposes.

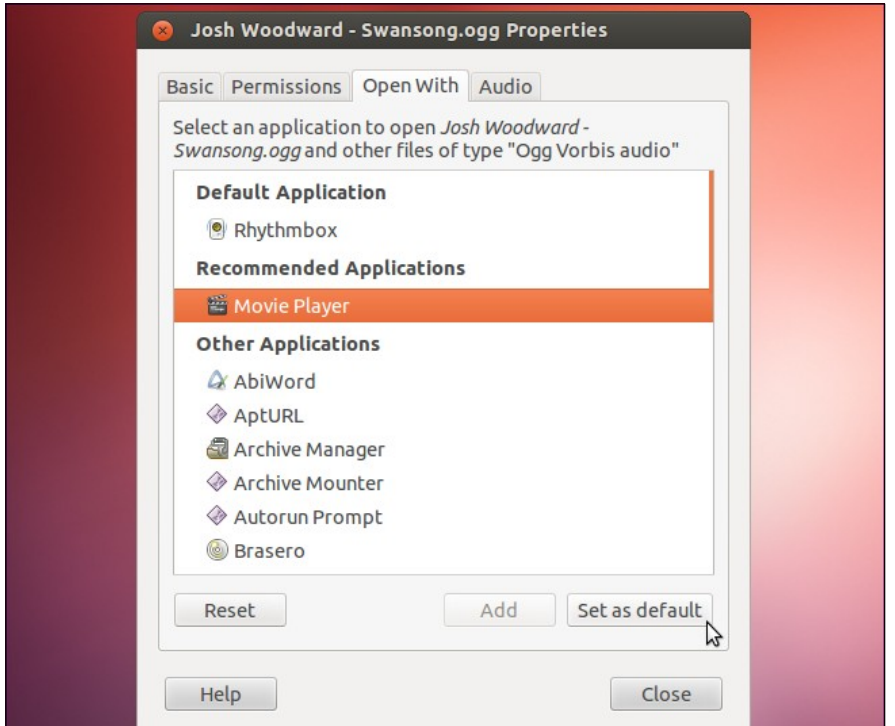
https://www.april.org/en/open-formats

How to identify a file type ?

Filename extensions

- See https://en.wikipedia.org/wiki/List_of_filename_extensions
- Look up or browse categories https://www.file-extension.info/
- Windows / linux desktops: applications association to filename extension





How to identify a file type ?

Filename extensions

- See https://en.wikipedia.org/wiki/List_of_filename_extensions
- Look up or browse categories https://www.file-extension.info/
- Windows / linux desktops: applications association to filename extension
- reason why .htm vs .html ? the 8.3 filename format
- OS hide the extensions + associate applications
  = creates security issues

How to identify a file type ?

File header

may contain metadata about the file and its content

- ex., Exif metadata: image format, size, resolution color space, ...
- ex., AVI header format: https://www.filefix.org/format/avi.html#header

Camera manufacturer	Canon
Camera model	Canon EOS 1200D
Author	Praveen. P
Exposure time	1/60 sec (0.016666666666667)
F-number	f/11
ISO speed rating	200
Date and time of data generation	22:29, 22 November 2018
Lens focal length	41 mm
Show extended details	

ex., Exif metadata
https://en.wikipedia.org/wiki/Exif

Character-based documents may be opened in text editors, and their header interpreted

ex., head of a XML file

```
<?xml version="1.0" encoding="UTF-8"?>
 <!DOCTYPE html
 PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN"
"http://www.w3.org/TR/xhtml1/DTD/xhtml1-strict.dtd">
```

ex., head of a iCalendar file
https://en.wikipedia.org/wiki/ICalendar

```
BEGIN:VCALENDAR
VERSION:2.0
PRODID:-//hacksw/handcal//NONSGML v1.0//EN
BEGIN:VEVENT
```

ex., head of a ISO 10303-21 STEP file
https://en.wikipedia.org/wiki/ISO_10303-21

```
ISO-10303-21;
HEADER;
FILE_DESCRIPTION(
/* description */ ('A minimal AP214 example wit
/* implementation_level */ '2;1');
FILE_NAME(
/* name */ 'demo',
/* time_stamp */ '2003-12-27T11:57:53',
/* author */ ('Lothar Klein'),
/* organization */ ('LKSoft'),
```

How to identify a file type ?

File's first bytes: Magic numbers

the first few bytes may be distinctive enough

Hex signature	ISO 8859-1	Offset	Extension	Description
23 21	#!	0		Script or data to be passed to the program following the shebang (#!)
52 49 46 46 ?? ?? ?? ?? 41 56 49 20	RIFF????AVI SP	0	avi	Audio Video Interleave video format
FF FB FF F3 FF F2	ÿû ÿó ÿò	0	mp3	MPEG-1 Layer 3 file without an ID3 tag or with an ID3v1 tag (which is appended at the end of the file)
49 44 33	ID3	0	mp3	MP3 file with an ID3v2 container
21 3C 61 72 63 68 3E 0A	!<arch>LF	0	deb	linux deb file
37 7A BC AF 27 1C	7z¼¯'FS	0	7z	7-Zip File Format
1F 8B	US ‹	0	gz tar.gz	GZIP compressed file[1]
FD 37 7A 58 5A 00	ý7zXZNUL	0	xz tar.xz	XZ compression utility using LZMA2 compression
4E 45 53 1A	NESSUB	0	nes	Nintendo Entertainment System ROM file[2]

Example of file signatures - https://en.wikipedia.org/wiki/List_of_file_signatures

How to identify a file type ?

File's first bytes: Magic numbers

the first few bytes may be distinctive enough

ex., a file that starts with a Byte Order Mark (BOM) tells the system:

- Highly probable that the text stream is Unicode
- Describes the byte order, or endianness, of the text stream: Big-endian (BE) vs Little-endian (LE)
- Distinguish between Unicode character encoding (UTF-8, UTF-16, UTF-32, ...)

Encoding	Representation (hexadecimal)	Representation (decimal)	Bytes as CP1252 characters
UTF-8[a]	EF BB BF	239 187 191	ï»¿
UTF-16 (BE)	FE FF	254 255	þÿ
UTF-16 (LE)	FF FE	255 254	ÿþ
UTF-32 (BE)	00 00 FE FF	0 0 254 255	^@^@þÿ (^@ is the null character)
UTF-32 (LE)	FF FE 00 00	255 254 0 0	ÿþ^@^@ (^@ is the null character)

https://en.wikipedia.org/wiki/Byte_order_mark

How to identify a file type ?

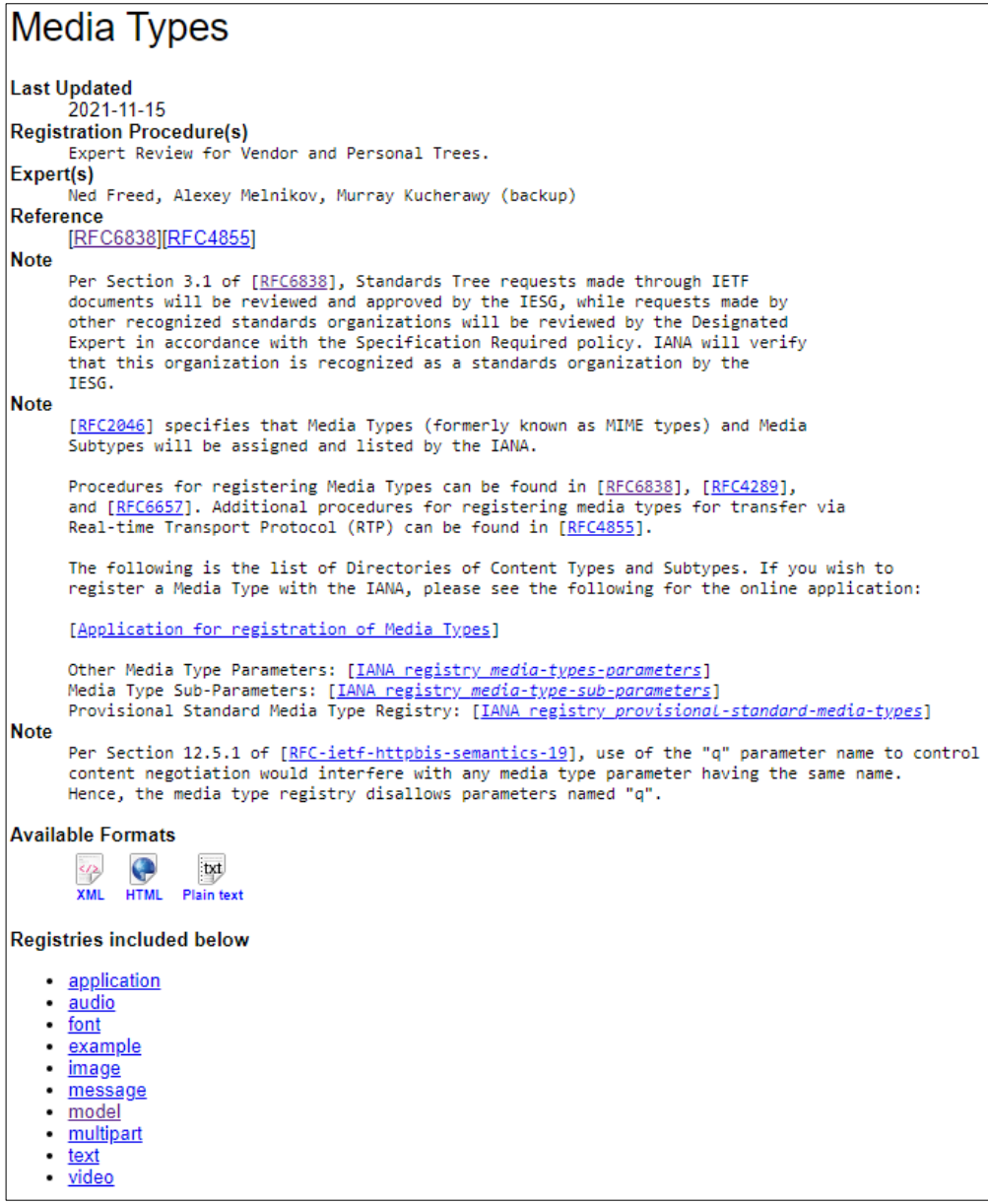
Example: the gzip file format starts with:

- a 10-byte header, containing:
  - magic number (0x1f8b),
  - the compression method (0x08 for DEFLATE),
  - 1-byte of header flags,
  - a 4-byte timestamp,
  - compression flags and
  - the operating system ID.
- optional extra headers as allowed by the header flags, including the original filename, and a comment field,

How to identify a file type ?

MIME type = Media types

- managed by Internet Assigned Numbers Authority (IANA)
- example: text/html, image/gif, font/woff2, ...
- used by many internet protocols
- demo with your browser: with the Network tab of the developer tools open, see HTTP response headers of requests when accessing https://fonts.google.com/



Media Types

Last Updated
2021-11-15
Registration Procedure(s)
Expert Review for Vendor and Personal Trees.
Expert(s)
Ned Freed, Alexey Melnikov, Murray Kucherawy (backup)
Reference
[RFC6838][RFC4855]
Note
Per Section 3.1 of [RFC6838], Standards Tree requests made through IETF documents will be reviewed and approved by the IESG, while requests made by other recognized standards organizations will be reviewed by the Designated Expert in accordance with the Specification Required policy. IANA will verify that this organization is recognized as a standards organization by the IESG.
Note
[RFC2046] specifies that Media Types (formerly known as MIME types) and Media Subtypes will be assigned and listed by the IANA.

Procedures for registering Media Types can be found in [RFC6838], [RFC4289], and [RFC6657]. Additional procedures for registering media types for transfer via Real-time Transport Protocol (RTP) can be found in [RFC4855].

The following is the list of Directories of Content Types and Subtypes. If you wish to register a Media Type with the IANA, please see the following for the online application:

[Application for registration of Media Types]

Other Media Type Parameters: [IANA registry *media-types-parameters*]
Media Type Sub-Parameters: [IANA registry *media-type-sub-parameters*]
Provisional Standard Media Type Registry: [IANA registry *provisional-standard-media-types*]
Note
Per Section 12.5.1 of [RFC-ietf-httpbis-semantics-19], use of the "q" parameter name to control content negotiation would interfere with any media type parameter having the same name. Hence, the media type registry disallows parameters named "q".

Available Formats
XML HTML Plain text

Registries included below

- application
- audio
- font
- example
- image
- message
- model
- multipart
- text
- video

https://www.iana.org/assignments/media-types/media-types.xhtml

iana.org/assignments/media-types/font/woff2

font/woff2

Type name: font

Subtype name: woff2

Required parameters: None

Optional parameters: None

Encoding considerations: Binary

Interoperability considerations: WOFF 2.0 is an improvement on WOFF 1.0. The two formats have different Internet Media Types and different @font-face formats, and they may be used in parallel.

Published specification: This media type registration is extracted from the WOFF 2.0 specification [W3C.CR-WOFF2-20150414] at W3C.

Applications that use this media type: WOFF 2.0 is used by web browsers, often in conjunction with HTML and CSS.

Additional information:

Magic number(s): The signature field in the WOFF header MUST contain the "magic number" 0x774F4632 ('wOF2')

File extension(s): woff2

Macintosh file type code(s): (no code specified)

Macintosh Universal Type Identifier code: "org.w3.woff2"

@font-face Format: woff2

Fragment Identifiers: See Section 4.2.

Person & email address to contact for further information:
Chris Lilley (www-font&w3.org).

Intended usage: COMMON

Restrictions on usage: None

Author: The WOFF2 specification is a work product of the World Wide Web Consortium's WebFonts working group.

Change controller: The W3C has change control over this specification.

iana.org/assignments/media-types/font/woff2

```
Type name:  font

Subtype name:  woff2

Required parameters:  None

Optional parameters:  None

Encoding considerations:  Binary

Interoperability considerations:  WOFF 2.0 is an improvement on WOFF
   1.0.  The two formats have different Internet Media Types and
   different @font-face formats, and they may be used in parallel.

Published specification:  This media type registration is extracted
   from the WOFF 2.0 specification [W3C.CR-WOFF2-20150414] at W3C.

Applications that use this media type:  WOFF 2.0 is used by web
   browsers, often in conjunction with HTML and CSS.

Additional information:

   Magic number(s):  The signature field in the WOFF header MUST
      contain the "magic number" 0x774F4632 ('wOF2')

   File extension(s):  woff2

   Macintosh file type code(s):  (no code specified)

   Macintosh Universal Type Identifier code:  "org.w3.woff2"

   @font-face Format:  woff2

   Fragment Identifiers:  See Section 4.2.

Person & email address to contact for further information:
   Chris Lilley (www-font&w3.org).

Intended usage:  COMMON

Restrictions on usage:  None

Author:  The WOFF2 specification is a work product of the World Wide
   Web Consortium's WebFonts working group.

Change controller:  The W3C has change control over this
   specification.
```

font/woff2

Macintosh file type code(s)

Apple Computer, Inc., "Mac OS: File Type and Creator Codes, and File Formats", Apple Knowledge Article 55381, June 1993, last accessed 17/05/2008
https://web.archive.org/web/20080517021842/http://www.info.apple.com/kbnum/n55381

PICT	picture	Stores a PICT image contained in the file
PREF	preference	Stores the environment settings for an application
snd	sound	Stores a sound used in the file
STR	string	Stores a string or hexadecimal data used in the file
STR#	string list	Stores multiple strings used in the file
styl	style	Defines style information, such as the font, color and size of text
TEXT	text	Stores text

https://en.wikipedia.org/wiki/Resource_fork#Major_resource_types

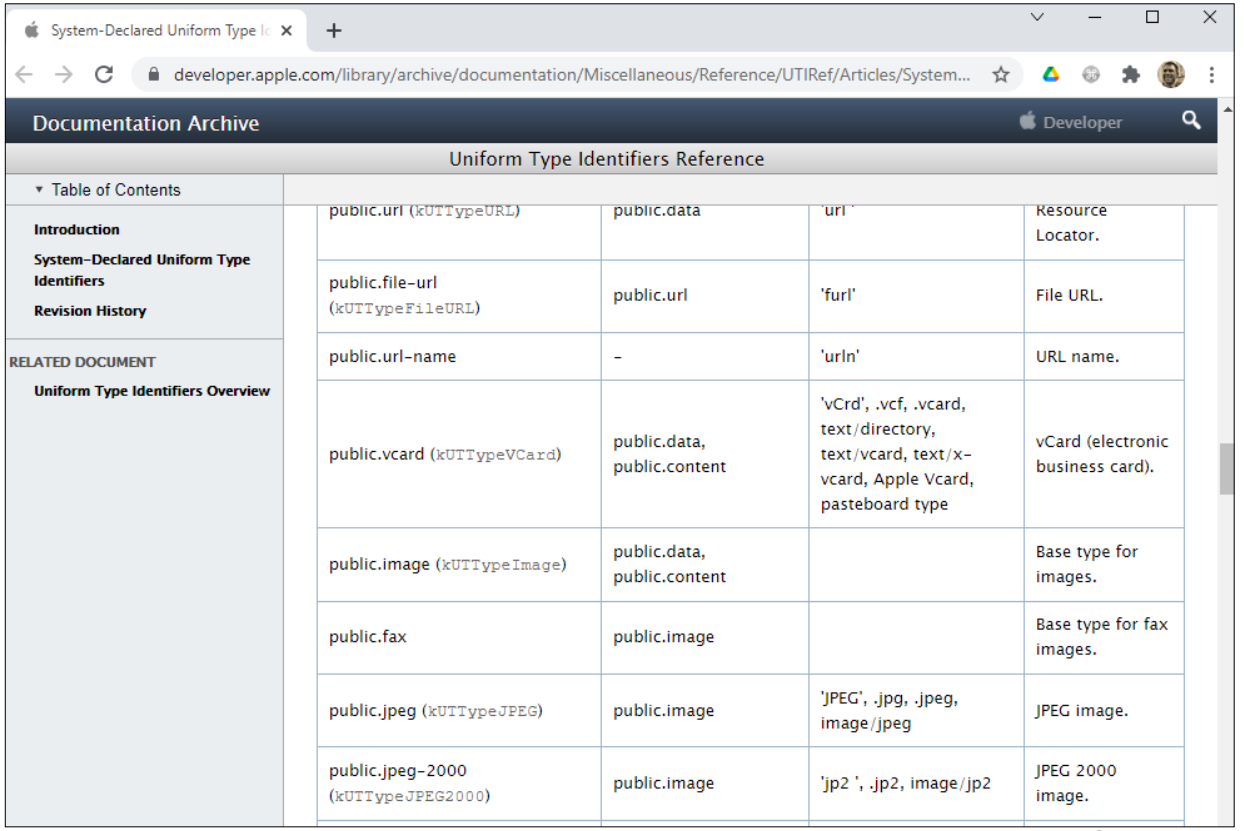
iana.org/assignments/media-types/font/woff2

```
Type name:  font

Subtype name:  woff2

Required parameters:  None

Optional parameters:  None

Encoding considerations:  Binary

Interoperability considerations:  WOFF 2.0 is an improvement on WOFF
   1.0.  The two formats have different Internet Media Types and
   different @font-face formats, and they may be used in parallel.

Published specification:  This media type registration is extracted
   from the WOFF 2.0 specification [W3C.CR-WOFF2-20150414] at W3C.

Applications that use this media type:  WOFF 2.0 is used by web
   browsers, often in conjunction with HTML and CSS.

Additional information:

   Magic number(s):  The signature field in the WOFF header MUST
      contain the "magic number" 0x774F4632 ('wOF2')

   File extension(s):  woff2

   Macintosh file type code(s):  (no code specified)

   Macintosh Universal Type Identifier code:  "org.w3.woff2"

   @font-face Format:  woff2

   Fragment Identifiers:  See Section 4.2.

Person & email address to contact for further information:
   Chris Lilley (www-font&w3.org).

Intended usage:  COMMON

Restrictions on usage:  None

Author:  The WOFF2 specification is a work product of the World Wide
   Web Consortium's WebFonts working group.

Change controller:  The W3C has change control over this
   specification.
```

font/woff2

Macintosh Universal Type Identifier code

- reverse-DNS naming structure
- public UTIs, and organization-specific UTIs
- multi-heritance



public.url (kUTTypeURL)	public.data	'url '	Resource Locator.
public.file-url (kUTTypeFileURL)	public.url	'furl'	File URL.
public.url-name	-	'urln'	URL name.
public.vcard (kUTTypeVCard)	public.data, public.content	'vCrd', .vcf, .vcard, text/directory, text/vcard, text/x-vcard, Apple Vcard, pasteboard type	vCard (electronic business card).
public.image (kUTTypeImage)	public.data, public.content		Base type for images.
public.fax	public.image		Base type for fax images.
public.jpeg (kUTTypeJPEG)	public.image	'JPEG', .jpg, .jpeg, image/jpeg	JPEG image.
public.jpeg-2000 (kUTTypeJPEG2000)	public.image	'jp2 ', .jp2, image/jp2	JPEG 2000 image.

https://developer.apple.com/library/archive/documentation/Miscellaneous/Reference/UTIRef/

Data Interoperability and Semantics

Part 2. Data Formats

Part 2.2. Delimiter-separated values

ICM – Toolbox Engineering and Interoperability of Software Systems – Course unit on Data Interoperability and Semantics
M1 Cyber Physical and Social Systems – Course unit on Data Interoperability and Semantics
Maxime Lefrançois https://maxime-lefrancois.info
Course unit URL: https://ci.mines-stetienne.fr/cps2/course/data

Delimiter-separated values

separator character

- comma

  Comma-separated values (CSV)

  ex: Microsoft Excel csv export[1]
- semicolon

  ex: Microsoft Excel csv export[2]
- tab

  Tab-separated values (TSV)
- colon

  ex: Linux `$ cat /etc/passwd`

```
File: /etc/passwd
root:x:0:0::/root:/bin/zsh
bin:x:1:1::/:/usr/bin/nologin
daemon:x:2:2::/:/usr/bin/nologin
mail:x:8:12::/var/spool/mail:/usr/bin/nologin
ftp:x:14:11::/srv/ftp:/usr/bin/nologin
http:x:33:33::/srv/http:/usr/bin/nologin
```

persons.csv - Notepad

File Edit Format View Help

```
Family Name,Given Name,VIAF ID
Ackersdijck,Willem Cornelis,17959345
Adelung,Friedrich von,22963658
Afzelius,Arvid August,49972119
Amerling,Karel,13331054
Anton,Karl Gottlob von,183632821
Arwidsson,Adolf Ivar,8184878
Asbjørnsen,Peter Christen,116587918
Attems,Heinrich,37665468
Atterbom,Per Daniel Amadeus,46819248
Balabin,Viktor Petrovich,44473845
Banks,Joseph,46830189
Beck,Friedrich,44338671
Becker,Reinhold von,42101066
Bernhart,Johann Baptist,69674335
Bertram,Johann,32890043
Bilderdijk,Willem,14882166
Boisserée,Sulpiz,7483155
Bopp,Franz,61614118
Borovský,Karel Havlíček,100277614
Bosković,Jovan,161354270
Buslaev,Fyodor,10074560
Cenowa,Florian Stanislaw,44466031
Chomiakov.Aleksei.66492873
```

[1] unless the decimal point of the locale is a comma (ex., France)

[2] when the decimal point of the locale is a comma

Delimiter-separated values

RFC 4180

separator character

records separated by CRLF

```
aaa,bbb,ccc CRLF
zzz,yyy,xxx CRLF
```

or

```
aaa,bbb,ccc CRLF
zzz,yyy,xxx
```

optional header file

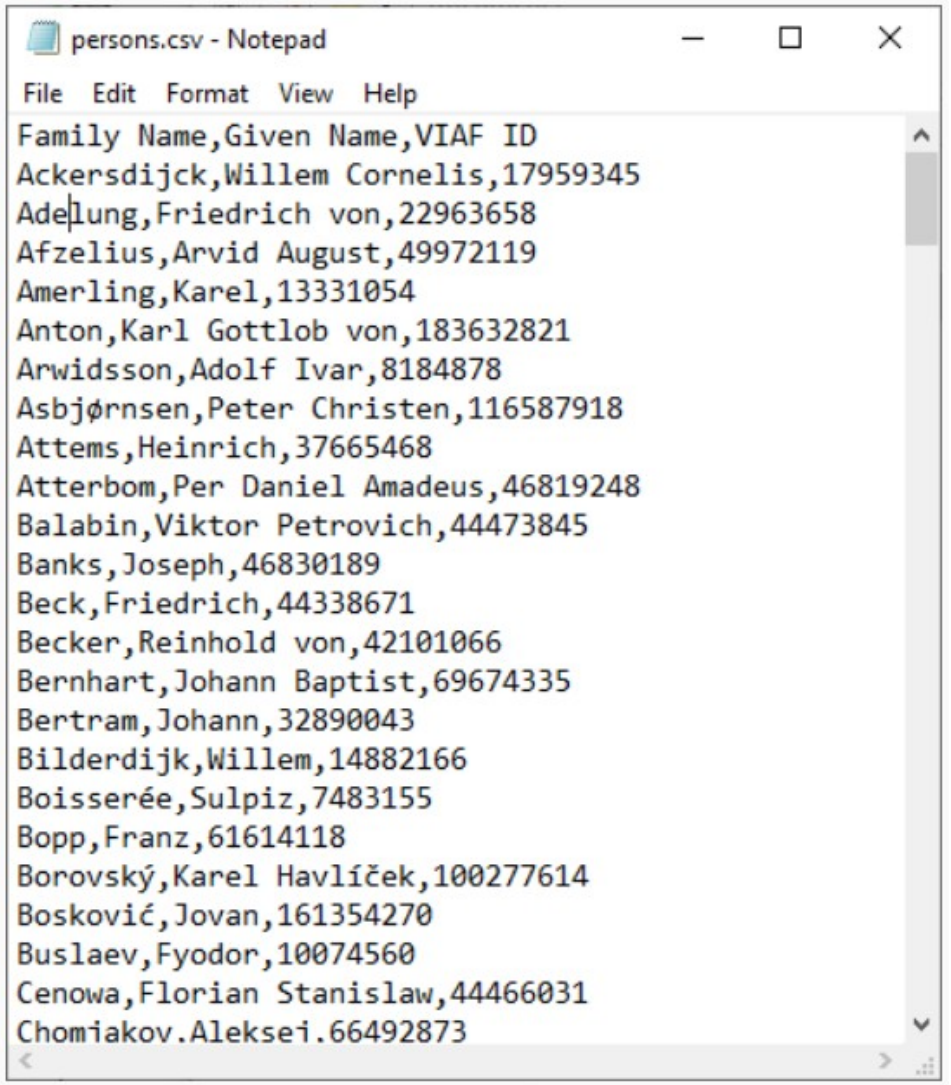
```
field_name,field_name,field_name CRLF
aaa,bbb,ccc CRLF
zzz,yyy,xxx CRLF
```

fields enclosed in double quote

```
"aaa","b CRLF
bb","ccc" CRLF
zzz,yyy,xxx
```

escaping cell character

```
"aaa","b""bb","ccc"
```



Delimiter-separated values

- Comparison Java libraries:
  https://github.com/uniVocity/csv-parsers-comparison
  - CPU: AMD Ryzen 7 1700 Eight-Core Processor @ 4.0 GHz
  - RAM: 32 GB
  - Storage: 1TB SSD drive
  - OS: Arch Linux 64-bit
  - JDK: 9.0.4 64-bit (Linux)
  - JDK: 1.8.0_144 64-bit (Linux)
  - JDK: 1.7.0_80 64-bit (Linux)
  - JDK: 1.6.0_45 64-bit (Linux)

Processing 3,173,958 rows of non RFC 4180 compliant input. No quoted values.

JDK 9

Parser	Average time	% Slower than best	Best time	Worst time
uniVocity CSV parser	739 ms	Best time!	707 ms	768 ms
SimpleFlatMapper CSV parser	861 ms	16%	848 ms	901 ms
Jackson CSV parser	1212 ms	64%	1169 ms	1238 ms
Product Collections parser	1409 ms	90%	1389 ms	1451 ms
Java CSV Parser	1498 ms	102%	1490 ms	1508 ms
JCSV Parser	1681 ms	127%	1660 ms	1710 ms
Oster Miller CSV parser	1772 ms	139%	1762 ms	1780 ms
Gen-Java CSV	1799 ms	143%	1790 ms	1805 ms
Simple CSV parser	1861 ms	151%	1832 ms	1900 ms
SuperCSV	1893 ms	156%	1858 ms	1964 ms
OpenCSV	2022 ms	173%	2007 ms	2037 ms
Apache Commons CSV	2424 ms	228%	2409 ms	2442 ms
Way IO Parser	2577 ms	248%	2532 ms	2638 ms

Data Interoperability and Semantics

Part 2. Data Formats

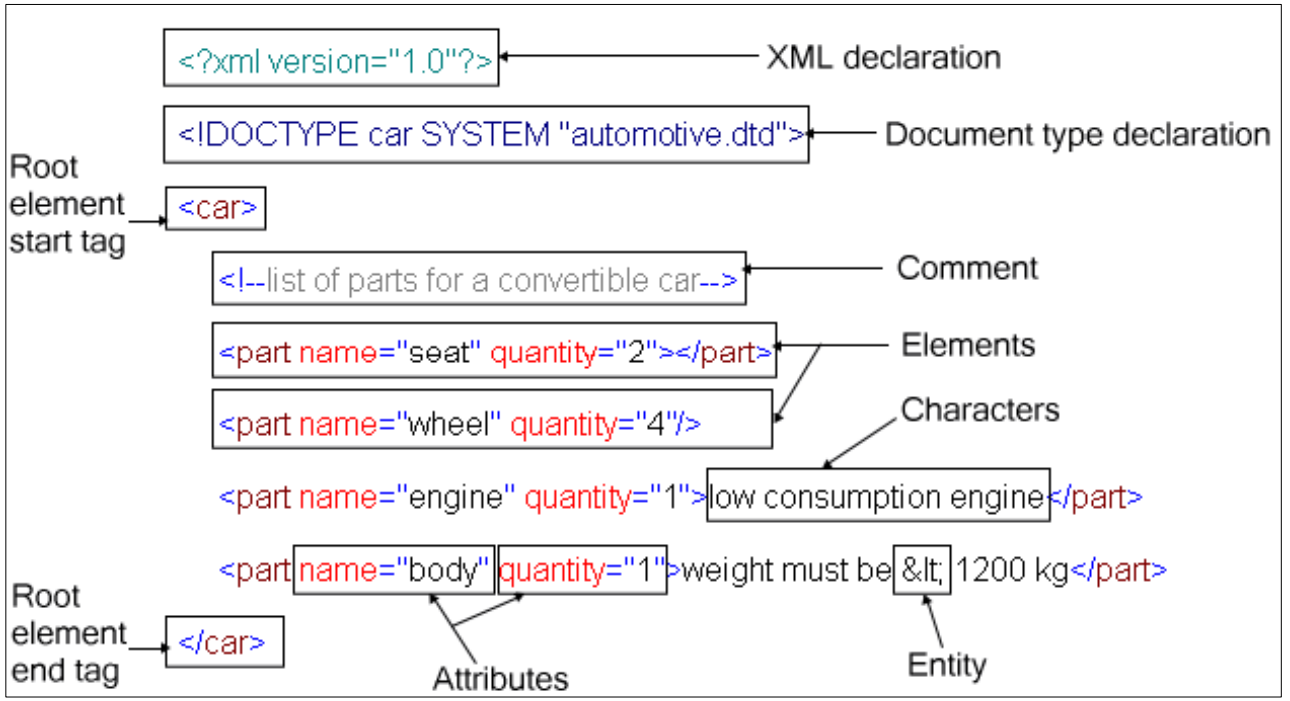
Part 2.3. Extensible Markup Language (XML)

ICM – Toolbox Engineering and Interoperability of Software Systems – Course unit on Data Interoperability and Semantics
M1 Cyber Physical and Social Systems – Course unit on Data Interoperability and Semantics
Maxime Lefrançois https://maxime-lefrancois.info
Course unit URL: https://ci.mines-stetienne.fr/cps2/course/data

Extensible Markup Language

XML (file format)

Filename extension	`.xml`
Internet media type	`application/xml` `text/xml` [1]
Uniform Type Identifier (UTI)	public.xml
UTI conformation	public.text
Magic number	`<?xml`
Developed by	World Wide Web Consortium
Type of format	Markup language
Extended from	SGML
Extended to	Numerous languages, including XHTML · RSS · Atom · KML
Standard	1.0 (Fifth Edition) (November 26, 2008; 12 years ago) 1.1 (Second Edition) (August 16, 2006; 15 years ago)
Open format?	Yes



- v1.0 in 1998, still extensively used in many verticals
- numerous formats based on XML (418 registered on IANA) https://en.wikipedia.org/wiki/List_of_XML_markup_languages
  application/atom+xml application/rdf+xml ...
- verbosity, complexity and redundancy

Extensible Markup Language

XML (file format)

Filename extension	`.xml`
Internet media type	`application/xml` `text/xml` [1]
Uniform Type Identifier (UTI)	public.xml
UTI conformation	public.text
Magic number	`<?xml`
Developed by	World Wide Web Consortium
Type of format	Markup language
Extended from	SGML
Extended to	Numerous languages, including XHTML · RSS · Atom · KML
Standard	1.0 (Fifth Edition) (November 26, 2008; 12 years ago) 1.1 (Second Edition) (August 16, 2006; 15 years ago)
Open format?	Yes

Characters and escaping

- unicode implementations: `<?xml version="1.0" encoding="UTF-8"?>`
- escaping characters: `<` '<' - `&` '&' - `❤` '❤' - etc.

Syntactical correctness

- *well formed vs ill-formed*
- one root tag
- correct nesting
- tag names (approx) start with letter, then alphanumeric or ':'

Schemas and validation

- *valid vs invalid*
- DTD, or XML Schema

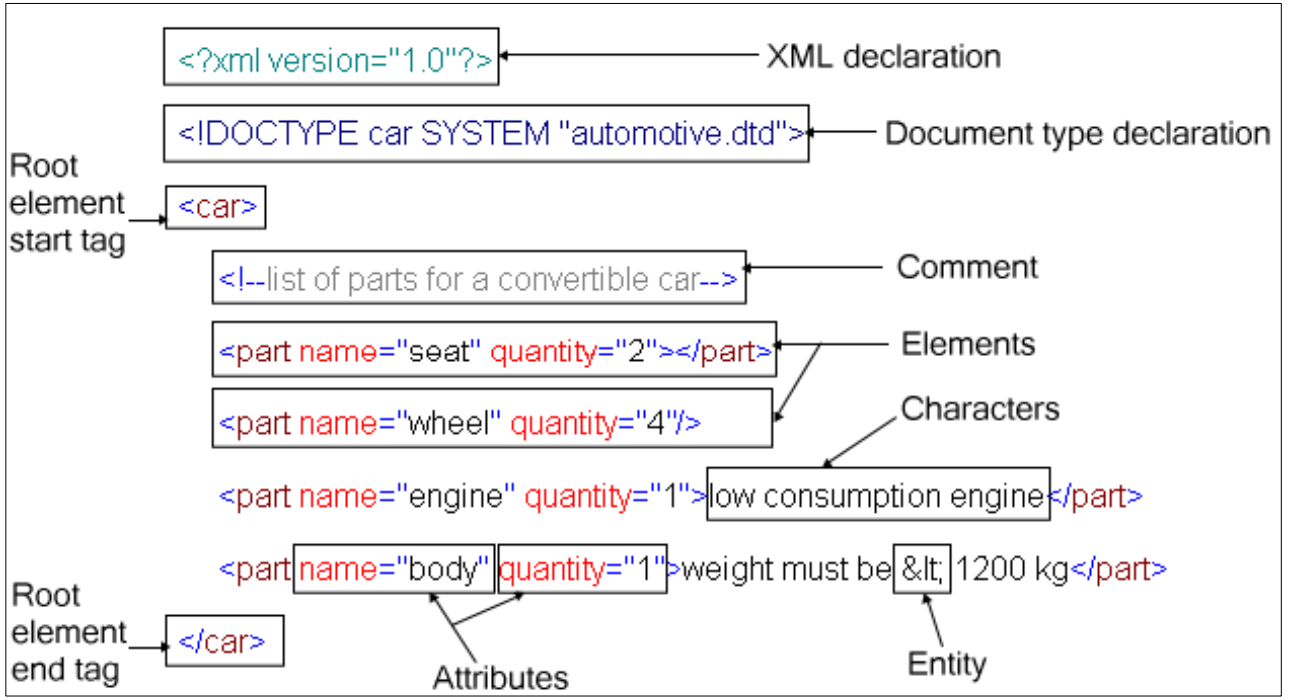
Namespaces

- `xmlns:ns1="http://example.org/ns1"`
  `xmlns:ns2="http://example.org/ns2"`
- allows to use different schemas together: `<ns1:Tag> <ns2:Tag>`

Extensible Markup Language

XML (file format)

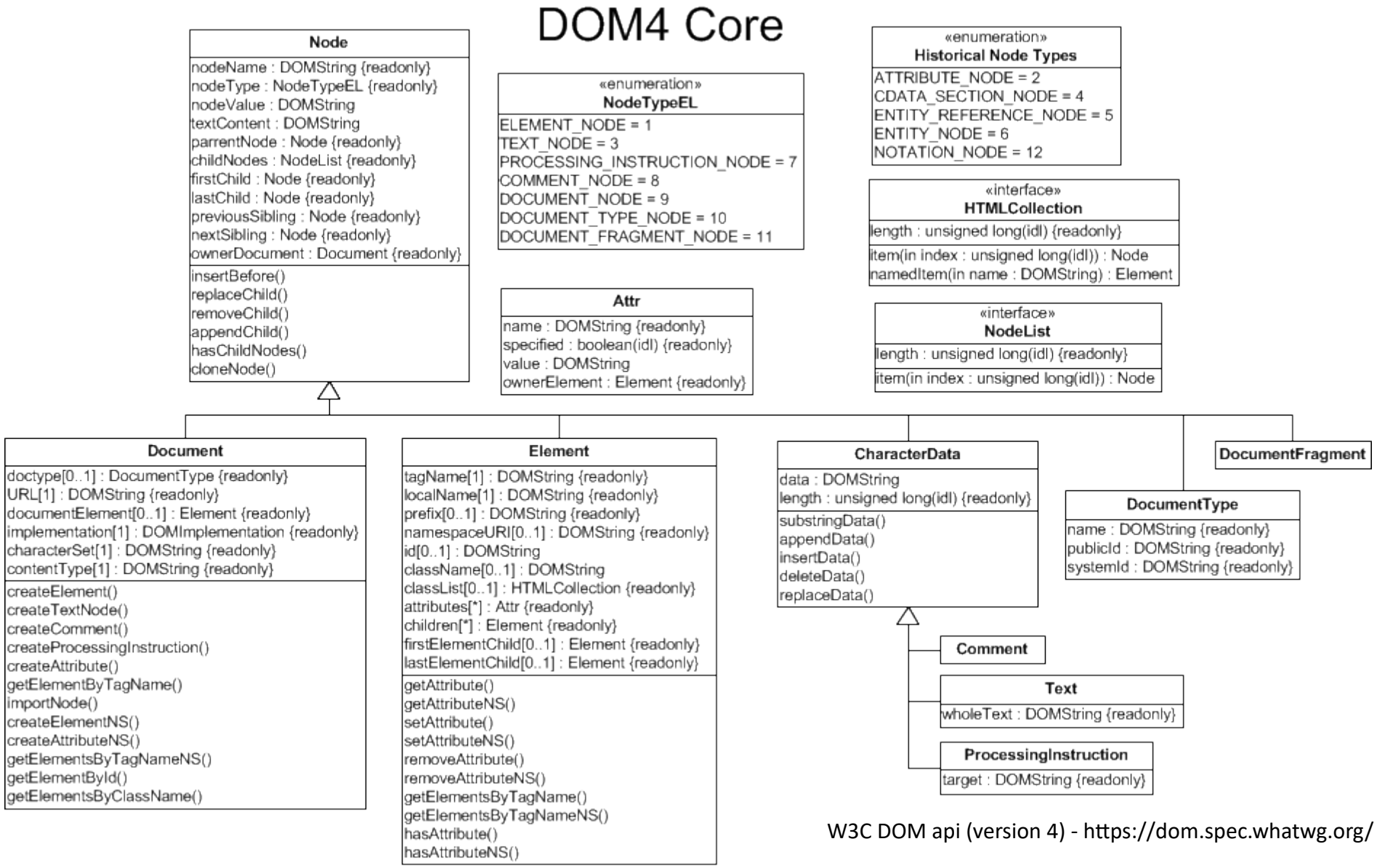
Filename extension	`.xml`
Internet media type	`application/xml` `text/xml` [1]
Uniform Type Identifier (UTI)	public.xml
UTI conformation	public.text
Magic number	`<?xml`
Developed by	World Wide Web Consortium
Type of format	Markup language
Extended from	SGML
Extended to	Numerous languages, including XHTML · RSS · Atom · KML
Standard	1.0 (Fifth Edition) (November 26, 2008; 12 years ago) 1.1 (Second Edition) (August 16, 2006; 15 years ago)
Open format?	Yes



Rules of thumb for modelling with XML:

- limit the number of tag types and attribute types!
- use attributes for simple datatypes only
- order of elements is important
- group elements when appropriate
- anticipate how you are going to write XPath queries

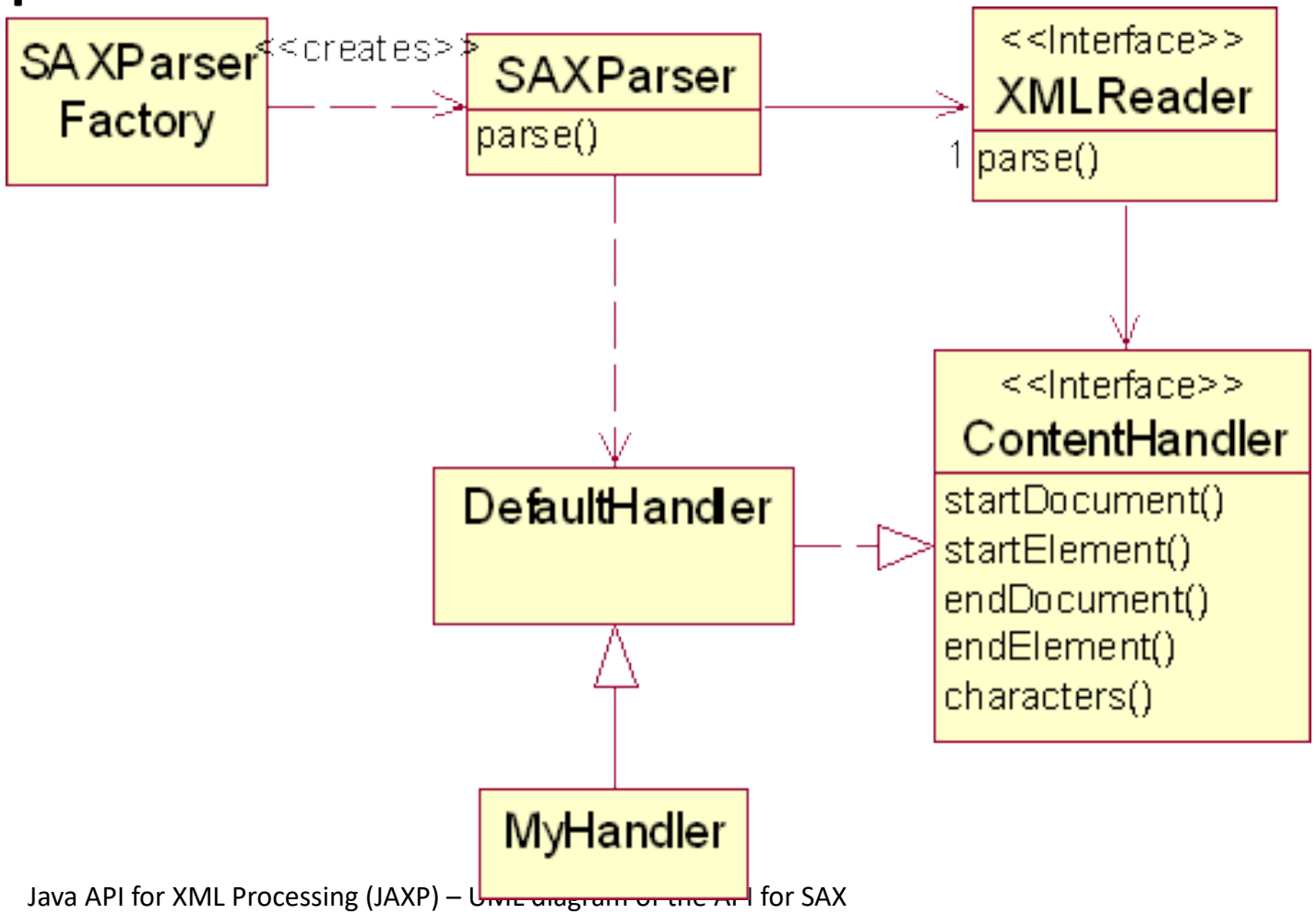
XML manipulation: Document Object Model (DOM)



W3C DOM api (version 4) - https://dom.spec.whatwg.org/

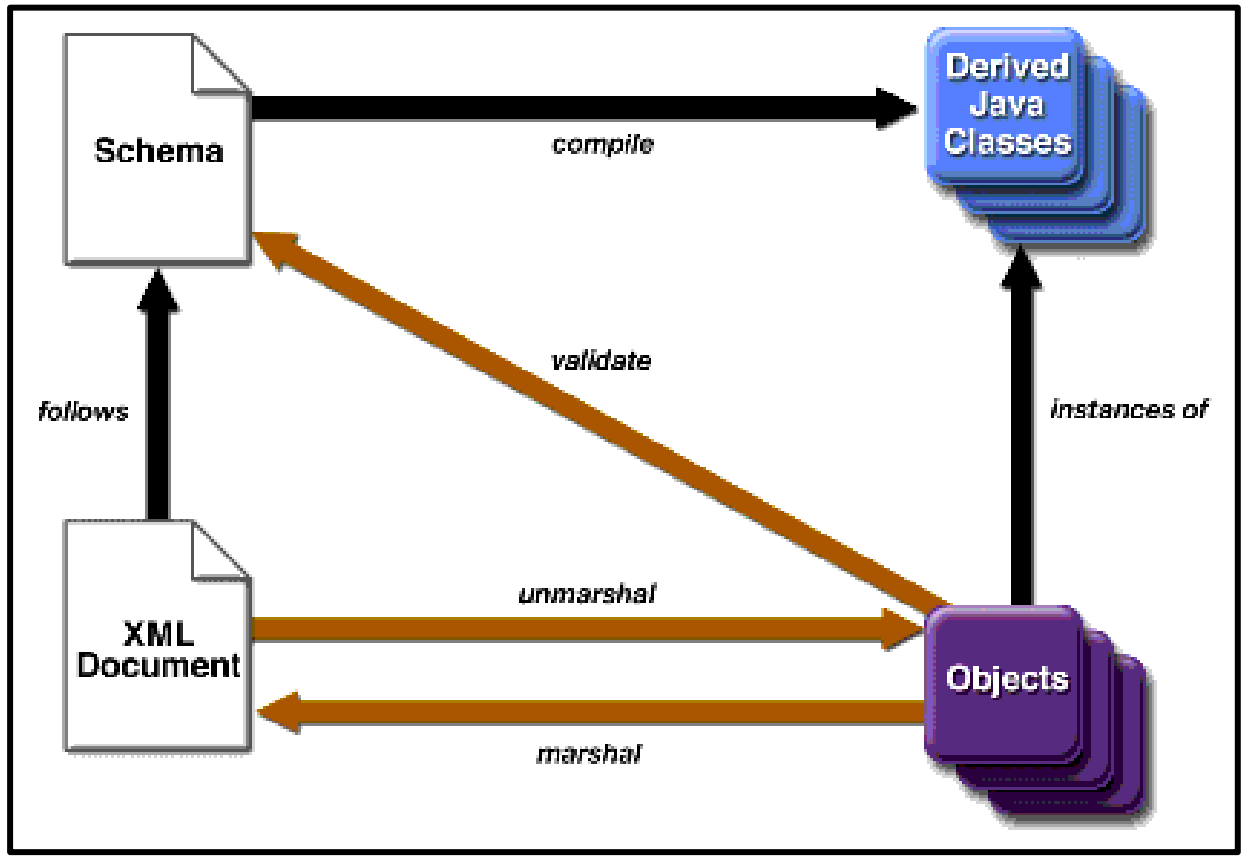
SAX (Simple API for XML)

XML traversal API



Java API for XML Processing (JAXP) – UML diagram of the API for SAX

XML – OOP data binding



XML-OOP data binding

ex Java: https://zetcode.com/java/jaxb/

Employee.java

```java
@XmlRootElement(name = "employee")
@XmlAccessorType(XmlAccessType.FIELD)
public class Employee implements Serializable
{
    private Integer id;
    private String firstName;
    private String lastName;
}
```

employee.xml

```xml
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<employee>
    <firstName>Lokesh</firstName>
    <id>1</id>
    <lastName>Gupta</lastName>
</employee>
```

Java Architecture for XML Binding (JAXB) example

https://howtodoinjava.com/jaxb/jaxb-annotations/

XPath: Declarative tree-traversal language

XPATH locators

You can test all locator examples using Firebug and Firepath on the following page: https://vpl.bibliocommons.com/search?q=java&t=keyword
See more articles about XPATH on my blog,http://test-able.blogspot.ca .

//input

finds all INPUT elements that are included in the web page

//input[@aria-label='Search Catalogue']

finds all INPUT elements that have an ARIA-LABEL attribute with the "Search Catalogue" value

//button[@data-id='sortSelector']/span

1. finds all BUTTON elements that have a DATA-ID attribute with the "sortSelector" value
2. for each element from the BUTTON list, find the SPAN elements that are direct children

(//div[@class='listItem clearfix '])[2]//span

1. finds the DIV elements that have the CLASS attribute's value equal to "listItem clearfix"
2. gets the second element from the DIV elements list
3. finds all SPAN elements (direct children or not) that are inside of the second DIV

//input[@id='globalQuery' and @name='q']

finds all INPUT elements that
1. have an ID attribute with the "globalQuery" value
2. have a NAME attribute with the "q" value

More XPATH locators

You can test all locator examples using Firebug and Firepath on the following page: https://vpl.bibliocommons.com/search?q=java&t=keyword

(//div[@class='listItem clearfix '])[3]

1. finds all DIV elements that have the CLASS attribute's value equal to "listItem clearfix"
2. selects the 3rd DIV element from the list of DIV elements

//input[@id='globalQuery' and @name='q']

finds all INPUT elements that
1. have an ID attribute with the "globalQuery" value
2. have a NAME attribute with the "q" value

//a[contains(@href, 'show_circulation')]

finds all A elements that have "show_circulation" included in the value of the HREF attribute

//a[starts-with(@href, '/item/show_circulation')]

finds all A elements that have the HREF attribute's value starting with "/item/show_circulation"

//input[not(@name='q')]

finds all INPUT elements that have the NAME attribute's value different than "q"

https://www.w3.org/TR/xpath/
https://cheatography.com/alexsiminiuc/cheat-sheets/xpath/pdf/
https://devhints.io/xpath

(//div[@class='listItem clearfix '])[last()]

1. finds all DIV elements that have the CLASS attribute's value equal to "listItem clearfix"
2. select the last DIV element from the list of DIV elements

count(//div[@class='listItem clearfix '])

provides the count of all DIV elements that have the CLASS attribute's value equal to "listItem clearfix"

//a[@class='extendSearch']/@testid

1. finds the A elements that have the CLASS attribute's value equal to "extendSearch"
2. gets the value of the TESTID attribute for the A elements

//a[@class='extendSearch']/text()

1. finds the A elements that have the CLASS attribute's value equal to "extendSearch"
2. get the value of the A elements

//div[normalize-space(@class)='listItem clearfix']

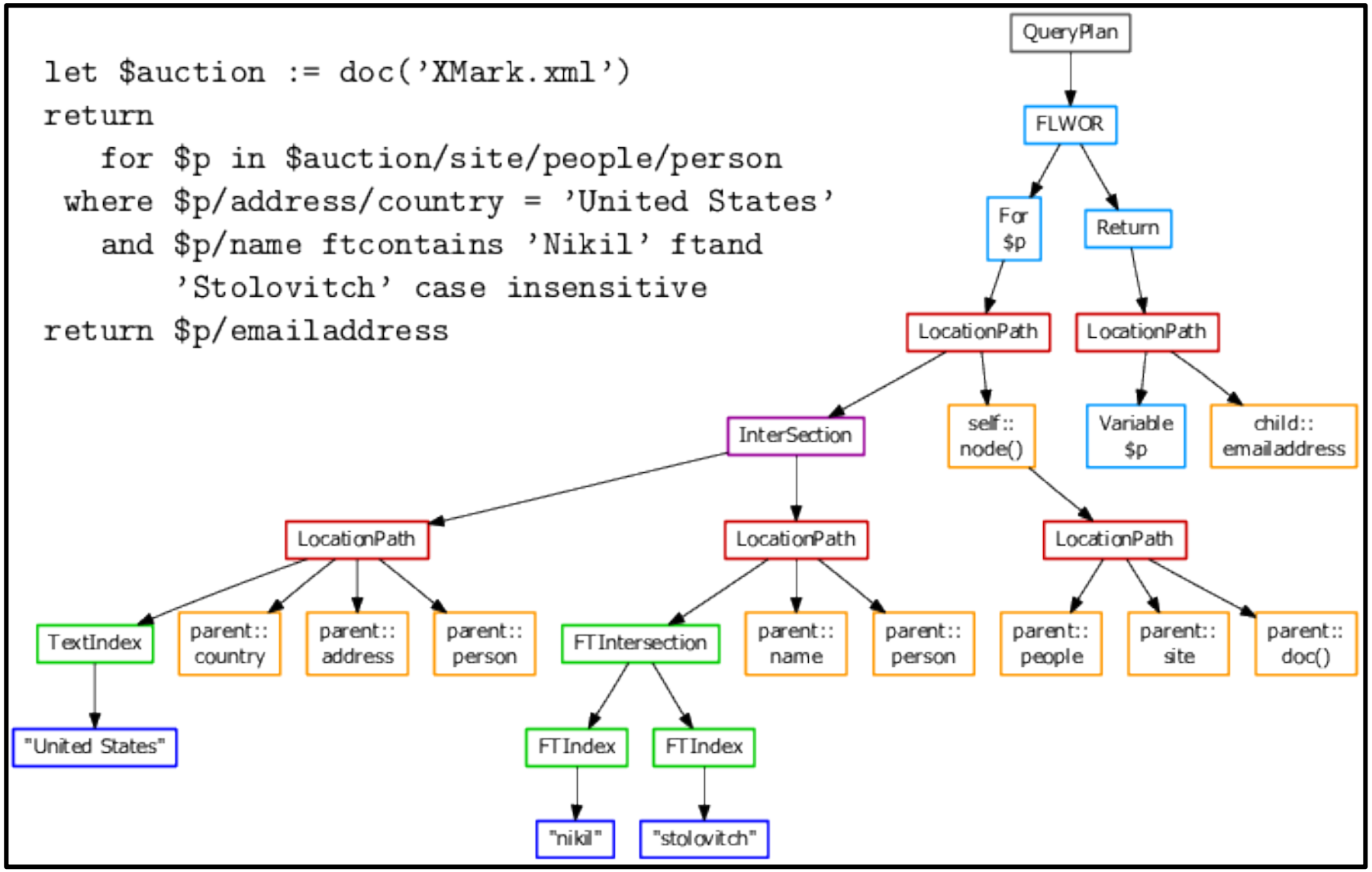
1. finds the DIV elements that have a CLASS attribute
2. removes all spaces from the value of the CLASS attribute
2. finds the DIV elements that have the value of the CLASS attribute (excluding spaces) equal to "listItem clearfix"

//a[string-length(@href) > 70]

finds all A elements that have the HREF attribute value length greater than 70 characters

XQuery: Declarative query language

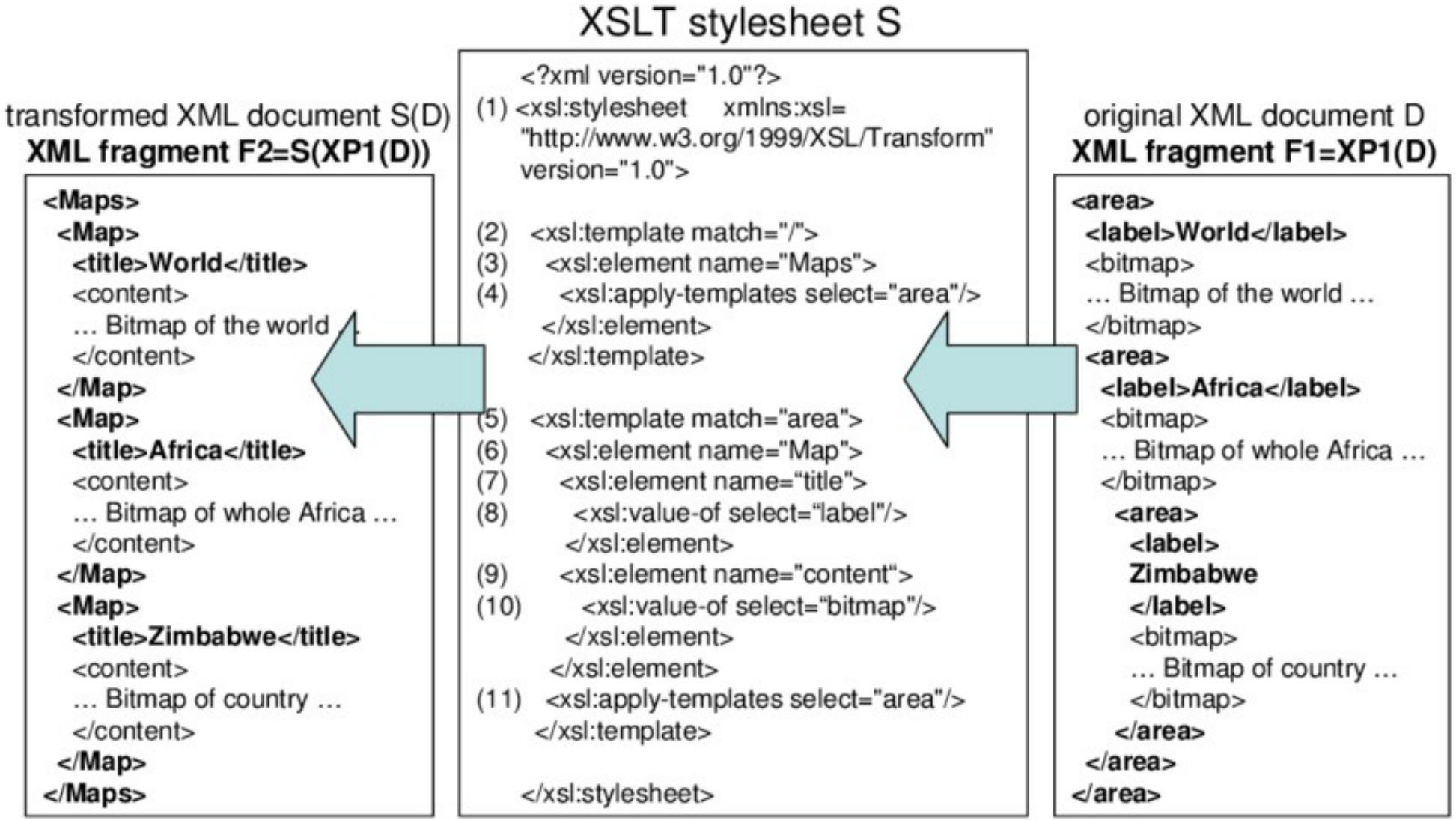
```
let $auction := doc('XMark.xml')
return
   for $p in $auction/site/people/person
 where $p/address/country = 'United States'
   and $p/name ftcontains 'Nikil' ftand
       'Stolovitch' case insensitive
return $p/emailaddress
```



XSLT: Declarative transformation language

XSLT

Paradigm	Declarative
Developer	World Wide Web Consortium (W3C)
First appeared	1998
Stable release	3.0 / June 8, 2017; 4 years ago
Filename extensions	.xslt
Website	www.w3.org/TR/xslt-30/
Major implementations	
libxslt, Saxon, Xalan	
Influenced by	
DSSSL	



Data Interoperability and Semantics

Part 2. Data Formats

Part 2.4. JavaScript Object Notation (JSON)

ICM – Toolbox Engineering and Interoperability of Software Systems – Course unit on Data Interoperability and Semantics
M1 Cyber Physical and Social Systems – Course unit on Data Interoperability and Semantics
Maxime Lefrançois https://maxime-lefrancois.info
Course unit URL: https://ci.mines-stetienne.fr/cps2/course/data

JavaScript Object Notation

JavaScript Object Notation

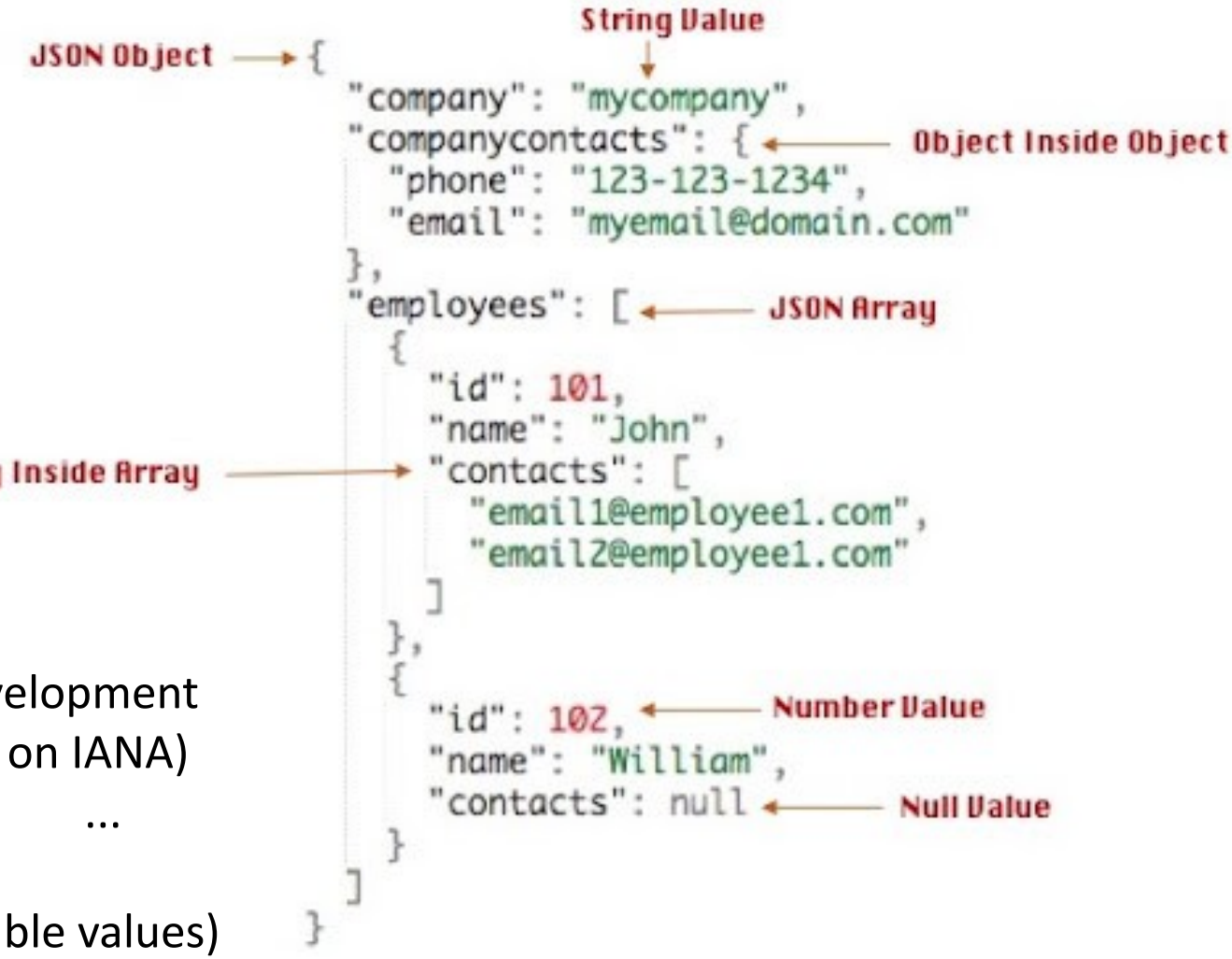
Filename extension	`.json`
Internet media type	`application/json`
Type code	TEXT
Uniform Type Identifier (UTI)	public.json
Type of format	Data interchange
Extended from	JavaScript
Standard	STD 90 (RFC 8259), ECMA-404, ISO/IEC 21778:2017
Open format?	Yes
Website	json.org

- since early 2000s. Now used a lot in software development
- many mediatypes based on JSON (118 registered on IANA)
  application/geo+json application/ld+json ...
- simple, human-readable
- attribute–value pairs and arrays (or other serializable values)

JavaScript Object Notation

JavaScript Object Notation

Filename extension	.json
Internet media type	application/json
Type code	TEXT
Uniform Type Identifier (UTI)	public.json
Type of format	Data interchange
Extended from	JavaScript
Standard	STD 90 (RFC 8259), ECMA-404, ISO/IEC 21778:2017
Open format?	Yes
Website	json.org



- since early 2000s. Now used a lot in software development
- many mediatypes based on JSON (118 registered on IANA)
  application/geo+json application/ld+json ...
- simple, human-readable
- attribute–value pairs and arrays (or other serializable values)

JavaScript Object Notation

JavaScript Object Notation

Filename extension	`.json`
Internet media type	`application/json`
Type code	TEXT
Uniform Type Identifier (UTI)	public.json
Type of format	Data interchange
Extended from	JavaScript
Standard	STD 90 (RFC 8259), ECMA-404, ISO/IEC 21778:2017
Open format?	Yes
Website	json.org

Characters and escaping

- full Unicode character set
- escaping characters \b \f \n \r \t \" \

Data types

- number
- string
- boolean
- array
- object
- null

Semantics

While JSON provides a syntactic framework for data interchange, unambiguous data interchange also requires agreement between producer and consumer on the semantics of specific use of the JSON syntax. One example of where such an agreement is necessary is the serialization of data types defined by the JavaScript syntax that are not part of the JSON standard, e.g., Date, Function, Regular Expression, and "undefined"

JSON Schema

JSON Schema Validator

newtonsoft.com

An online, interactive JSON Schema validator. Supports JSON Schema Draft 3, Draft 4, Draft 6 and Draft 7.

View source code

Select schema: Empty schema

```
{
  "$schema": "http://json-schema.org/draft-07/schema#",

  "definitions": {
    "person": {
      "type": "object",
      "properties": {
        "name": { "type": "string" },
        "children": {
          "type": "array",
          "items": { "$ref": "#/definitions/person" },
          "default": []
        }
      }
    },

  "type": "object",

  "properties": {
    "person": { "$ref": "#/definitions/person" }
  }
}
```

Input JSON:

```
{
  "personsdfsd": {
    "nameklk": "Elizabeth",
    "childrensss": [
      {
        "namess": "Charles",
        "childrenss": [
          {
            "namess": "William",
            "children": [
              { "namesss": "George" },
              { "name": "Charlotte" }
            ]
          },
          {
            "namesss": "Harry"
          }
        ]
      }
    ]
  }
}
```

✔ No errors found. JSON validates against the schema

JSONPath: Declarative tree-traversal language

XPath	JSONPath	Description
/	$	the root object/element
.	@	the current object/element
/	. or []	child operator
..	n/a	parent operator
//	..	recursive descent. JSONPath borrows this syntax from E4X.
*	*	wildcard. All objects/elements regardless their names.
@	n/a	attribute access. JSON structures don't have attributes.
[]	[]	subscript operator. XPath uses it to iterate over element collections and for predicates. In Javascript and JSON it is the native array operator.
|	[,]	Union operator in XPath results in a combination of node sets. JSONPath allows alternate names or array indices as a set.
n/a	[start:end:step]	array slice operator borrowed from ES4.
[]	?()	applies a filter (script) expression.
n/a	()	script expression, using the underlying script engine.
()	n/a	grouping in Xpath

Examples:

```
$.store.book[0].title
$['store']['book'][0]['title']
$.store.book[(@.length-1)].title
$.store.book[?(@.price < 10)].title
```

https://goessner.net/articles/JsonPath/

JSONPath: Declarative tree-traversal language

```
{ "store": {
    "book": [
      { "category": "reference",
        "author": "Nigel Rees",
        "title": "Sayings of the Century",
        "price": 8.95
      },
      { "category": "fiction",
        "author": "Evelyn Waugh",
        "title": "Sword of Honour",
        "price": 12.99
      },
      { "category": "fiction",
        "author": "Herman Melville",
        "title": "Moby Dick",
        "isbn": "0-553-21311-3",
        "price": 8.99
      },
      { "category": "fiction",
        "author": "J. R. R. Tolkien",
        "title": "The Lord of the Rings",
        "isbn": "0-395-19395-8",
        "price": 22.99
      }
    ],
    "bicycle": {
      "color": "red",
      "price": 19.95
    }
  }
}
```

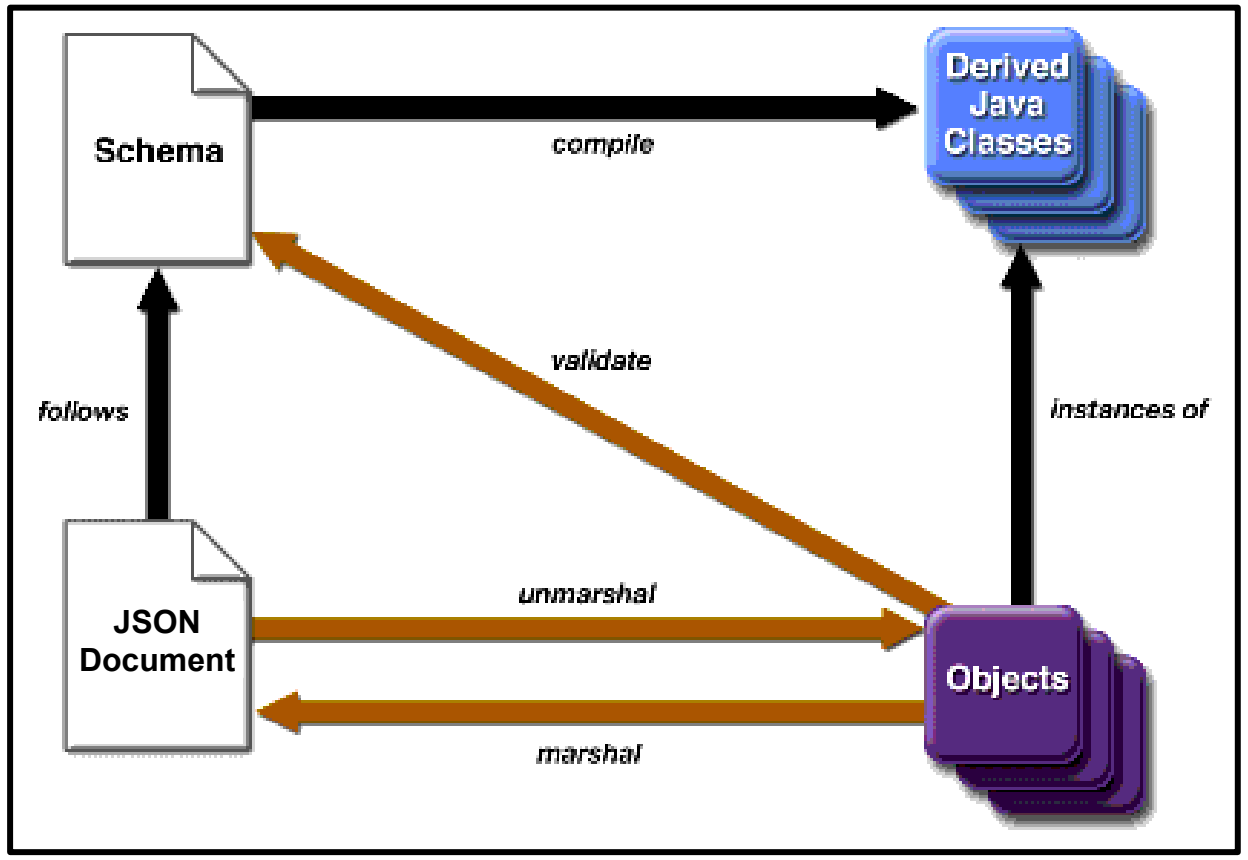
XPath	JSONPath	Result
/store/book/author	$.store.book[*].author	the authors of all books in the store
//author	$..author	all authors
/store/*	$.store.*	all things in store, which are some books and a red bicycle.
/store//price	$.store..price	the price of everything in the store.
//book[3]	$..book[2]	the third book
//book[last()]	$..book[(@.length-1)] $..book[-1:]	the last book in order.
//book[position()<3]	$..book[0,1] $..book[:2]	the first two books
//book[isbn]	$..book[?(@.isbn)]	filter all books with isbn number
//book[price<10]	$..book[?(@.price<10)]	filter all books cheapier than 10
//*	$..*	all Elements in XML document. All members of JSON structure.

https://goessner.net/articles/JsonPath/

APIs for JSON Processing

example for java: https://javaee.github.io/jsonp/

JSON – OOP data binding



JSON-OOP data binding

Parsing Speed	Speed. MB/MS	Parsing Time
GSON	100%	0%
Jackson	58%	70.87%
JSON.simple	79%	126.58%
JSONP	44%	25.49%

Comparison of Java JSON libraries: Jackson vs. Gson vs. JSON-B vs. JSON-P vs. org.JSON vs. Jsonpath
https://itsallbinary.com/jackson-vs-gson-vs-json-b-vs-json-p-vs-org-json-vs-jsonpath-java-json-libraries-features-comparison/

Data Interoperability and Semantics

Part 2. Data Formats

Part 2.5. Configuration file formats

ICM – Toolbox Engineering and Interoperability of Software Systems – Course unit on Data Interoperability and Semantics
M1 Cyber Physical and Social Systems – Course unit on Data Interoperability and Semantics
Maxime Lefrançois https://maxime-lefrancois.info
Course unit URL: https://ci.mines-stetienne.fr/cps2/course/data

.properties files

.properties	
Filename extension	.properties

Mainly for Java-related conf files

Ex: i18n[1] and l10n[2]

- `key=value`,
- `key = value`,
- `key:value`,
- `key value`

```
# You are reading a comment in ".properties" file.
! The exclamation mark can also be used for comments.
# Lines with "properties" contain a key and a value separated by a delimiting character.
# There are 3 delimiting characters: '=' (equal), ':' (colon) and whitespace (space, \t and \f).
website = https://en.wikipedia.org/
language : English
topic .properties files
# A word on a line will just create a key with no value.
empty
# White space that appears between the key, the value and the delimiter is ignored.
# This means that the following are equivalent (other than for readability).
hello=hello
hello = hello
# Keys with the same name will be overwritten by the key that is the furthest in a file.
# For example the final value for "duplicateKey" will be "second".
duplicateKey = first
duplicateKey = second
# To use the delimiter characters inside a key, you need to escape them with a \.
# However, there is no need to do this in the value.
delimiterCharacters\:\=\ = This is the value for the key "delimiterCharacters\:\=\ "
# Adding a \ at the end of a line means that the value continues to the next line.
multiline = This line \
continues
# If you want your value to include a \, it should be escaped by another \.
path = c:\\wiki\\templates
# This means that if the number of \ at the end of the line is even, the next line is not included in the value.
# In the following example, the value for "evenKey" is "This is on one line\".
evenKey = This is on one line\\
# This line is a normal comment and is not included in the value for "evenKey"
# If the number of \ is odd, then the next line is included in the value.
# In the following example, the value for "oddKey" is "This is line one and\#This is line two".
oddKey = This is line one and\\\
# This is line two
# White space characters are removed before each line.
# Make sure to add your spaces before your \ if you need them on the next line.
# In the following example, the value for "welcome" is "Welcome to Wikipedia!".
welcome = Welcome to \
          Wikipedia!
# If you need to add newlines and carriage returns, they need to be escaped using \n and \r respectively.
# You can also optionally escape tabs with \t for readability purposes.
valueWithEscapes = This is a newline\n and a carriage return\r and a tab\t.
# You can also use Unicode escape characters (maximum of four hexadecimal digits).
# In the following example, the value for "encodedHelloInJapanese" is "こんにちは".
encodedHelloInJapanese = \u3053\u3093\u306b\u3061\u306f
# But with more modern file encodings like UTF-8, you can directly use supported characters.
helloInJapanese = こんにちは
```

[1] i18n: internationalization
[2] l10n: localization

INI files

INI

Filename extension	.ini
Internet media type	text/plain, application/textedit, zz-application/zz-winassoc-ini
Type of format	Initialization/Configuration File

https://en.wikipedia.org/wiki/INI_file

- Developed for Windows for initialization: BOOT.INI, SYSTEM.INI, WIN.INI, …
- Similar in Linux systems with .conf, .cfg, …
- Syntax used by many programs
  - php.ini
  - ssh.conf
  - git.conf
  - …

```
; last modified 1 April 2001 by John Doe
[owner]
name = John Doe
organization = Acme Widgets Inc.

[database]
; use IP address in case network name resolution is not working
server = 192.0.2.62
port = 143
file = "payroll.dat"
```

```
[section]
domain = wikipedia.org

[section.subsection]
foo = bar
```

TOML files

TOML

Filename extension	`.toml`
Internet media type	*Not registered*[1]
Developed by	Tom Preston-Werner Community
Initial release	23 February 2013; 9 years ago
Latest release	v1.0.0 January 11, 2021; 18 months ago
Type of format	Data interchange
Open format?	Yes
Website	toml.io

https://en.wikipedia.org/wiki/TOML

Tom's Obvious, Minimal Language.

https://toml.io/en/

- `key = "value"` pairs
- `[section names]`,
- `# comments`
- Datatypes:
  - String,
  - Integer,
  - Float,
  - Boolean,
  - Datetime,
  - Array,
  - and Table

```toml
# This is a TOML document.

title = "TOML Example"

[owner]
name = "Tom Preston-Werner"
dob = 1979-05-27T07:32:00-08:00 # First class dates

[database]
server = "192.168.1.1"
ports = [ 8000, 8001, 8002 ]
connection_max = 5000
enabled = true

[servers]

  # Indentation (tabs and/or spaces) is allowed but not require
  [servers.alpha]
  ip = "10.0.0.1"
  dc = "eqdc10"

  [servers.beta]
  ip = "10.0.0.2"
  dc = "eqdc10"

[clients]
data = [ ["gamma", "delta"], [1, 2] ]

# Line breaks are OK when inside arrays
hosts = [
  "alpha",
  "omega"
]
```

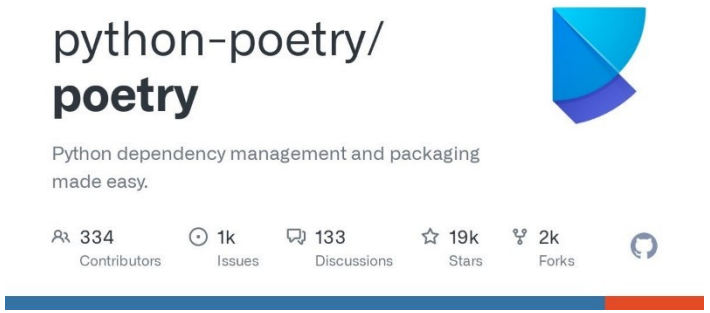
TOML files

TOML

Filename extension	`.toml`
Internet media type	*Not registered*[1]
Developed by	Tom Preston-Werner Community
Initial release	23 February 2013; 9 years ago
Latest release	v1.0.0 January 11, 2021; 18 months ago
Type of format	Data interchange
Open format?	Yes
Website	toml.io

https://en.wikipedia.org/wiki/TOML

Example:
Python packaging and dependency management with Poetry




```toml
[project]
name = "infer_pyproject"
version = "0.1.0"
description = "Create a pyproject.toml file for an existing project."
authors = [
    "Martin Thoma <info@martin-thoma.de>"
]
license = "MIT"
readme = "README.md"
python = "^3.6"
homepage = "https://github.com/MartinThoma/infer_pyproject"
repository = "https://github.com/MartinThoma/infer_pyproject"
documentation = "https://github.com/MartinThoma/infer_pyproject"

keywords = ["packaging", "dependency", "infer", "pyproject.toml"]

classifiers = [
    "Topic :: Software Development"
]

# Requirements
[dependencies]
Click = "^7.0"

[dev-dependencies]
black = { version = "^18.3-alpha.0", python = "^3.6" }

[scripts]
poetry = "infer_pyproject.cli:main"

[build-system]
requires = [
    "setuptools >= 35.0.2",
    "setuptools_scm >= 2.0.0, <3"
]
build-backend = "setuptools.build_meta"
```

https://martin-thoma.com/pyproject-toml/

Data Interoperability and Semantics

Part 2. Data Formats

Part 2.6. YAML Ain’t Markup Language (YAML)

ICM – Toolbox Engineering and Interoperability of Software Systems – Course unit on Data Interoperability and Semantics
M1 Cyber Physical and Social Systems – Course unit on Data Interoperability and Semantics
Maxime Lefrançois https://maxime-lefrancois.info
Course unit URL: https://ci.mines-stetienne.fr/cps2/course/data

YAML Ain’t Markup Language (YAML)

YAML

Filename extensions	`.yaml`, `.yml`
Internet media type	*Not registered*
Initial release	11 May 2001; 20 years ago
Latest release	1.2 (Third Edition) (1 October 2009; 12 years ago)
Type of format	Data interchange
Open format?	Yes
Website	yaml.org

- since early 2000s
- used a lot for software configuration
- human-readable, superset of JSON

```
--- !clarkevans.com/^invoice
invoice: 34843
date   : 2001-01-23
bill-to: &id001
    given  : Chris
    family : Dumars
    address:
        lines: |
            458 Walkman Dr.
            Suite #292
        city    : Royal Oak
        state   : MI
        postal  : 48046
ship-to: *id001
product:
    - sku         : BL394D
      quantity    : 4
      description : Basketball
      price       : 450.00
    - sku         : BL4438H
      quantity    : 1
      description : Super Hoop
      price       : 2392.00
tax  : 251.42
total: 4443.52
comments: >
    Late afternoon is best.
    Backup contact is Nancy
    Billsmer @ 338-4338.
```

SCALAR

COLLECTIONS

MULTI-LINE COLLECTIONS

LISTS/DICTIONARIES

MULTI-LINE FORMATTING

YAML Ain’t Markup Language (YAML)

YAML

Filename extensions	`.yaml`, `.yml`
Internet media type	*Not registered*
Initial release	11 May 2001; 20 years ago
Latest release	1.2 (Third Edition) (1 October 2009; 12 years ago)
Type of format	Data interchange
Open format?	Yes
Website	yaml.org

- Whitespace indentation, tabs forbidden
- end-of-line comments (#)
- lists

```yaml
--- # Favorite movies
- Casablanca
- North by Northwest
- The Man Who Wasn't There
```

or

```yaml
--- # Shopping list
[milk, pumpkin pie, eggs, juice]
```

- associative arrays

```yaml
--- # Indented Block
  name: John Smith
  age: 33
--- # Inline Block
{name: John Smith, age: 33}
```

YAML Ain’t Markup Language (YAML)

YAML

Filename extensions	`.yaml`, `.yml`
Internet media type	*Not registered*
Initial release	11 May 2001; 20 years ago
Latest release	1.2 (Third Edition) (1 October 2009; 12 years ago)
Type of format	Data interchange
Open format?	Yes
Website	yaml.org

- strings

```
data: |
   There once was a tall man from Ealing
   Who got on a bus to Darjeeling
       It said on the door
       "Please don't sit on the floor"
   So he carefully sat on the ceiling
```

or

```
data: >
   Wrapped text
   will be folded
   into a single
   paragraph

   Blank lines denote
   paragraph breaks
```

- datatypes

```
---
a: 123                  # an integer
b: "123"                # a string, disambiguated by quotes
c: 123.0                # a float
d: !!float 123          # also a float via explicit data type prefixed by (!!)
e: !!str 123            # a string, disambiguated by explicit type
f: !!str Yes            # a string via explicit type
g: Yes                  # a boolean True (yaml1.1), string "Yes" (yaml1.2)
h: Yes we have No bananas  # a string, "Yes" and "No" disambiguated by context.
```

YAML Ain’t Markup Language (YAML)

YAML

Filename extensions	`.yaml`, `.yml`
Internet media type	*Not registered*
Initial release	11 May 2001; 20 years ago
Latest release	1.2 (Third Edition) (1 October 2009; 12 years ago)
Type of format	Data interchange
Open format?	Yes
Website	yaml.org

- anchors and references

```yaml
--- # Sequencer protocols for Laser eye surgery
- step:  &id001                  # defines anchor label &id001
    instrument:      Lasik 2000
    pulseEnergy:     5.4
    pulseDuration:   12
    repetition:      1000
    spotSize:        1mm

- step: &id002
    instrument:      Lasik 2000
    pulseEnergy:     5.0
    pulseDuration:   10
    repetition:      500
    spotSize:        2mm
- step: *id001                   # refers to the first step (with anchor &id001)
- step: *id002                   # refers to the second step
- step: *id002
```

YAML Ain’t Markup Language (YAML)

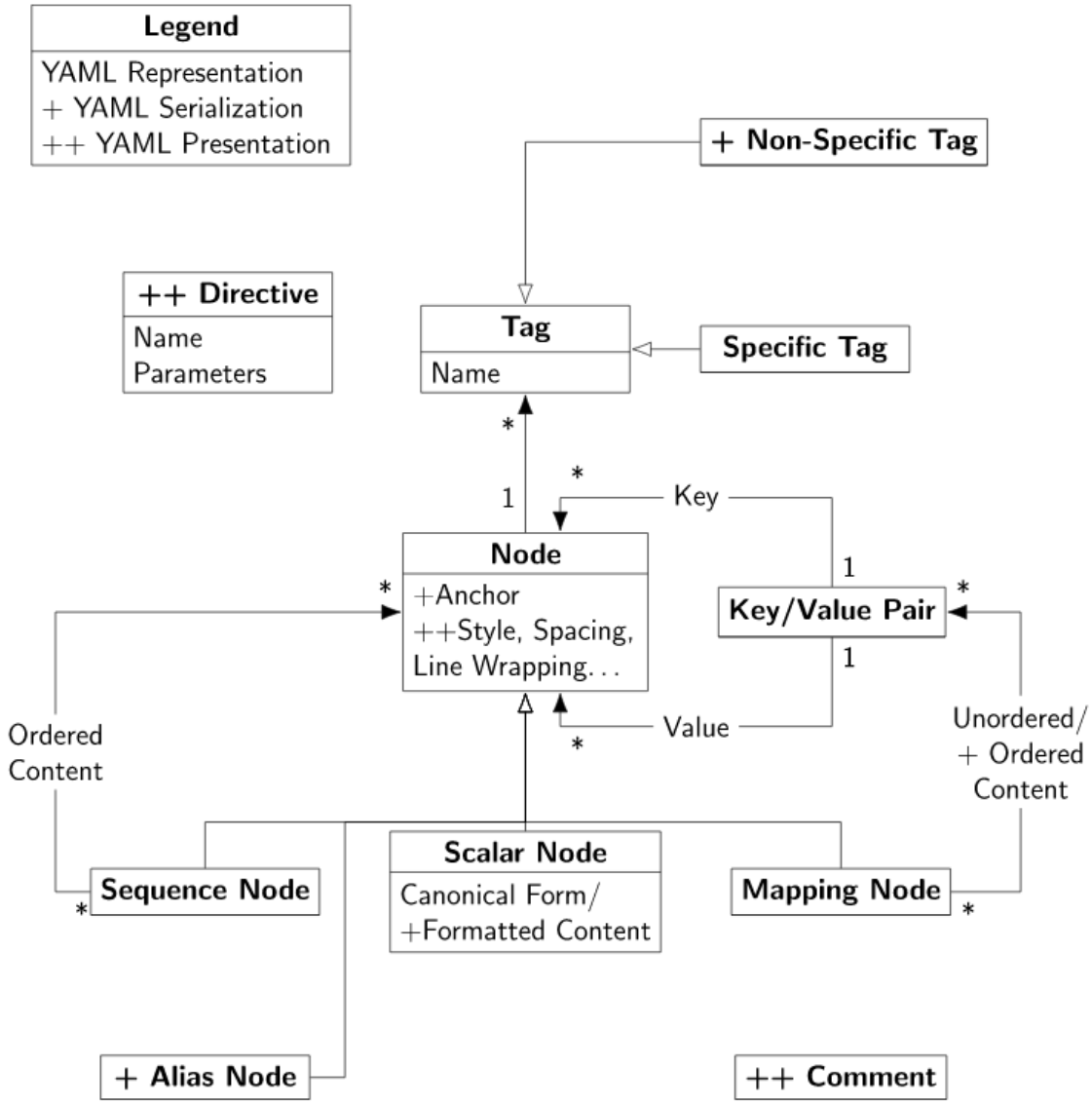
https://yaml.org/refcard.html

```
%YAML 1.1   # Reference card
---
Collection indicators:
  '? ' : Key indicator.
  ': ' : Value indicator.
  '- ' : Nested series entry indicator.
  ', ' : Separate in-line branch entries.
  '[]' : Surround in-line series branch.
  '{}' : Surround in-line keyed branch.
Scalar indicators:
  '''' : Surround in-line unescaped scalar ('' escaped ').
  '"'  : Surround in-line escaped scalar (see escape codes below).
  '|'  : Block scalar indicator.
  '>'  : Folded scalar indicator.
  '-'  : Strip chomp modifier ('|-' or '>-').
  '+'  : Keep chomp modifier ('|+' or '>+').
  1-9  : Explicit indentation modifier ('|1' or '>2').
         # Modifiers can be combined ('|2-', '>+1').
Alias indicators:
  '&'  : Anchor property.
  '*'  : Alias indicator.
Tag property: # Usually unspecified.
  none    : Unspecified tag (automatically resolved by application).
  '!'     : Non-specific tag (by default, "!!map"/"!!seq"/"!!str").
  '!foo'  : Primary (by convention, means a local "!foo" tag).
  '!!foo' : Secondary (by convention, means "tag:yaml.org,2002:foo").
  '!h!foo': Requires "%TAG !h! <prefix>" (and then means "<prefix>foo").
  '!<foo>': Verbatim tag (always means "foo").
Document indicators:
  '%'  : Directive indicator.
  '---': Document header.
  '...': Document terminator.
Misc indicators:
  ' #' : Throwaway comment indicator.
  '`@' : Both reserved for future use.
Special keys:
  '='  : Default "value" mapping key.
  '<<' : Merge keys from another mapping.
Core types: # Default automatic tags.
  '!!map' : { Hash table, dictionary, mapping }
  '!!seq' : { List, array, tuple, vector, sequence }
  '!!str' : Unicode string
More types:
  '!!set' : { cherries, plums, apples }
  '!!omap': [ one: 1, two: 2 ]
Language Independent Scalar types:
  { ~, null }              : Null (no value).
  [ 1234, 0x4D2, 02333 ]   : [ Decimal int, Hexadecimal int, Octal int ]
  [ 1_230.15, 12.3015e+02 ]: [ Fixed float, Exponential float ]
  [ .inf, -.Inf, .NAN ]    : [ Infinity (float), Negative, Not a number ]
  { Y, true, Yes, ON  }    : Boolean true
  { n, FALSE, No, off }    : Boolean false
  ? !!binary >
      R0lG...BADS=
  : >-
      Base 64 binary value.
Escape codes:
 Numeric   : { "\x12": 8-bit, "\u1234": 16-bit, "\U00102030": 32-bit }
 Protective: { "\\": '\', "\"": '"', "\ ": ' ', "\<TAB>": TAB }
 C         : { "\0": NUL, "\a": BEL, "\b": BS, "\f": FF, "\n": LF, "\r": CR,
             "\t": TAB, "\v": VTAB }
 Additional: { "\e": ESC, "\_": NBSP, "\N": NEL, "\L": LS, "\P": PS }
```

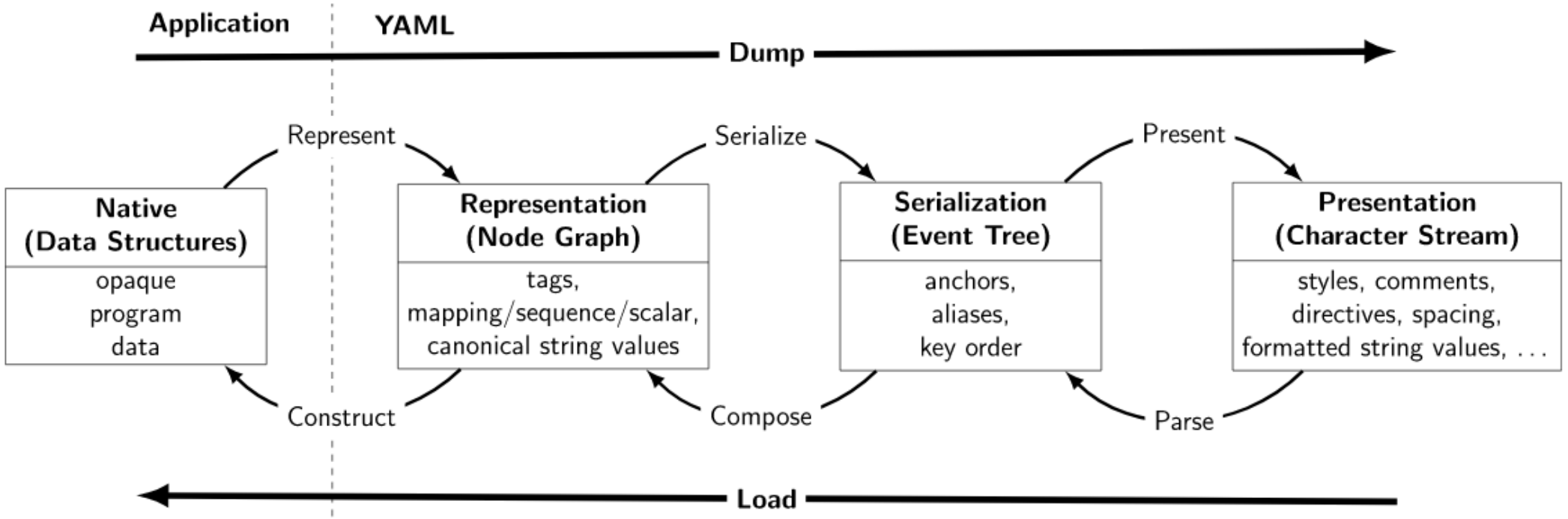
YAML Ain’t Markup Language (YAML)

YAML

Filename extensions	`.yaml`, `.yml`
Internet media type	*Not registered*
Initial release	11 May 2001; 20 years ago
Latest release	1.2 (Third Edition) (1 October 2009; 12 years ago)
Type of format	Data interchange
Open format?	Yes
Website	yaml.org



APIs for YAML Processing



- Tutorial for Java:
  https://www.baeldung.com/java-snake-yaml
- Package for JavaScript:
  https://www.npmjs.com/package/yaml
- Module for Python:
  https://pyyaml.org/wiki/PyYAMLDocumentation

Data Interoperability and Semantics

Part 2. Data Formats

Part 2.7. Lightweight markup languages

ICM – Toolbox Engineering and Interoperability of Software Systems – Course unit on Data Interoperability and Semantics
M1 Cyber Physical and Social Systems – Course unit on Data Interoperability and Semantics
Maxime Lefrançois https://maxime-lefrancois.info
Course unit URL: https://ci.mines-stetienne.fr/cps2/course/data

Lightweight markup languages

Comparing language features

Language	HTML export tool	HTML import tool	Tables	Link titles	`class` attribute	`id` attribute	Release date
AsciiDoc	Yes	Yes	Yes	Yes	Yes	Yes	2002-11-25[1]
BBCode	No	No	Yes	No	No	No	1998
Creole	No	No	Yes	No	No	No	2007-07-04[2]
Gemtext	Yes	?	No	Yes	No	No	2020
GitHub Flavored Markdown	Yes	No	Yes	Yes	No	No	2011-04-28+
Jira Formatting Notation	Yes	No	Yes	Yes	No	No	2002+[3]
Markdown	Yes	Yes	No	Yes	Yes/No	Yes/No	2004-03-19[4][5]
Markdown Extra	Yes	Yes	Yes[6]	Yes	Yes	Yes	2013-04-11[7]
MarkedText	Yes	No	Yes	No	No	No	2021-01
MediaWiki	Yes	Yes	Yes	Yes	Yes	Yes	2002[8]
MultiMarkdown	Yes	No	Yes	Yes	No	No	2009-07-13
Org-mode	Yes	Yes[9]	Yes	Yes	Yes	Yes	2003[10]
PmWiki	Yes[11]	Yes	Yes	Yes	Yes	Yes	2002-01
POD	Yes	?	No	Yes	?	?	1994
reStructuredText	Yes	Yes[9]	Yes	Yes	Yes	auto	2002-04-02[12]
Slack	No	No	No	Yes	No	No	2013+[13][14]
TiddlyWiki	Yes	No	Yes	Yes	Yes	No	2004-09[15]
Textile	Yes	No	Yes	Yes	Yes	Yes	2002-12-26[16]
Texy	Yes	Yes	Yes	Yes	Yes	Yes	2004[17]
txt2tags	Yes	Yes[18]	Yes[19]	Yes	Yes/No	Yes/No	2001-07-26[20]
WhatsApp	No	No	No	No	No	No	2016-03-16[21]

Lightweight markup languages

HTML output	strongly emphasized	emphasized text	<code>code</code>	semantic
	bold text	italic text	<tt>monospace text</tt>	presentational
AsciiDoc	*bold text*	'italic text'	+monospace text+	Can double operators to apply formatting where there is no word bounda example **b**old t**ex**t yields **b**old t**ex**t).
		italic text	`monospace text`	
ATX	*bold text*	_italic text_	|monospace text|	email style
BBCode	[b]bold text[/b]	[i]italic text[/i]	[code]monospace text[/code]	Formatting works across line breaks.
Creole	**bold text**	//italic text//	{{{monospace text}}}	Triple curly braces are for *nowiki* which is optionally monospace.
Gemtext	N/A	N/A	```alt text monospace text ```	Text immediately following the first three backticks is alt-text.
Jira Formatting Notation	*bold text*	_italic text_	{{monospace text}}	
Markdown[42]	**bold text**	*italic text*	`monospace text`	semantic HTML tags
	__bold text__	_italic text_		
MarkedText	**bold text**	//italic text//	;;monospace text;;	semantic HTML tags
MediaWiki	'''bold text'''	''italic text''	<code>monospace text</code>	mostly resorts to inline HTML
Org-mode	*bold text*	/italic text/	=code=	
			~verbatim~	
PmWiki	'''bold text'''	''italic text''	@@monospace text@@	
reST	**bold text**	*italic text*	``monospace text``	
Setext	**bold text**	~italic text~	N/A	
Textile[43]	*strong*	_emphasis_	@monospace text@	semantic HTML tags
	bold text	__italic text__		presentational HTML tags
Texy!	**bold text**	*italic text*	`monospace text`	semantic HTML tags by default, optional support for presentational tags
		//italic text//		
TiddlyWiki	''bold text''	//italic text//	`monospace text`	
			``monospace text``	
txt2tags	**bold text**	//italic text//	``monospace text``	
POD	B<bold text>	I<italic text>	C<monospace text>	Indented text is also shown as monospaced code.
Slack	*bold text*	_italic text_	`monospace text`	```block of monospaced text```
WhatsApp	*bold text*	_italic text_	```monospace text```	

Lightweight markup languages

Paragraphs and Line Breaks

- Para = one or more consecutive lines
- Empty lines = end of paragraph
- > 2 spaces at EoL = HTML line break

Headers

- # This is an H1
- ## This is an H2
- ###### This is an H6

Alternative: ==s for H1, --s for H2

Emphasis

- *This is an *
  and _so is this_
- **This is a **
  and __so is this__
- Use same closing marker as opening
- Emphasis possible in middle of word
- * or _ for a literal * or _

Blockquotes

- Prefix > for each blockquote line...
  ... or even hard-wrapped paragraph
- Additional > for nested blockquotes
- Blockquotes can contain Markdown

Images

- ![Alt text](URL "Title")
- ![Alt text][id]
  [id]: URL "Title"
- See syntax for Links

Links

Inline	★ This is [an example](http://example.com/) link ★ This is [example](http://example.com/ "Title") with a title
Reference	★ This is a [reference link][id] with an id [id]: http://example.com/ "Title" ★ This is a [reference link][] without an id [reference link]: http://example.com/ (Title)

- Link id can be letters, numbers, spaces, and punctuation
- Link id is case-insensitive
- In ref links, link definition can be indented up to 3 spaces
- In ref links, use parentheses, single or double quotes for title
- In ref links, link URL can be surrounded by < >
- In ref links, title can be on separate line with indentation
- In ref links, link definitions can appear anywhere in document

Lists

Unordered	Asterisks, pluses & hyphens interchangeably	+ Red * Green - Blue
Ordered	Numbers, not necessarily sequential. Always start with 1.	1. Red 1. Green 1. Blue

- List marker must be followed by 1/more spaces or a tab
- List marker can be indented upto 3 spaces
- Hanging indent supported but not required
- Blank lines between list items implies paragraph
- Paragraph start must be indented by 4 spaces or a tab
- To avoid unintended list numbering (e.g 1986. What a great season) backslash the period (1986\. What a great season).

Markdown

➔ http://daringfireball.net/projects/markdown/
Cheat sheet from: Ahren Code ➔ http://ahren.org/code/

Code Block

- Indent code by 4 spaces or 1 tab
- One level indentation will be removed
- & and <, > are automatically HTML'ised
- Markdown not usable in code blocks

Code Spans

- Use backticks around code: `printf()`
- Use multiple backticks for literal backtick: ``ls `date+%h` | wc -l``
  - Use space if literal backtick is at start of code span: `` ` ``

Miscellaneous

Horizontal Rule / Line
3 or more asterisks, hyphens or underscores

List with embedded Blockquote
Indent > delimiters:

```
* A list item with a blockquote:
  > This is a blockquote
```

List with embedded Code block
Indent code block 8 spaces or two tabs:

```
* A list item with a code block:
      ... some code ....
```

Automatic Links
- Use <, > for auto-linking URLs
- Use <, > for auto-linking email addrs
- Email text is auto obscured (somewhat)

For all else use HTML markup
- blank lines around block level elements
- no indentation for tags
- no Markdown inside HTML blocks

Data Interoperability and Semantics

Part 2. Data Formats

Part 2.8. Compressed formats

ICM – Toolbox Engineering and Interoperability of Software Systems – Course unit on Data Interoperability and Semantics
M1 Cyber Physical and Social Systems – Course unit on Data Interoperability and Semantics
Maxime Lefrançois https://maxime-lefrancois.info
Course unit URL: https://ci.mines-stetienne.fr/cps2/course/data

Compression

A compressed file is an archive that contains one or more file that have been reduced in size.

many different file compression types: **zip**, arc, arj, rar, cab, **tar.gz**, ...

lossless file compression

reduce redundancy without loosing data

ex. AAABBBBBCC -> A3B5C2

lossy file compression

ok to loose some data

ex. video, audio, images

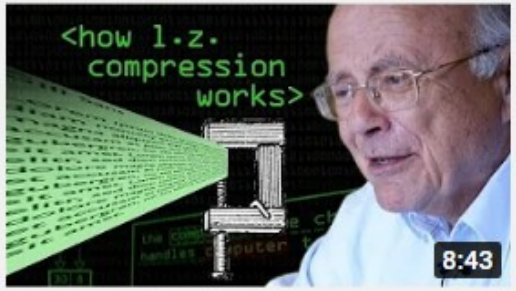
How compression works

Compression - Computerphile

Computerphile ✔ 384K views • 8 years ago

Most of us deal with data compression on a daily basis, but what is it and how does it work? Professor David Brailsford introduces compression with regards to text and pictures. http://www.faceboo...

CC



Elegant Compression in Text (The LZ 77 Method) - Computerphile

Computerphile ✔ 436K views • 8 years ago

Text compression methods such as LZ can reduce file sizes by up to 80%. Professor Brailsford explains the nuts and bolts of how it is done. Original Compression film: http://youtu.be/Lto-ajuqW3w...

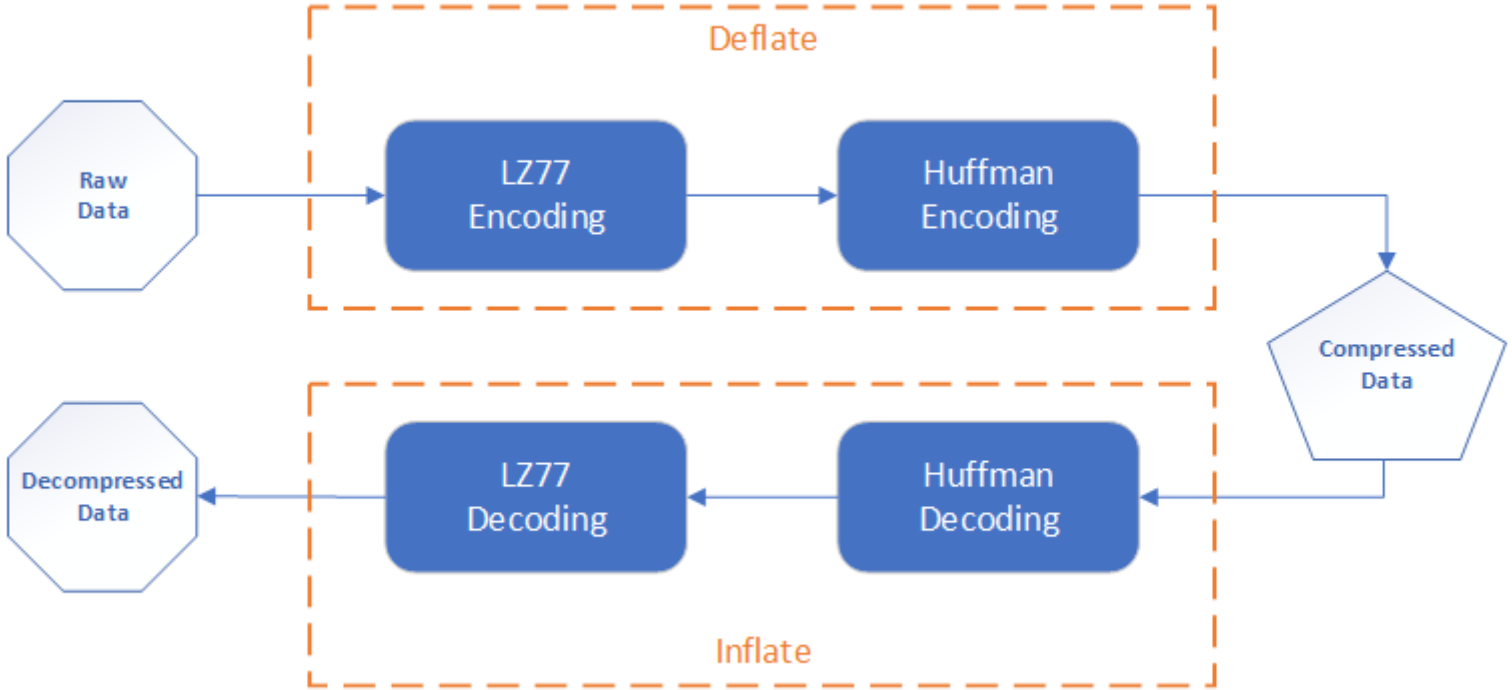
CC

How Huffman Trees Work - Computerphile

Computerphile ✔ 225K views • 8 years ago

How do we derive the most compact codes for a situation? Huffman Trees can help. Professor Brailsford explains how computer scientists like their trees to be upside down. "Entropy in Compression...

CC



gzip

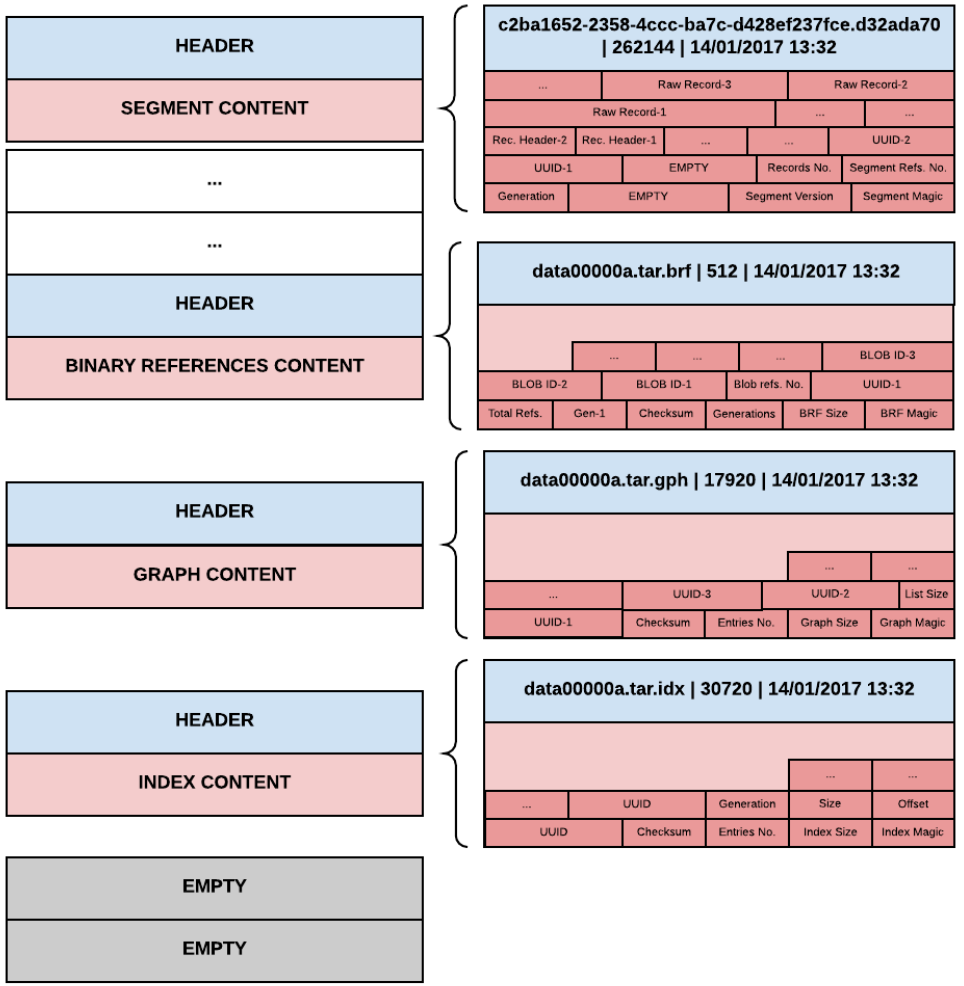
Filename extension	`.gz`
Internet media type	`application/gzip`[2]
Uniform Type Identifier (UTI)	org.gnu.gnu-zip-archive
Magic number	`1f 8b`
Developed by	Jean-loup Gailly and Mark Adler
Type of format	Data compression
Open format?	Yes
Website	gzip.org (obsolete)

ZIP file format

Filename extension	`.zip` `.zipx`
Internet media type	`application/zip`[1]
Uniform Type Identifier (UTI)	com.pkware.zip-archive
Magic number	*none* `PK\x03\x04` `PK\x05\x06` (empty) `PK\x07\x08` (spanned)
Developed by	PKWARE, Inc.
Initial release	14 February 1989; 32 years ago
Latest release	6.3.9 (15 July 2020; 15 months ago)
Type of format	Data compression
Extended to	JAR (EAR, RAR (Java), WAR) Office Open XML (Microsoft) Open Packaging Conventions OpenDocument (ODF) XPI (Mozilla extensions)
Standard	APPNOTE from PKWARE ISO/IEC 21320-1:2015 (a subset of ZIP file format 6.3.3)
Open format?	Yes

tar

Filename extension	`.tar`
Internet media type	`application/x-tar`
Uniform Type Identifier (UTI)	public.tar-archive
Magic number	`u s t a r \0 0 0` at byte offset 257 (for POSIX versions) `u s t a r \040 \040 \0` (for old GNU tar format)[1] absent in pre-POSIX versions
Latest release	various (various)
Type of format	File archiver
Standard	POSIX since POSIX.1, presently in the definition of *pax*[1]
Open format?	Yes



+

gzip

Filename extension	`.gz`
Internet media type	`application/gzip`[2]
Uniform Type Identifier (UTI)	org.gnu.gnu-zip-archive
Magic number	`1f 8b`
Developed by	Jean-loup Gailly and Mark Adler
Type of format	Data compression
Open format?	Yes
Website	gzip.org (obsolete)

Data Interoperability and Semantics

Part 2. Data Formats

Part 2.9 Multimedia formats

ICM – Toolbox Engineering and Interoperability of Software Systems – Course unit on Data Interoperability and Semantics
M1 Cyber Physical and Social Systems – Course unit on Data Interoperability and Semantics
Maxime Lefrançois https://maxime-lefrancois.info
Course unit URL: https://ci.mines-stetienne.fr/cps2/course/data

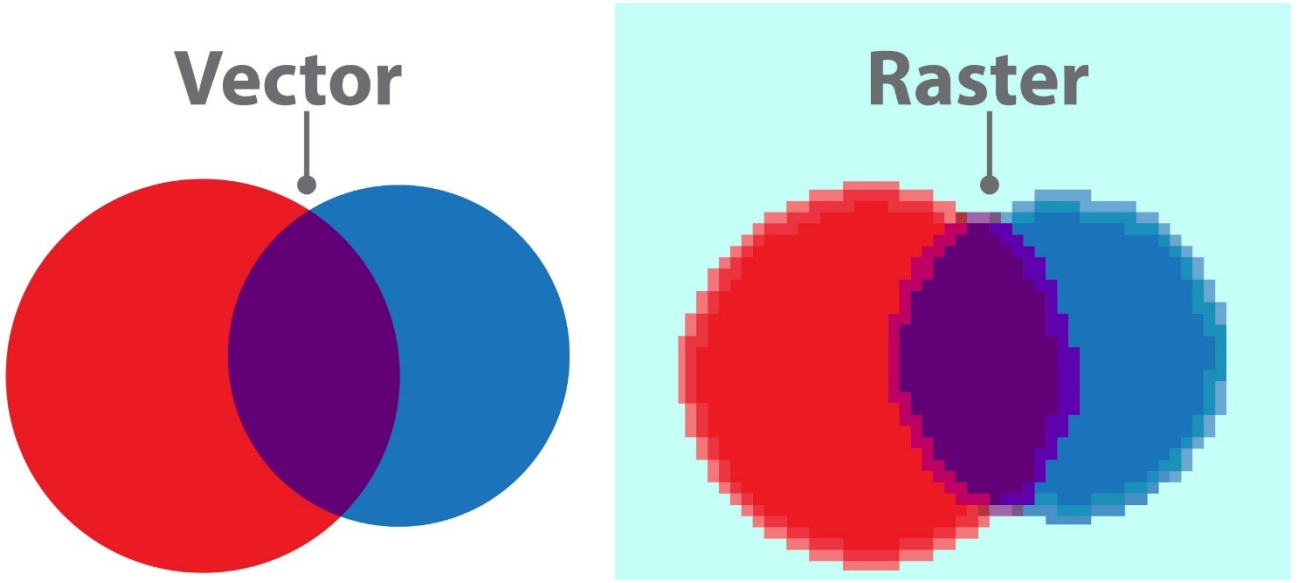
V · T · E	Multimedia compression and container formats	[hide]
Video compression	ISO, IEC, MPEG	MJPEG · Motion JPEG 2000 · MPEG-1 · MPEG-2 (Part 2) · MPEG-4 (Part 2 / ASP · Part 10 / AVC · Part 33 / IVC) · MPEG-H (Part 2 / HEVC) · MPEG-I (Part 3 / VVC) · MPEG-5 (Part 1 / EVC · Part 2 / LCEVC)
	ITU-T, VCEG	H.120 · DCT (H.261 · H.262 · H.263 · H.264 / AVC · H.265 / HEVC · H.266 / VVC · DV)
	SMPTE	VC-1 · VC-2 · VC-3 · VC-5 · VC-6
	TrueMotion	TrueMotion S · DCT (VP3 · VP6 · VP7 · VP8 · VP9 · AV1)
	Others	Apple Video · AVS · Bink · Cinepak · Daala · DVI · FFV1 · Huffyuv · Indeo · Lagarith · Microsoft Video 1 · MSU Lossless · OMS Video · Pixlet · ProRes (422 · 4444) · QuickTime (Animation · Graphics) · RealVideo · RTVideo · SheerVideo · Smacker · Sorenson Video/Spark · Theora · Thor · WMV · XEB · YULS
Audio compression	ISO, IEC, MPEG	MPEG-1 Layer II (Multichannel) · MPEG-1 Layer I · MPEG-1 Layer III (MP3) · AAC (HE-AAC · AAC-LD) · MPEG Surround · MPEG-4 ALS · MPEG-4 SLS · MPEG-4 DST · MPEG-4 HVXC · MPEG-4 CELP · MPEG-D USAC · MPEG-H 3D Audio
	ITU-T	G.711 (A-law · μ-law) · G.718 · G.719 · G.722 · G.722.1 · G.722.2 · G.723 · G.723.1 · G.726 · G.728 · G.729 · G.729.1
	IETF	Opus · iLBC · Speex · Vorbis
	3GPP	AMR · AMR-WB · AMR-WB+ · EVRC · EVRC-B · EVS · GSM-HR · GSM-FR · GSM-EFR
	ETSI	AC-3 · AC-4 · DTS
	Others	ACELP · ALAC · Asao · ATRAC · AVS · CELT · Codec 2 · DRA · FLAC · iSAC · MELP · Monkey's Audio · MT9 · Musepack · OptimFROG · OSQ · QCELP · RCELP · RealAudio · RTAudio · SBC · SD2 · SHN · SILK · Siren · SMV · SVOPC · TTA (True Audio) · TwinVQ · VMR-WB · VSELP · WavPack · WMA · MQA · aptX · aptX HD · aptX Low Latency · aptX Adaptive · LDAC · LHDC · LLAC
Image compression	IEC, ISO, IETF, W3C, ITU-T, JPEG	CCITT Group 4 · DCT (HEIC · HEVC · JPEG · JPEG XL · JPEG XR · JPEG XT · TIFF/EP) · Arithmetic (JBIG · JBIG2) · JPEG-LS · JPEG XS · JPEG 2000 · LZ (GIF · PNG) · TIFF · TIFF/IT
	Others	APNG · BPG · DCT (AVIF · AV1) · DjVu · EXR · FLIF · ICER · MNG · PGF · QTVR · WBMP · WebP
Containers	ISO, IEC	MPEG-ES (MPEG-PES) · MPEG-PS · MPEG-TS · ISO/IEC base media file format · MPEG-4 Part 14 (MP4) · Motion JPEG 2000 · MPEG-21 Part 9 · MPEG media transport
	ITU-T	H.222.0 · T.802
	IETF	RTP · Ogg
	SMPTE	GXF · MXF
	Others	3GP and 3G2 · AMV · ASF · AIFF · AVI · AU · BPG · Bink (Smacker) · BMP · DivX Media Format · EVO · Flash Video · HEIF · IFF · M2TS · Matroska (WebM) · QuickTime File Format · RatDVD · RealMedia · RIFF (**WAV**) · MOD and TOD · VOB, IFO and BUP
Collaborations	NETVC · MPEG LA · HEVC Advance · Alliance for Open Media	
Methods	Discrete cosine transform (DCT · MDCT) · Entropy (Arithmetic · Huffman · Modified) · FFT · LPC (ACELP · CELP · LSP · WLPC) · Lossless · Lossy · LZ (DEFLATE · LZW) · PCM (A-law · μ-law · ADPCM · DPCM) · Transform · Wavelet (Daubechies · DWT · Transform)	
Lists	Comparison of audio coding formats · Comparison of video codecs · List of codecs	
See Compression methods for techniques and Compression software for codecs		

Authority control: National libraries · France (data)

https://en.wikipedia.org/wiki/Template:Compression_formats

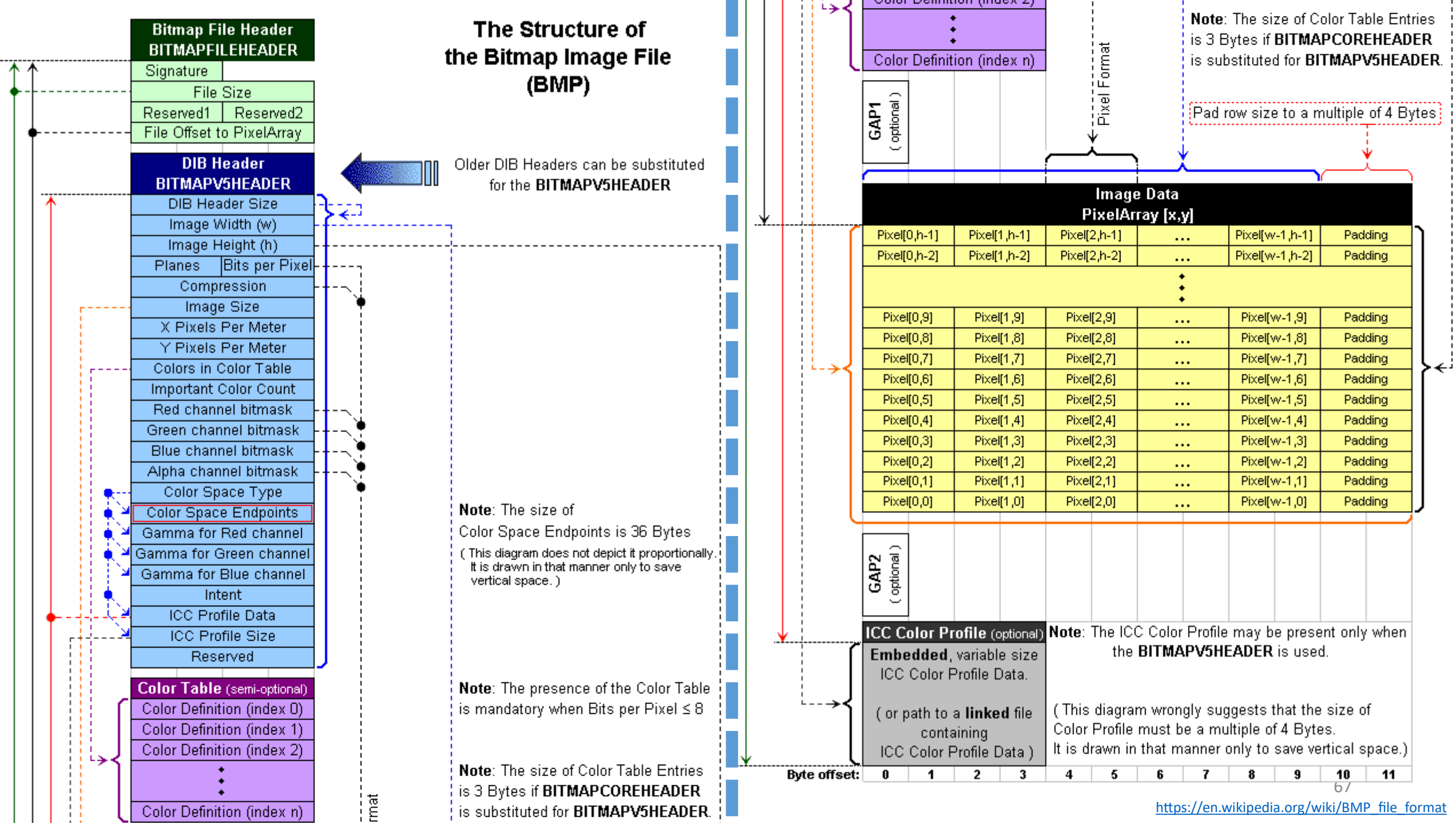
Image file formats

https://en.wikipedia.org/wiki/Image_file_formats



V · T · E	Graphics file formats [hide]
Raster	ANI · ANIM · APNG · ART · AVIF · BMP · BPG · BSAVE · CAL · CIN · CPC · CPT · DDS · DPX · ECW · EXR · FITS · FLIC · FLIF · FPX · GIF · HDRi · HEVC · ICER · ICNS · ICO / CUR · ICS · ILBM · JBIG · JBIG2 · JNG · JPEG · JPEG-LS · JPEG 2000 · JPEG XL · JPEG XR · JPEG XS · JPEG XT (JPEG-HDR) · KRA · MNG · MIFF · NRRD · PAM · PBM / PGM / PPM / PNM · PCX · PGF · PICtor · PNG · PSD / PSB · PSP · QTVR · RAS · RGBE (Logluv TIFF) · SGI · TGA · TIFF (TIFF/EP · TIFF/IT) · UFO / UFP · WBMP · WebP · XBM · XCF · XPM · XWD
Raw	CIFF · DNG
Vector	AI · CDR · CGM · DXF · EVA · EMF · EMF+ · Gerber · HVIF · IGES · PGML · SVG · VML · WMF · Xar
Compound	CDF · DjVu · EPS · PDF · PICT · PS · SWF · XAML
Metadata	Exchangeable image file format (Exif) · International Press Telecommunications Council § Photo metadata · Extensible Metadata Platform (XMP) · GIF § Metadata · Steganography
	Category · Comparison

The Structure of the Bitmap Image File (BMP)



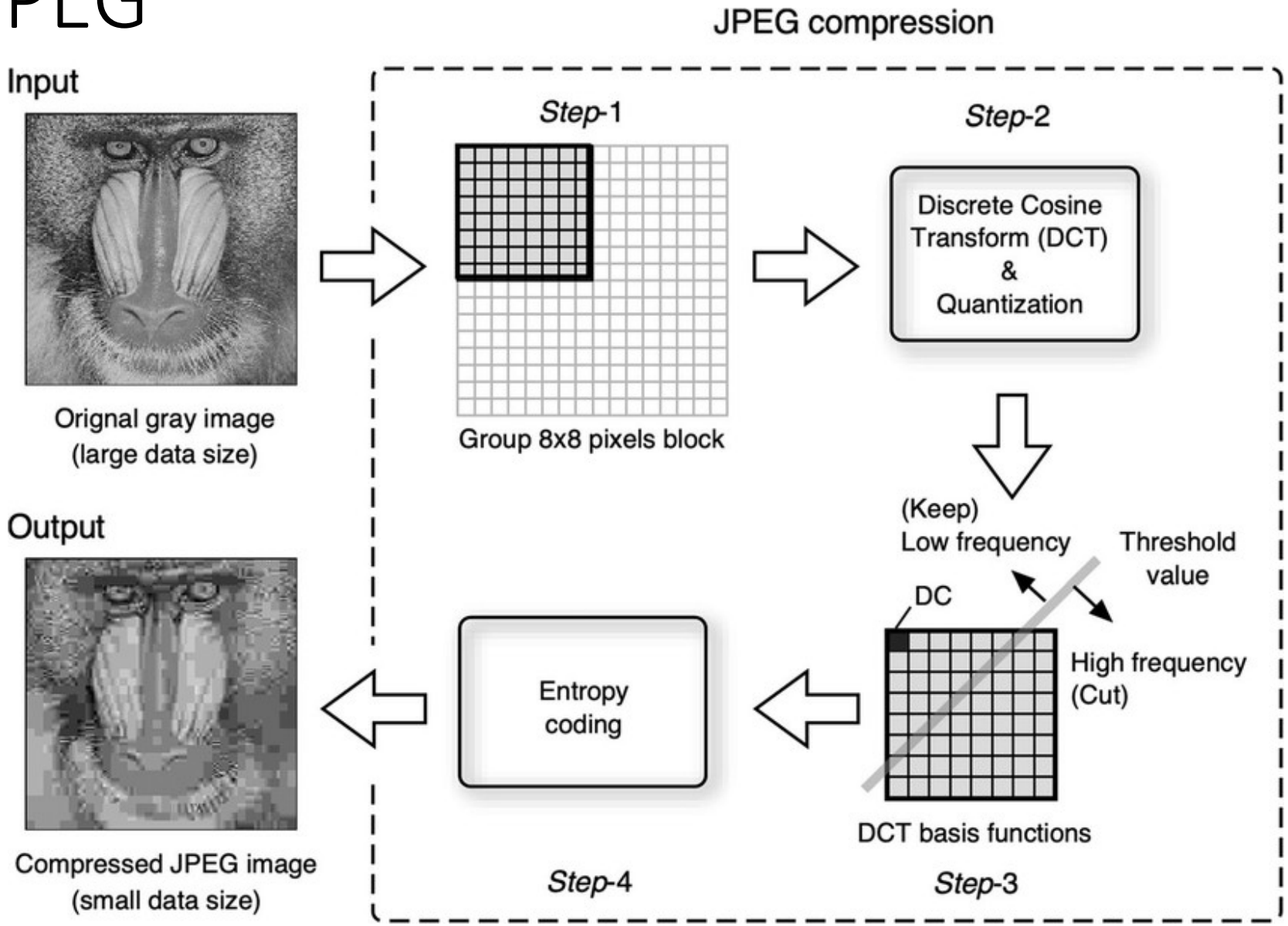
https://en.wikipedia.org/wiki/BMP_file_format

JPEG

JPEG

A photo of a European wildcat with the quality increasing, from left to right

Filename extension	`.jpg` , `.jpeg` , `.jpe` `.jif` , `.jfif` , `.jfi`
Internet media type	`image/jpeg`
Type code	`JPEG`
Uniform Type Identifier (UTI)	public.jpeg
Magic number	`ff d8 ff`
Developed by	Joint Photographic Experts Group, IBM, Mitsubishi Electric, AT&T, Canon Inc.,[1] ITU-T Study Group 16
Initial release	September 18, 1992; 29 years ago
Type of format	Lossy image compression format
Standard	ITU-T T.81, ITU-T T.83, ITU-T T.84, ITU-T T.86, ISO/IEC 10918
Website	www.jpeg.org/jpeg/



Portable Network Graphics

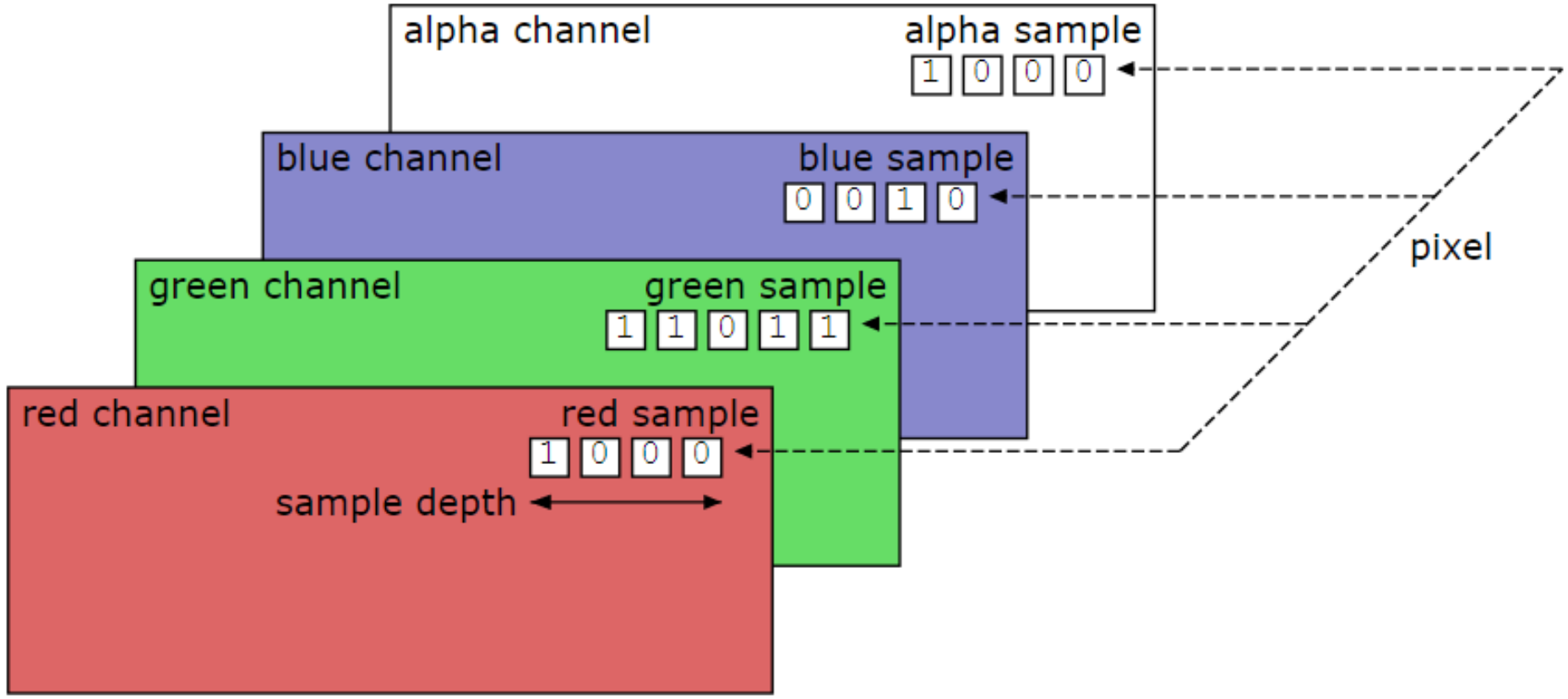
A PNG image with an 8-bit transparency channel, overlaid onto a checkered background, typically used in graphics software to indicate transparency

Filename extension	`.png`
Internet media type	`image/png`
Type code	PNGf PNG
Uniform Type Identifier (UTI)	public.png
UTI conformation	public.image
Magic number	`89 50 4e 47 0d 0a 1a 0a`
Developed by	PNG Development Group (donated to W3C)
Initial release	1 October 1996; 25 years ago
Type of format	Lossless bitmap image format
Extended to	APNG, JNG and MNG
Standard	ISO/IEC 15948,[1] IETF RFC 2083
Open format?	Yes

PNG

W3C REC and ISO/IEC 15948:2003

http://www.w3.org/TR/PNG



Portable Network Graphics

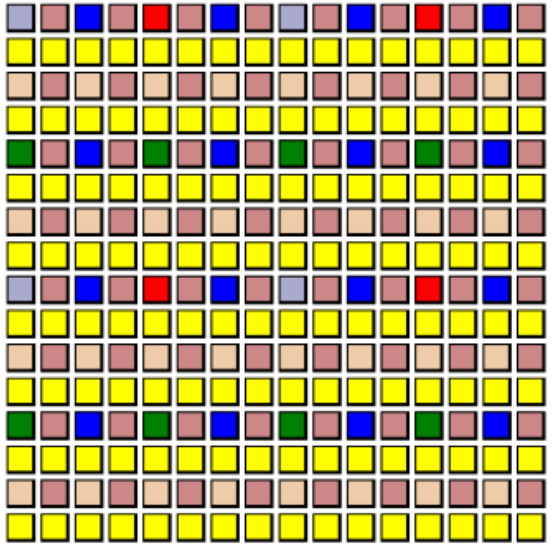
A PNG image with an 8-bit transparency channel, overlaid onto a checkered background, typically used in graphics software to indicate transparency

Filename extension	`.png`
Internet media type	`image/png`
Type code	PNGf PNG
Uniform Type Identifier (UTI)	public.png
UTI conformation	public.image
Magic number	`89 50 4e 47 0d 0a 1a 0a`
Developed by	PNG Development Group (donated to W3C)
Initial release	1 October 1996; 25 years ago
Type of format	Lossless bitmap image format
Extended to	APNG, JNG and MNG
Standard	ISO/IEC 15948,[1] IETF RFC 2083
Open format?	Yes

PNG

W3C REC and ISO/IEC 15948:2003

http://www.w3.org/TR/PNG



First reduced image

Second reduced image

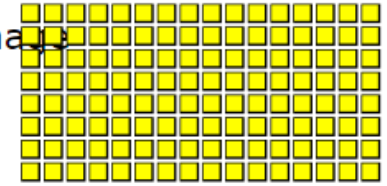
Third reduced image

Fourth reduced image

Fifth reduced image

Sixth reduced image

Seventh reduced image



Modern image file formats: WebP

- WebP lossless images are 26% smaller in size compared to PNGs.
- WebP lossy images are 25-34% smaller than comparable JPEG images at equivalent structural similarity index measure quality index.

Image	File Name	Original Size	Compressed JPG	WebP Format
	jpg-to-webp-1.jpg	480 KB	407 KB	43 KB
	jpg-to-webp-2.jpg	659 KB	578 KB	113 KB
	jpg-to-webp-3.jpg	787 KB	715 KB	127 KB
	jpg-to-webp-4.jpg	617 KB	543 KB	61 KB

WebP

An example WebP image

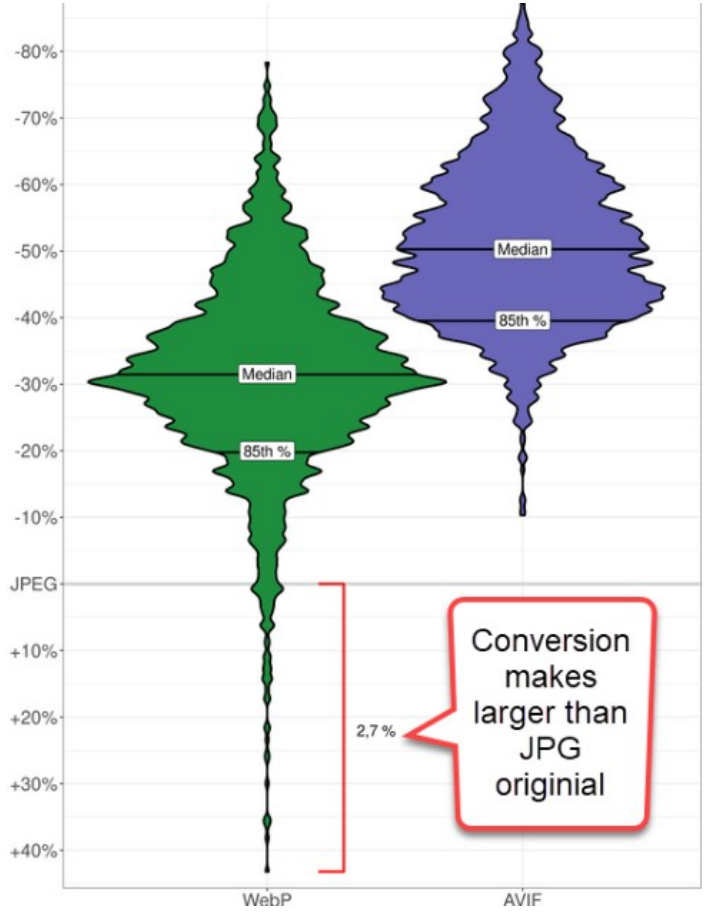
Filename extension	.webp[1]
Internet media type	image/webp
Magic number	52 49 46 46 xx xx xx xx 57 45 42 50
Developed by	Google
Initial release	30 September 2010; 11 years ago[2]
Type of format	Image format Lossless/lossy compression algorithm
Contained by	Resource Interchange File Format (RIFF)[3]
Open format?	Yes[4]
Website	developers.google.com/speed/webp

Modern image file formats: AVIF

AV1 Image File Format (AVIF)

Filename extension	`.avif, .avifs; .heif, .heifs; .hif`
Internet media type	`image/avif` , `image/avif-sequence`
Developed by	Alliance for Open Media
Initial release	v1.0.0, 19 February 2019
Type of format	Image format Lossless/lossy compression algorithm
Contained by	HEIF
Extended from	HEIF, ISOBMFF, AV1
Open format?	Yes
Website	aomediacodec.github.io/av1-avif/

- In most cases, AVIF generates a smaller image payload than WebP.
- In a few cases, WebP generates a larger version than the original JPEG.
- ImageEngine will recognize and deliver the most format with the smallest possible payload.

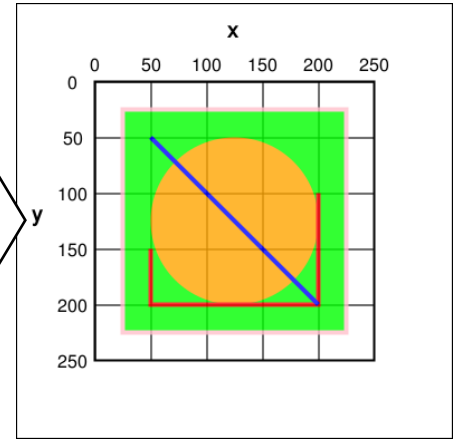


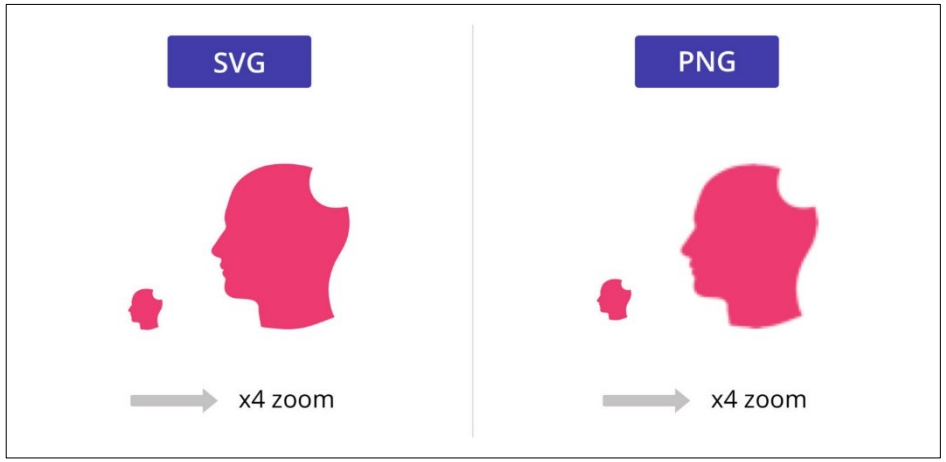
https://imageengine.io/blog/how-efficient-is-avif/

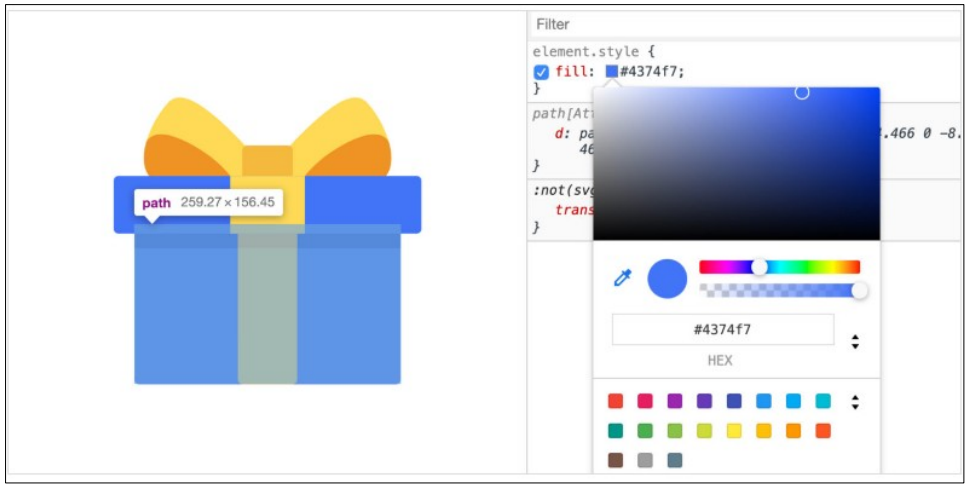
Scalable Vector Graphics

Internet media type	`image/svg+xml` [1][2]
Uniform Type Identifier (UTI)	`public.svg-image`
Developed by	W3C
Initial release	4 September 2001 (20 years ago)
Latest release	1.1 (Second Edition) (16 August 2011; 10 years ago)
Type of format	Vector graphics
Extended from	XML
Standard	W3C SVG
Open format?	Yes
Website	www.w3.org/Graphics/SVG/

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<!DOCTYPE svg PUBLIC "-//W3C//DTD SVG 1.1//EN" "http://www.w3.org/Graphics/SVG/1.1/DTD/svg11.dtd">
<svg width="391" height="391" viewBox="-70.5 -70.5 391 391" xmlns="http://www.w3.org/2000/svg"
xmlns:xlink="http://www.w3.org/1999/xlink">
<rect fill="#fff" stroke="#000" x="-70" y="-70" width="390" height="390"/>
<g opacity="0.8">
    <rect x="25" y="25" width="200" height="200" fill="lime" stroke-width="4" stroke="pink" />
    <circle cx="125" cy="125" r="75" fill="orange" />
    <polyline points="50,150 50,200 200,200 200,100" stroke="red" stroke-width="4" fill="none" />
    <line x1="50" y1="50" x2="200" y2="200" stroke="blue" stroke-width="4" />
</g>
</svg>
```







What is an SVG File Used For and Why Developers Should be Using Them
Published Jan 19, 2021 - https://deliciousbrains.com/svg-advantages-developers/

Recommendations for images in the browser

https://developer.mozilla.org/en-US/docs/Web/Media/Formats

Photographs

- AVIF*, WebP, or JPEG

Icons

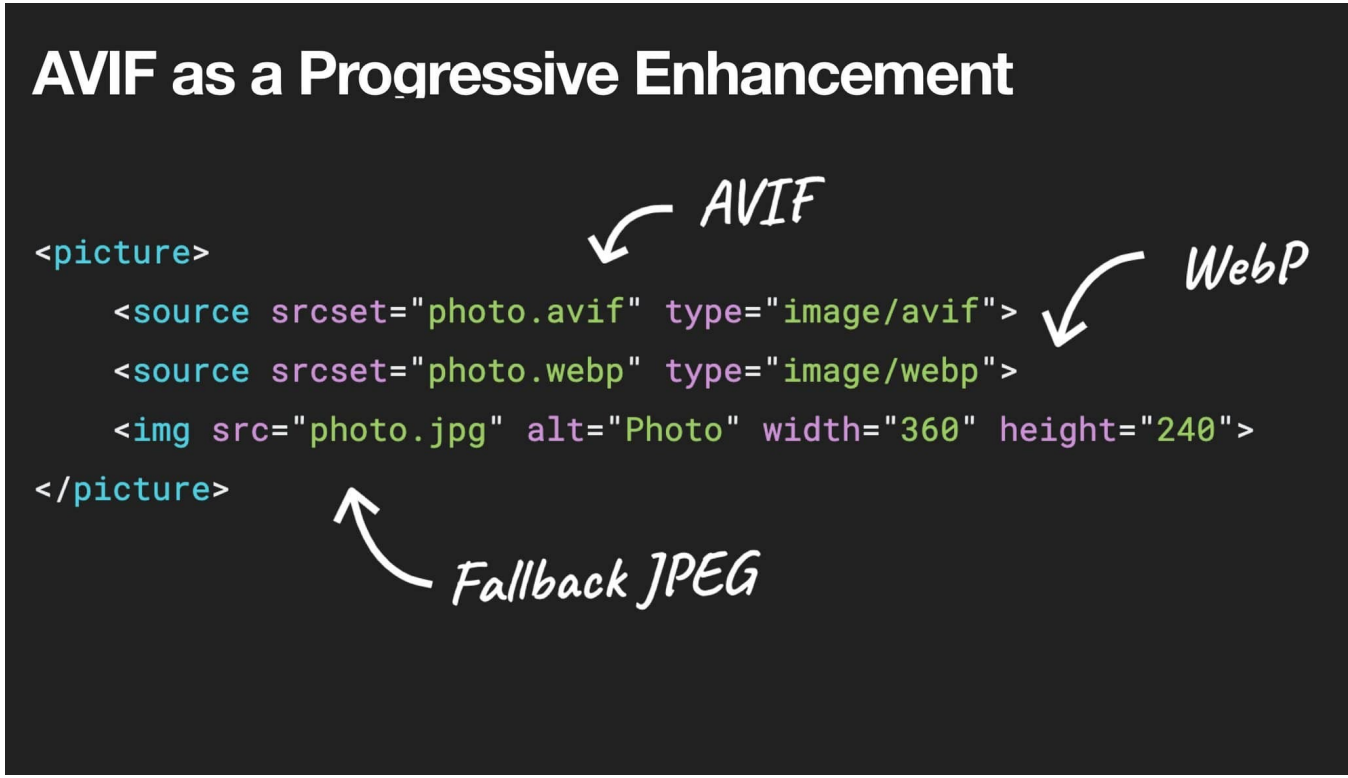
- SVG, Lossless WebP, or PNG

Screenshots

- Lossless WebP or PNG

Diagrams, drawings, and charts

- SVG, PNG



* I added AVIF here

Audio file formats

https://en.wikipedia.org/wiki/Audio_file_format

Waveform Audio File Format (WAVE/WAV)

Filename extension	`.wav` `.wave`
Internet media type	`audio/vnd.wave`,[1] `audio/wav`, `audio/wave`, `audio/x-wav`[2]
Type code	WAVE
Uniform Type Identifier (UTI)	com.microsoft.waveform-audio
Developed by	IBM & Microsoft
Initial release	August 1991; 30 years ago[3]
Latest release	Multiple Channel Audio Data and WAVE Files (7 March 2007; 14 years ago (update)[4][5])
Type of format	audio file format, container format
Extended from	RIFF
Extended to	BWF, RF64

MP3

Filename extension	`.mp3` `.bit` (before 1995)[1]
Internet media type	`audio/mpeg`[2] `audio/MPA`[3] `audio/mpa-robust`[4]
Developed by	Karlheinz Brandenburg, Ernst Eberlein, Heinz Gerhäuser, Bernhard Grill, Jürgen Herre and Harald Popp (all of Fraunhofer Society),[5] and others
Initial release	1991; 30 years ago
Type of format	Digital audio
Contained by	MPEG-ES
Standards	ISO/IEC 11172-3[6] ISO/IEC 13818-3[7]
Open format?	Yes[8]

Advanced Audio Coding

Filename extension	**MPEG/3GPP container** • `.m4a`, `.mp4`, `.3gp` **Apple container** • `.m4a`, `.m4b`, `.m4p`, `.m4r`, `.m4v` **ADTS stream** • `.aac`
Internet media type	`audio/aac` `audio/aacp` `audio/3gpp` `audio/3gpp2` `audio/mp4` `audio/mp4a-latm` `audio/mpeg4-generic`
Developed by	Bell, Fraunhofer, Dolby, Sony, Nokia, LG Electronics, NEC, NTT Docomo, Panasonic[1]
Initial release	1997; 24 years ago[2]
Type of format	Audio compression format, lossy compression
Contained by	MPEG-4 Part 14, 3GP and 3G2, ISO base media file format and Audio Data Interchange Format (ADIF)
Standard	ISO/IEC 13818-7, ISO/IEC 14496-3

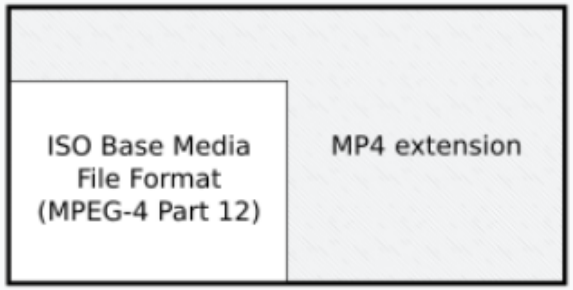
Video file formats

Audio Video Interleave

Filename extension	`.avi`
Internet media type	`video/vnd.avi`[1] `video/avi` `video/msvideo` `video/x-msvideo`
Type code	'Vfw '
Uniform Type Identifier (UTI)	public.avi
Developed by	Microsoft
Initial release	November 10, 1992; 29 years ago
Container for	Audio, Video
Extended from	Resource Interchange File Format
Open format?	No
Website	https://docs.microsoft.com/en-us/windows/win32/directshow/avi-file-format

MPEG-4 Part 14

MPEG-4 Part 14 - MP4 File Format



MPEG-4 Part 14 extends over ISO Base Media File Format (MPEG-4 Part 12).[1]

Filename extension	`.mp4`, `.m4a`, `.m4p`, `.m4b`, `.m4r` and `.m4v`[Note 1]
Internet media type	`video/mp4` `audio/mp4`
Type code	mpg4
Developed by	International Organization for Standardization
Initial release	2001; 20 years ago
Type of format	Media container
Container for	Audio, video and text
Extended from	QuickTime File Format and MPEG-4 Part 12
Standard	ISO/IEC 14496-14
Open format?	Yes

3GP

Filename extension	`.3gp`
Internet media type	`Video/3gpp`, `audio/3gpp`
Uniform Type Identifier (UTI)	public.3gpp
Developed by	3GPP
Type of format	media container
Container for	audio, video, text
Extended from	MPEG-4 Part 12

3G2

Filename extension	`.3g2`
Internet media type	`video/3gpp2`, `audio/3gpp2`
Uniform Type Identifier (UTI)	public.3gpp2
Developed by	3GPP2
Type of format	media container
Container for	audio, video, text
Extended from	MPEG-4 Part 12

WebM

Filename extension	`.webm`
Internet media type	`video/webm`, `audio/webm`
Developed by	Initially On2, Xiph, and Matroska; later Google
Initial release	May 18, 2010; 11 years ago[1]
Latest release	v1.9.0[2] (December 19, 2019; 22 months ago)
Type of format	Video file format
Container for	VP8/VP9/AV1 (video) Vorbis/Opus (audio)
Extended from	Limited *subset* of Matroska
Open format?	Yes[3]
Website	www.webmproject.org

Recommendations for audio and video files

https://developer.mozilla.org/en-US/docs/Web/Media/Formats

Audio-only files

If you need...	Consider using this container format
Compressed files for general-purpose playback	MP3 (MPEG-1 Audio Layer III)
Losslessly compressed files	FLAC with ALAC fallback
Uncompressed files	WAV

Video files

If you need...	Consider using this container format
General purpose video, preferably in an open format	WebM (ideally with MP4 fallback)
General purpose video	MP4 (ideally with WebM or Ogg fallback)
High compression optimized for slow connections	3GP (ideally with MP4 fallback)
Compatibility with older devices/browsers	QuickTime (ideally with AVI and/or MPEG-2 fallback)

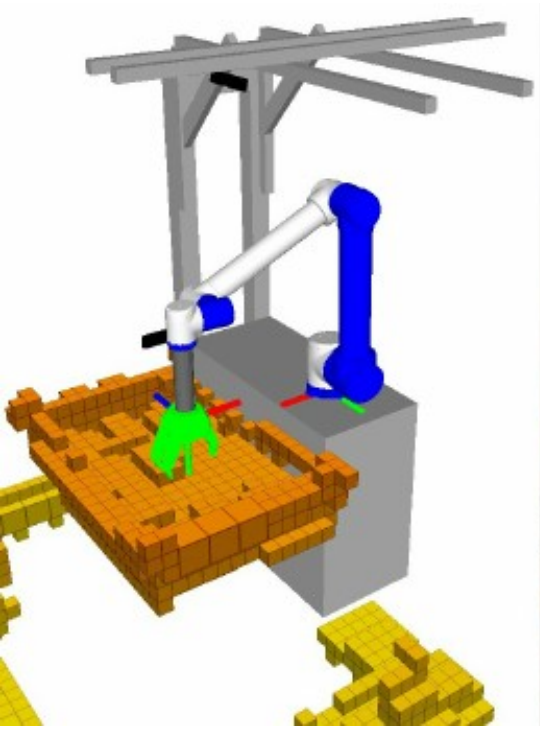
Data Interoperability and Semantics

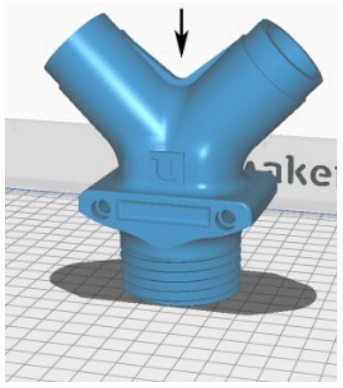
Part 2. Data Formats

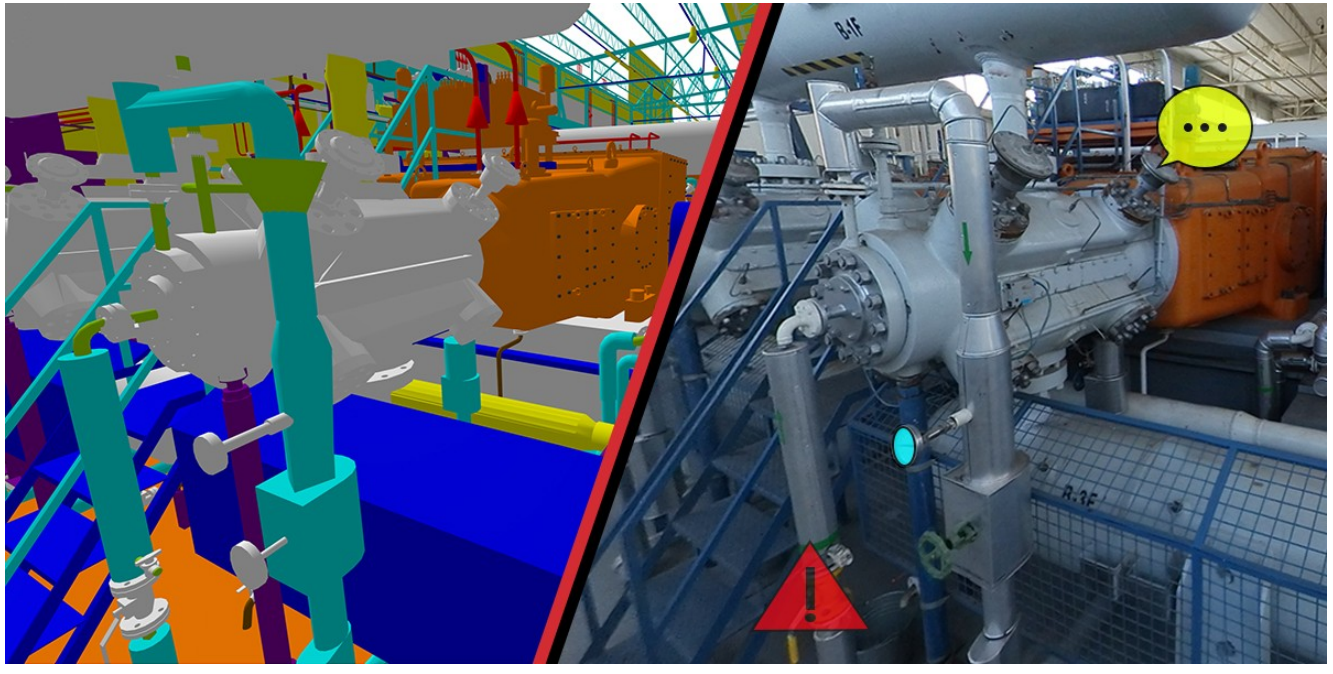
Part 2.10 3D models

(focusing on Web and Cyber-Physical Systems application domains)

ICM – Toolbox Engineering and Interoperability of Software Systems – Course unit on Data Interoperability and Semantics
M1 Cyber Physical and Social Systems – Course unit on Data Interoperability and Semantics
Maxime Lefrançois https://maxime-lefrancois.info
Course unit URL: https://ci.mines-stetienne.fr/cps2/course/data

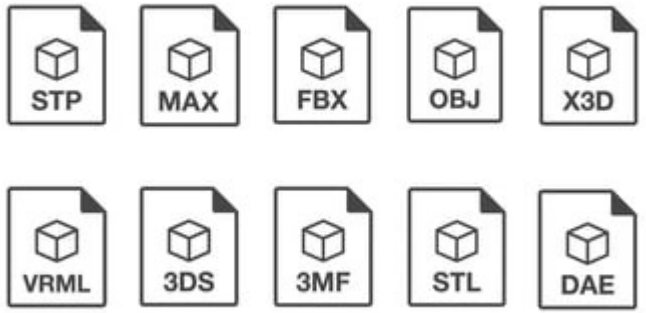






3D models

Many file formats



Key features of a 3D file:

- geometry,
- surface texture,
- scene details,
- animation of the model
- element properties,

https://all3dp.com/2/most-common-3d-file-formats-model/

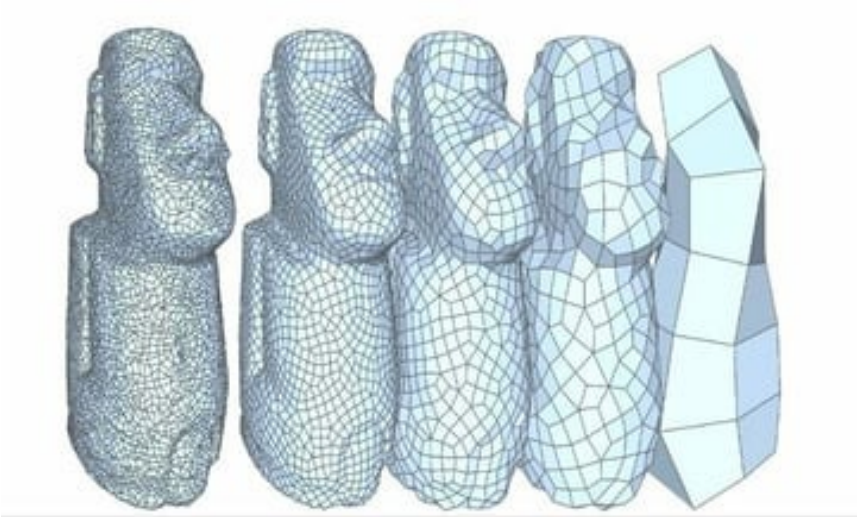
Shape Geometry

a) Approximate Mesh Encoding

« tesselations »: decompose a surface in polygons (usually triangles)

Good for 3D printing

Files may be huge.



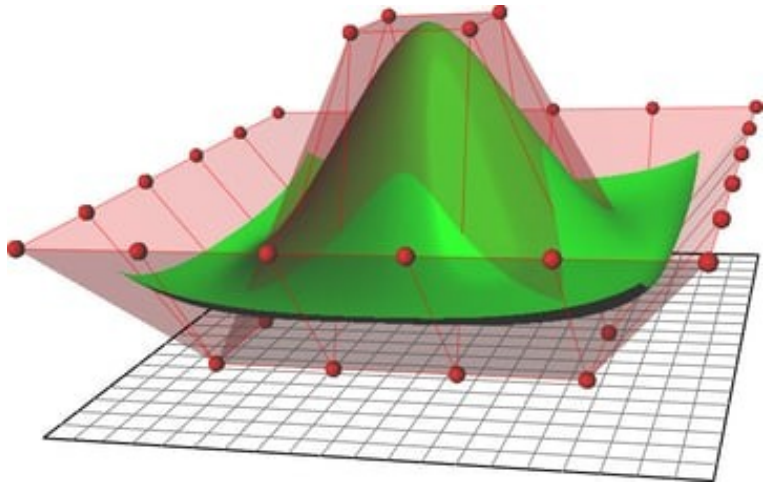
https://www.digitalpatterning.net/dpblog/how-do-i-know-the-pattern-is-right

b) Precise Mesh Encoding

Parametric surfaces made of control points and knots. Example: non-uniform rational basis spline (NURBS)

Exact at any resolution

Slow rendering

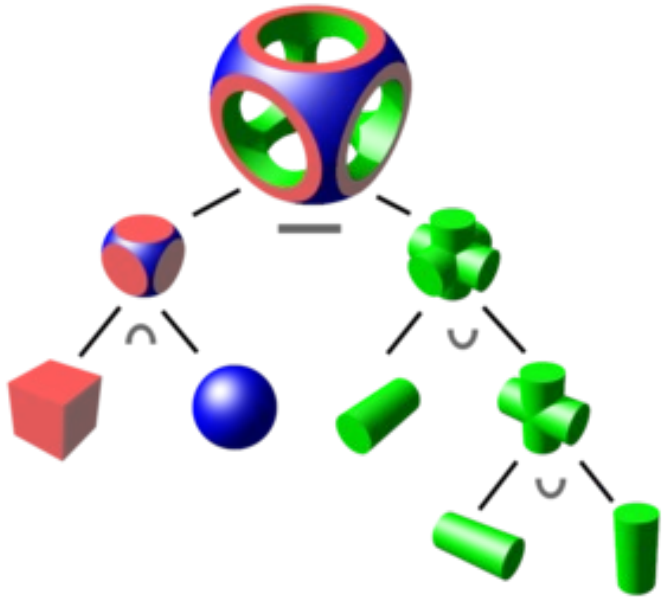


https://en.wikipedia.org/wiki/Non-uniform_rational_B-spline

c) Constructive Solid Geometry

Primitive shapes that are combined using Boolean operations

Good for CAD



https://en.wikipedia.org/wiki/Constructive_solid_geometry

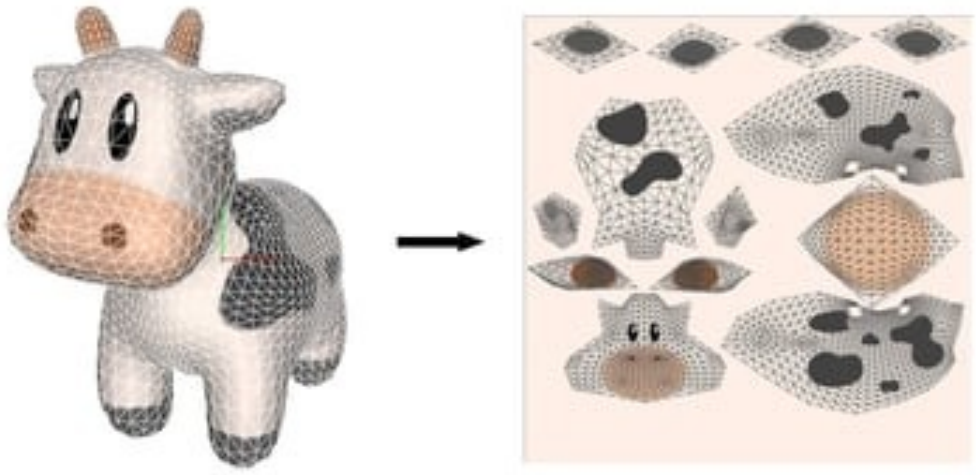
https://all3dp.com/2/most-common-3d-file-formats-model/

Surface Textures

https://www.creativeblog.com/3d-tips/find-high-res-textures-1232646

a) Texture Mapping

every point in the 3D model's surface (or the polygonal mesh) is mapped to a two-dimensional image.



https://metalbyexample.com/textures-and-samplers/

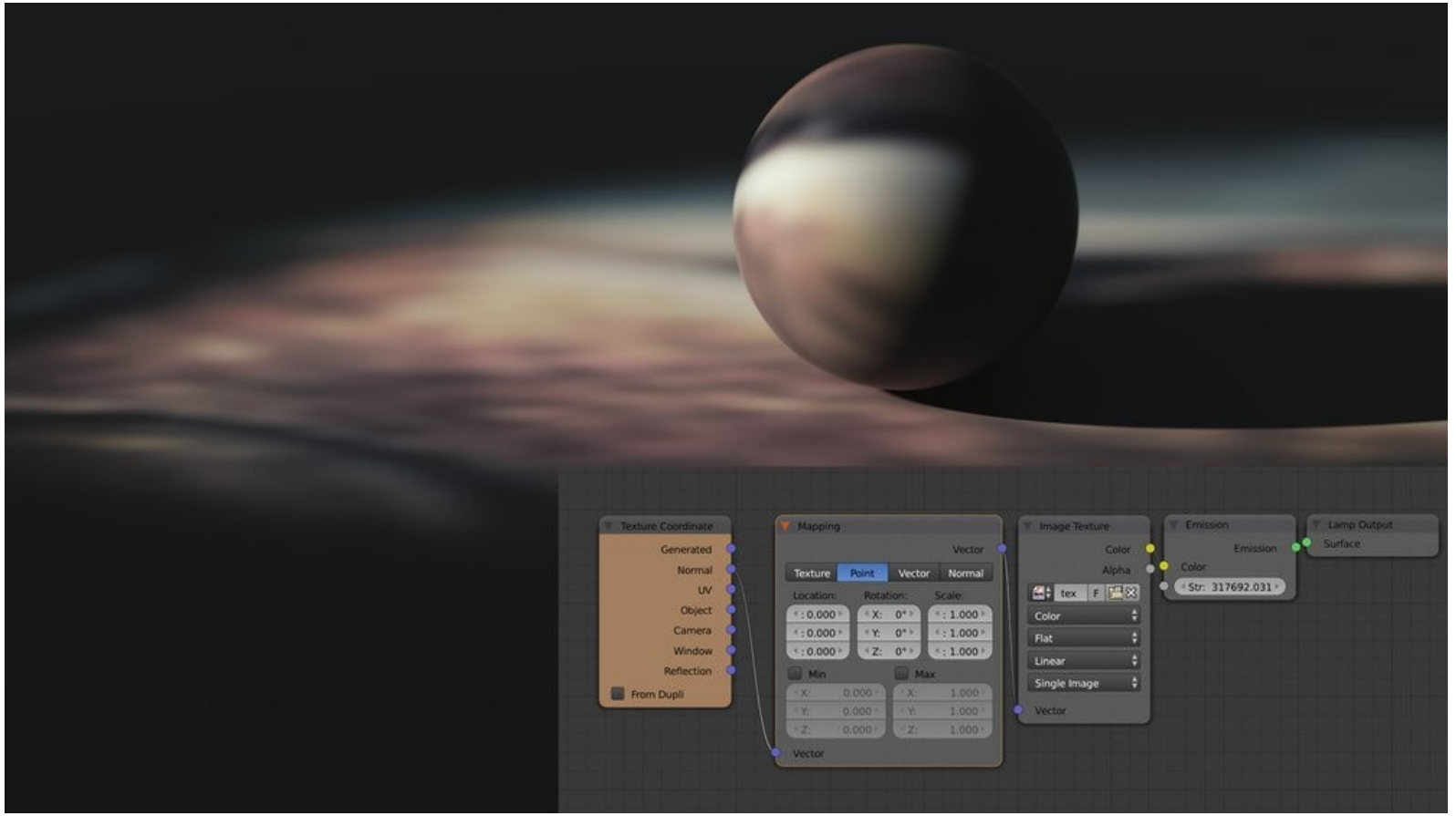
b) Face Attributes

each face of the mesh has a set of attributes
Color, texture, material type, reflection, refraction,…

https://www.blendernation.com/2021/11/05/realistic-glass-shader-in-blender-eevee-tutorial/

Scene detail

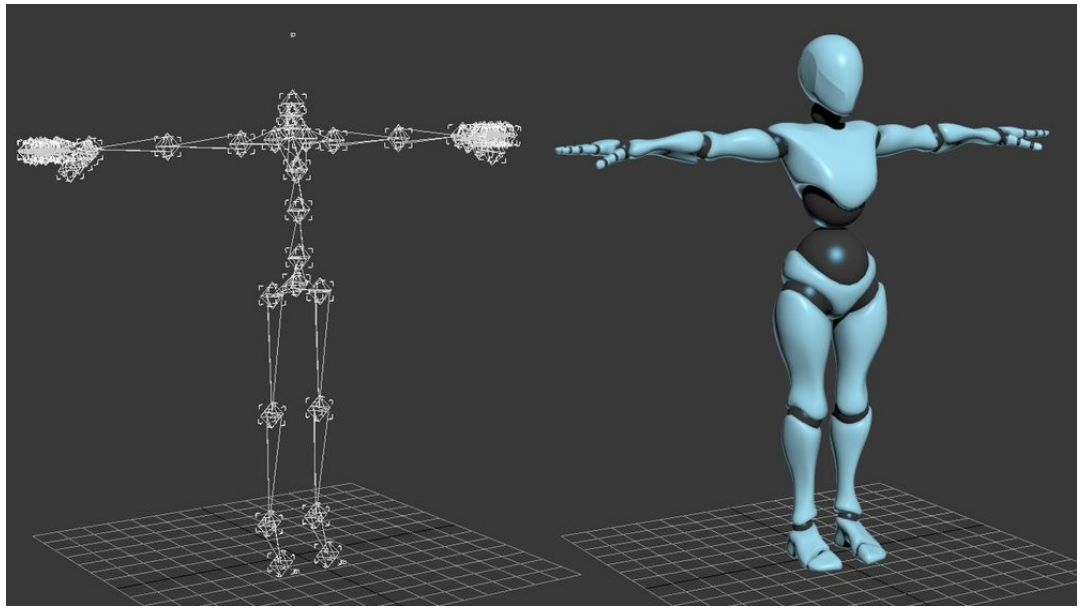
- layout of the 3D model in terms of cameras, light sources, and other nearby 3D models



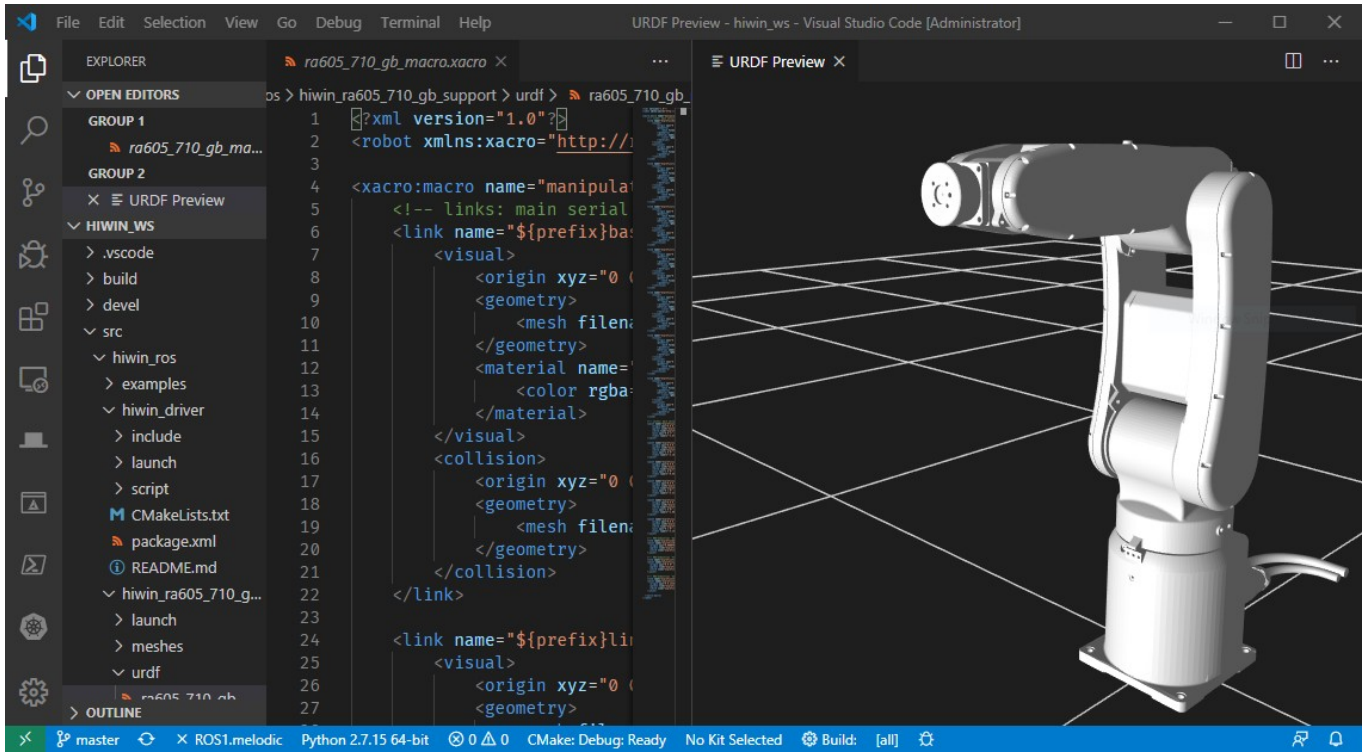
https://www.creativeshrimp.com/light-texture-tutorial-1-lighting-book.html

Animation

- skeletal animation: skeleton + joints



http://www.downloads.redway3d.com/downloads/public/documentation/wf_skeletal_animation.html



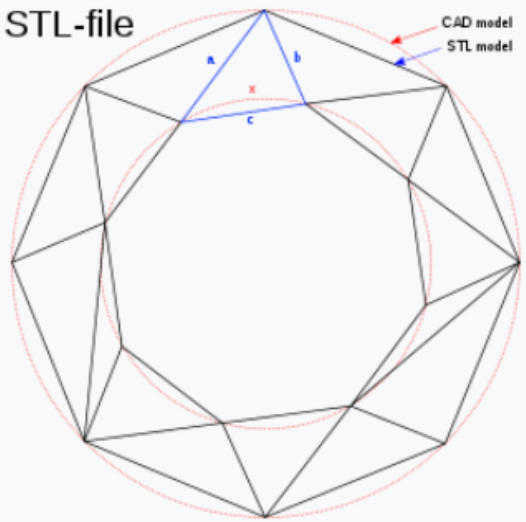
Unified Robot Description Format (URDF) in Visual Studio Code

https://microsoft.github.io/Win-RoS-Landing-Page/hiwin_case_study.html

Popular 3D file formats: **STL** (“stereolithography”)

STL

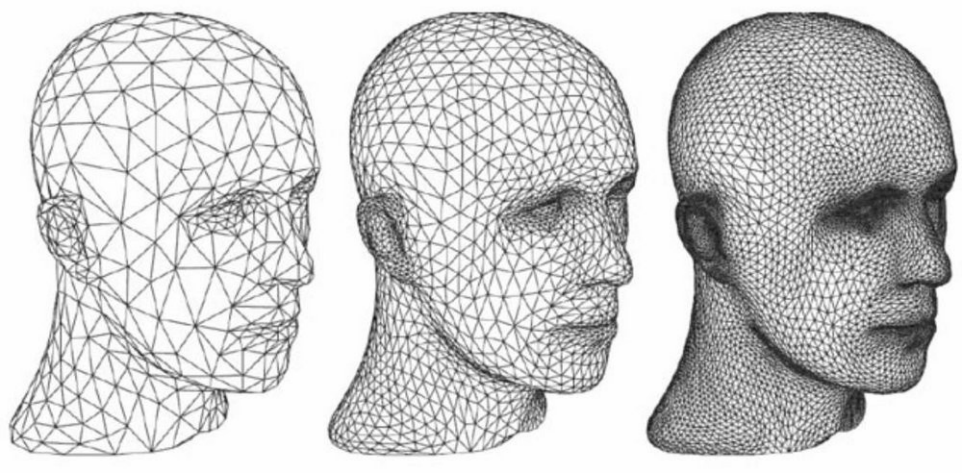
STL-file



A CAD representation of a torus (shown as two concentric red circles) and an STL approximation of the same shape (composed of triangular planes)

Filename extension	.stl
Internet media type	model/stl [1][2] model/x.stl-ascii model/x.stl-binary
Developed by	3D Systems
Initial release	1987
Type of format	Stereolithography

https://en.wikipedia.org/wiki/STL_(file_format)



FileInfo.com Example.STL

```
solid Default
  facet normal 0.000000e+00 0.000000e+00 1.000000e+00
    outer loop
      vertex -1.501741e+01 5.467951e+00 2.488822e+01
      vertex -1.501741e+01 3.843249e+00 2.488822e+01
      vertex -1.477975e+01 5.467951e+00 2.488822e+01
    endloop
  endfacet
  facet normal 0.000000e+00 0.000000e+00 1.000000e+00
    outer loop
      vertex -1.477975e+01 5.467951e+00 2.488822e+01
      vertex -1.501741e+01 3.843249e+00 2.488822e+01
      vertex -1.477975e+01 3.843249e+00 2.488822e+01
    endloop
  endfacet
  facet normal 0.000000e+00 0.000000e+00 1.000000e+00
    outer loop
      vertex -6.764211e+00 1.188298e-01 2.488822e+01
      vertex -6.764211e+00 1.782447e+00 2.488822e+01
      vertex -7.001871e+00 1.188298e-01 2.488822e+01
    endloop
  endfacet
  facet normal 0.000000e+00 0.000000e+00 1.000000e+00
    outer loop
      vertex -7.001871e+00 1.188298e-01 2.488822e+01
      vertex -6.764211e+00 1.782447e+00 2.488822e+01
      vertex -7.001871e+00 1.782447e+00 2.488822e+01
    endloop
  endfacet
  facet normal 0.000000e+00 0.000000e+00 1.000000e+00
```

- Popular for 3D printing
- Triangular mesh,
- No color information
- ASCII or Binary representation
- Superseded by OBJ, 3MF, AMF, …

Popular 3D file formats: Wavefront **OBJ** file format

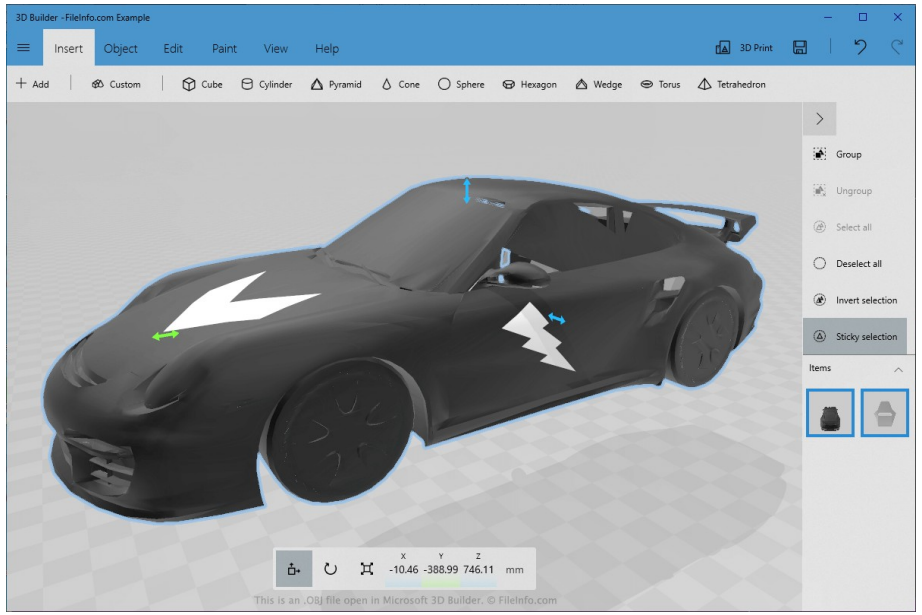
OBJ geometry format

Filename extension	.obj
Internet media type	model/obj [1]
Developed by	Wavefront Technologies
Type of format	3D model format

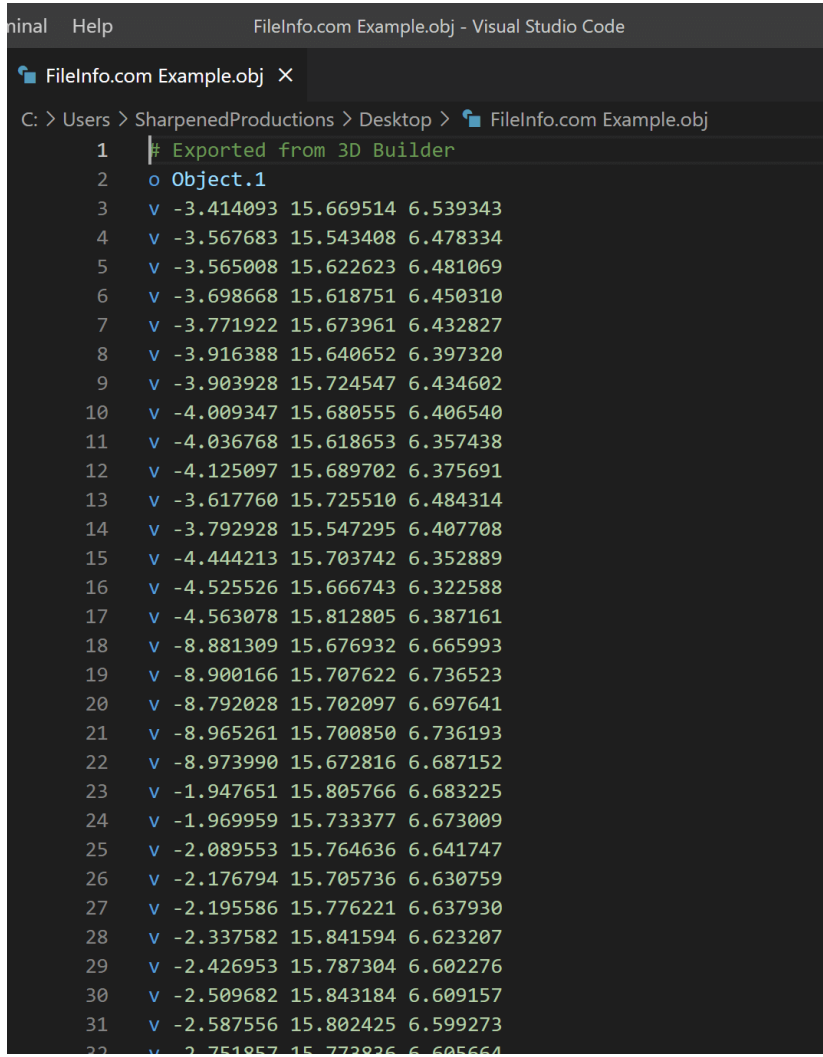
MTL material format

Filename extension	.mtl
Internet media type	model/mtl [3]
Magic number	ASCII: newmtl Hex: 6E65776D746C [4]
Developed by	Wavefront Technologies
Type of format	3D texture format

https://en.wikipedia.org/wiki/Wavefront_.obj_file



OBJ file open in Microsoft 3D Builder



- Popular for 3D printing and 3D graphics
- Triangular mesh
- + other kinds of interpolation: Tayolr, B-splines…
- No animation, no deformation
- Reference to companion file MTL (Material Template Library)

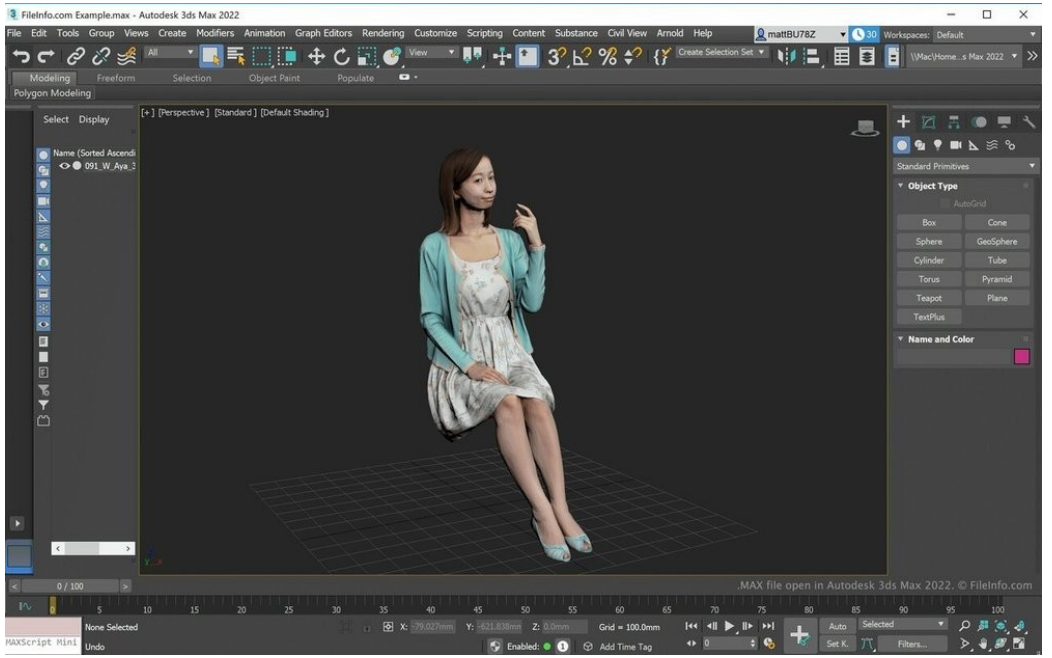
https://people.sc.fsu.edu/~jburkardt/data/mtl/mtl.html

Popular 3D file formats: Autodesk **3DS** Max file format

3DS Max File

Filename extension	.3ds
Internet media type	application/x-3ds, image/x-3ds
Magic number	4D 4D (hex), MM (ASCII)
Developed by	Autodesk Inc.
Type of format	3D file formats

https://en.wikipedia.org/wiki/.3ds



- Autodesk 3D Studio MAX 1996
- Popular for architecture, engineering, education, manufacturing
- geometry, appearance, scene, and animation
- Triangular mesh, total 65536 triangles
- Directional light sources are not supported

Reference: http://paulbourke.net/dataformats/3ds/
Tree structure made of chunks
In binary :

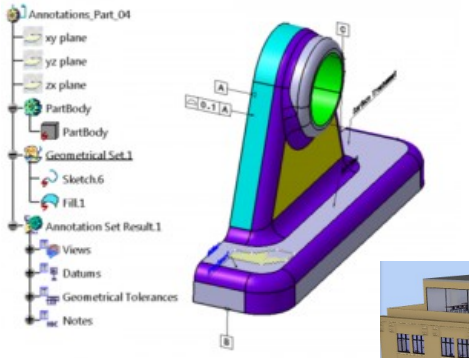
- 6 byte Chunk header:
  - 1-2 = chunk ID
  - 3-6 = little-endian length of the chunk
- Next bytes: chunk's data, including sub-chunks

Popular 3D file formats: ISO 10303-21 STEP-files

STEP

Filename extension	.stp, .step, .stpnc, .p21, .210
Internet media type	model/step, model/step+xml, model/step+zip, model/step-xml+zip
Magic number	ISO-10303-21
Developed by	ISO
Initial release	1994
Website	Specification

https://en.wikipedia.org/wiki/ISO_10303-21



```
ISO-10303-21;
HEADER;
FILE_DESCRIPTION(
/* description */ ('A minimal AP214 example with a single part'),
/* implementation_level */ '2;1');
FILE_NAME(
/* name */ 'demo',
/* time_stamp */ '2003-12-27T11:57:53',
/* author */ ('Lothar Klein'),
/* organization */ ('LKSoft'),
/* preprocessor_version */ ' ',
/* originating_system */ 'IDA-STEP',
/* authorization */ ' ');
FILE_SCHEMA (('AUTOMOTIVE_DESIGN { 1 0 10303 214 2 1 1}'));
ENDSEC;
DATA;
#10=ORGANIZATION('O0001','LKSoft','company');
#11=PRODUCT_DEFINITION_CONTEXT('part definition',#12,'manufacturing');
#12=APPLICATION_CONTEXT('mechanical design');
#13=APPLICATION_PROTOCOL_DEFINITION('','automotive_design',2003,#12);
#14=PRODUCT_DEFINITION('0',$,#15,#11);
#15=PRODUCT_DEFINITION_FORMATION('1',$,#16);
#16=PRODUCT('A0001','Test Part 1','',(#18));
#17=PRODUCT_RELATED_PRODUCT_CATEGORY('part',$,(#16));
#18=PRODUCT_CONTEXT('',#12,'');
#19=APPLIED_ORGANIZATION_ASSIGNMENT(#10,#20,(#16));
#20=ORGANIZATION_ROLE('id owner');
ENDSEC;
END-ISO-10303-21;
```

- Popular for in many engineering domains that rely on CAD
  - Mechanical engineering
  - AECOO industry (Architecture, Engineering, Construction, Owner Operator)
- ASCII, not storage-efficient
- Schema defined separately in the EXPRESS data modeling language ISO 10303-11

Data Interoperability and Semantics

</ Part 2. Data Formats >

ICM – Toolbox Engineering and Interoperability of Software Systems – Course unit on Data Interoperability and Semantics
M1 Cyber Physical and Social Systems – Course unit on Data Interoperability and Semantics
Maxime Lefrançois https://maxime-lefrancois.info
Course unit URL: https://ci.mines-stetienne.fr/cps2/course/data