Data Interoperability and Semantics

Syllabus and Course Organization

Course objectives

This lecture aims at ensuring you are familiar with data and information, at the core of every information technology (IT). The course focuses on four main topics:

1. Encoding base data types
2. Data formats
3. Data schemas and semantics
4. The value of data
5. The European strategy for data

You will practice with tutorials, and gain experience through a project

ICM – Computer Science Major – 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

Grading policy

Practical Work x2 (PW)
Written Exam x2 (WE)
Grade = (PW + WE)/2

ICM – Computer Science Major – 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

Syllabus and Course Organization

Course objectives

This lecture aims at ensuring you are familiar with data and information, at the core of every information technology (IT). The course focuses on four main topics:

1. Encoding base data types
2. Data formats
3. Data schemas and semantics
4. The value of data
5. The European strategy for data

You will practice with tutorials, and gain experience through a project

ICM – Computer Science Major – 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

Grading policy

Practical Work x2 (PW)
Written Exam x2 (WE)
Grade = (PW + WE)/2

ICM – Computer Science Major – 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

< Part 1. Encoding base data types >

Data Interoperability and Semantics
Outline

- < Part 1. Encoding base data types >
  - Part 1.1. Reminders: binary and hexadecimal strings
  - Part 1.2. Endianness
  - Example: MCF88 LoRa temperature, humidity and pressure sensor payload
  - Part 1.3. Computer number formats
  - Part 1.4. Character encoding
  - Part 1.5. Base32 and Base64 encoding
  - Part 1.6. Date and time
  - Part 1.7. XML Schema Datatypes
  - Part 1.8. Codes: countries, languages, …
  - Part 1.9. Quantities and Units of measure
  - Part 1.10. Colors

Data Interoperability and Semantics

Part 1. Encoding base data types

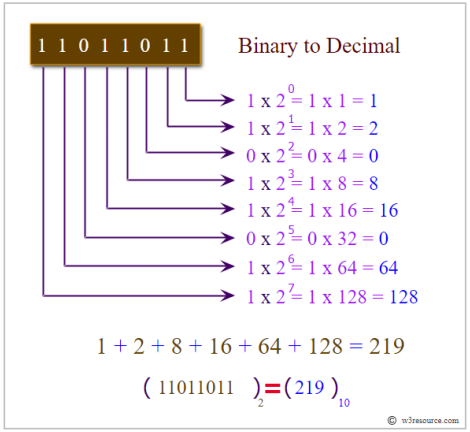
Part 1.1. Reminders: binary and hexadecimal strings

Numbering systems for computers

System	Base	Digits	ex python
Binary	**2**	**0,1**	0b"01111011"
Octal	**8**	**0,1,2,3,4,5,6,7**	0o"173"
Decimal	**10**	**0,1,2,3,4,5,6,7,8,9**	123
Hexadecimal	**16**	**0,1,2,3,4,5,6,7,8,9,A,B,C,D,E,F**	0x"7B"

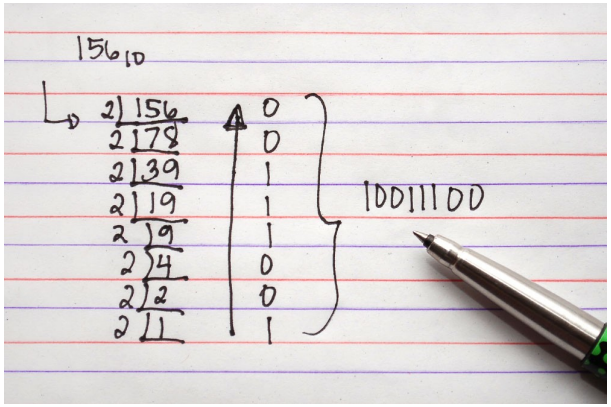
It has been debated a lot at the beginning
https://www.youtube.com/watch?v=thrx3SBEpL8

Tips: binary to decimal



https://www.w3resource.com/javascript-exercises/javascript-math-exercise-2.php

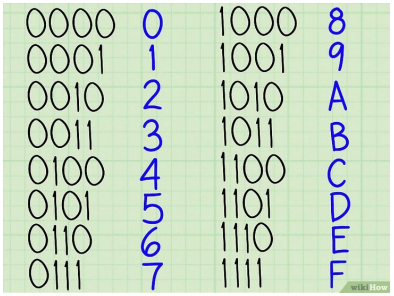
Tips: decimal to binary



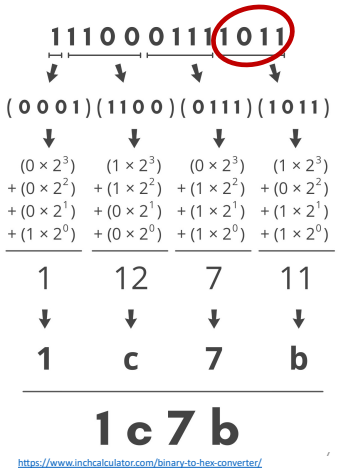
https://tex.stackexchange.com/questions/96399/how-can-i-illustrate-decimal-to-binary-conversion

Tips: binary to hexadecimal

Nibble - *In computing, a nibble is a four-bit aggregation, or half an octet. It is also known as half-byte or tetrade. In a networking or telecommunication context, the nibble is often called a semi-octet, quadbit, or quartet.*



Binary nibble to hexadecimal digit



https://www.inchcalculator.com/binary-to-hex-converter/

```c
#include <stdio.h>
int main()
{
  unsigned char a = 0x65;
  unsigned char b = 0x09;

  // a = 0x65(0110 0101), b = 0x09(0000 1001)
  printf("a = %#.2X, b = %#.2x\n", a, b);

  // & bitwise AND operator
  // the result is 0x01(0000 0001)
  printf("a&b = %#.2X\n", a&b);

  // | bitwise OR operator
  // the result is 0x6D(0110 1101)
  printf("a|b = %#.2X\n", a|b);

  // ^ bitwise exclusive OR operator
  // the result is 0x6C(0110 1100)
  printf("a^b = %#.2X\n", a^b);

  // << Left shift operator
  // the result is 0xCA(1100 1010)
  printf("a << 1 = %#.2X\n", a << 1);

  // >> Right shift operator
  // the result is 0x32(0011 0010)
  printf("a >> 1 = %#.2X\n", a >> 1);

  // ~ bitwise One's Complement operator
  // the result is 0X9A (1001 1010)
  // or more precisely 0XFFFFFF9A (~ promotes to int)
  printf("~a = %#.2X\n", ~a);

  // Get 3 bits starting at position 2 (start index 0)
  // the result is 1(0001)
  printf("(a >> 2) & 0x7 = %#.1X\n", (a>>2)&7);

  return 0;
}
```

Programming with binary strings

```
$ gcc main.c
$ ./a.out
a = 0X65, b = 0x09
a&b = 0X01
a|b = 0X6D
a^b = 0X6C
a << 1 = 0XCA
a >> 1 = 0X32
~a = 0XFFFFFF9A
(a >> 2) & 0x7 = 0X1
```

Data Interoperability and Semantics

Part 1. Encoding base data types

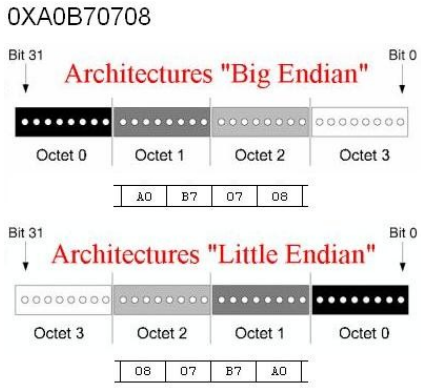
Part 1.2. Endianness

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

Endianness

*In computing, endianness is the order or sequence of bytes of a word of digital data in computer memory. Endianness is primarily expressed as **big-endian (BE)** or **little-endian (LE)**.*

Acronyms
- LSB - Least significant byte
- MSB - Most significant byte

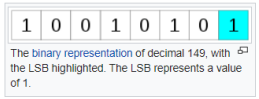


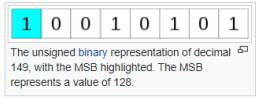
https://www.sqlpac.com/fr/documents/sybase-ase-12.5.3-dump-load-cross-platforms.html

Bit endianness or bit-level endianness

Bit endianness or bit-level endianness refers to the transmission order of bits over a serial medium

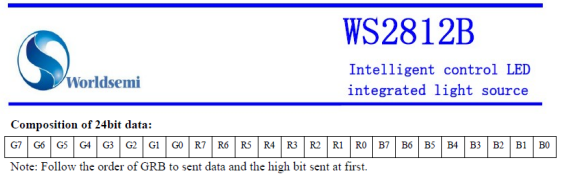
See course Programming Connected Devices:
- Least significant bit first: used in RS-232, Ethernet, USB…
- Most significant bit first: used in I²C





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

Example: WS2812B color leds



Data Interoperability and Semantics

Part 1. Encoding base data types

Example: MCF88 LoRa temperature, humidity and pressure sensor payload

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

Author: Colognato Stefano

Created: 29/09/2016
Modified: 30/11/2018
MCF88 DATA FRAME FORMAT 1.17

1.2 TEMPERATURE/PRESSURE/HUMIDITY

HOME

name	size [byte]	hex value	mean
Uplink ID	1 byte	04	Temperature/Pressure/Humidity
Data	10 byte	XX XX	Measure 1, refer to Note1
	10 byte	XX XX	Measure 2, refer to Note1
	10 byte	XX XX	Measure 3, refer to Note1
Batt %	1 byte (optional)	XX	Battery percentage
RFU	4 byte (optional)	XX XX XX XX	Optional RFU byte

Note1:

The 10 bytes for each measurement are divided as follows:

- 4 bytes are for the date and time. The MSB (most significant byte) is on the right so they must be read from the right. The 4 byte in reverse order are as follows:
  - 7 bit for the offset of the year, starting from the year 2000
  - 4 bit per month
  - 5 bit for day of the month
  - 5 bits for hour
  - 6 bits for minutes
  - 5 bits for half the seconds. The seconds range is from 0 to 31, so the result should be multiplied by 2 to find the actual seconds of the measurement.
- 2 bytes for temperature. The temperature is represented by a signed integer with the least significant byte first. The temperature is expressed in hundreds of a °C degree.
- 1 byte for humidity. Relative humidity is an unsigned integer corresponding to twice the percentage of humidity.
- 3 bytes for pressure. Pressure is an unsigned integer with the least significant byte first; it is expressed in Pascal.

Example

Sample payload:
04dc7e3721b40a47608801dd7e3721b10a43608801e07e3721b20a425d8801

Example: MCF88 LoRa sensors

Author: Colognato Stefano

Created: 29/09/2016
Modified: 30/11/2018
MCF88 DATA FRAME FORMAT 1.17

1.2 TEMPERATURE/PRESSURE/HUMIDITY

HOME

name	size [byte]	hex value	mean
Uplink ID	1 byte	04	Temperature/Pressure/Humidity
Data	10 byte	XX XX	Measure 1, refer to Note1
	10 byte	XX XX	Measure 2, refer to Note1
	10 byte	XX XX	Measure 3, refer to Note1
Batt %	1 byte (optional)	XX	Battery percentage
RFU	4 byte (optional)	XX XX XX XX	Optional RFU byte

Note1:

The 10 bytes for each measurement are divided as follows:

- 4 bytes are for the date and time. The MSB (most significant byte) is on the right so they must be read from the right. The 4 byte in reverse order are as follows:
  - 7 bit for the offset of the year, starting from the year 2000
  - 4 bit per month
  - 5 bit for day of the month
  - 5 bits for hour
  - 6 bits for minutes
  - 5 bits for half the seconds. The seconds range is from 0 to 31, so the result should be multiplied by 2 to find the actual seconds of the measurement.
- 2 bytes for temperature. The temperature is represented by a signed integer with the least significant byte first. The temperature is expressed in hundreds of a °C degree.
- 1 byte for humidity. Relative humidity is an unsigned integer corresponding to twice the percentage of humidity.
- 3 bytes for pressure. Pressure is an unsigned integer with the least significant byte first; it is expressed in Pascal.

Example

Sample payload:
04dc7e3721b40a47608801dd7e3721b10a43608801e07e3721b20a425d8801
Remove the first byte and divide the other 30 into 3 parts by 10 byte that correspond to 3 measurements.
The 3 measurements will be:
- *dc7e3721b40a47608801*
- *dd7e3721b10a43608801*
- *e07e3721b20a425d8801*

Author: Colognato Stefano

Created: 29/09/2016
Modified: 30/11/2018
MCF88 DATA FRAME FORMAT 1.17

1.2 TEMPERATURE/PRESSURE/HUMIDITY

HOME

name	size [byte]	hex value	mean
Uplink ID	1 byte	04	Temperature/Pressure/Humidity
Data	10 byte	XX XX	Measure 1, refer to Note1
	10 byte	XX XX	Measure 2, refer to Note1
	10 byte	XX XX	Measure 3, refer to Note1
Batt %	1 byte (optional)	XX	Battery percentage
RFU	4 byte (optional)	XX XX XX XX	Optional RFU byte

Note1:

The 10 bytes for each measurement are divided as follows:

- 4 bytes are for the date and time. The MSB (most significant byte) is on the right so they must be read from the right. The 4 byte in reverse order are as follows:
  - 7 bit for the offset of the year, starting from the year 2000
  - 4 bit per month
  - 5 bit for day of the month
  - 5 bits for hour
  - 6 bits for minutes
  - 5 bits for half the seconds. The seconds range is from 0 to 31, so the result should be multiplied by 2 to find the actual seconds of the measurement.
- 2 bytes for temperature. The temperature is represented by a signed integer with the least significant byte first. The temperature is expressed in hundreds of a °C degree.
- 1 byte for humidity. Relative humidity is an unsigned integer corresponding to twice the percentage of humidity.
- 3 bytes for pressure. Pressure is an unsigned integer with the least significant byte first; it is expressed in Pascal.

Example

Sample payload:
04dc7e3721b40a47608801dd7e3721b10a43608801e07e3721b20a425d8801
Remove the first byte and divide the other 30 into 3 parts by 10 byte that correspond to 3 measurements.
The 3 measurements will be:
- *dc7e3721b40a47608801*
- *dd7e3721b10a43608801*
- *e07e3721b20a425d8801*
Decipher the first measurement dividing it by groups and applying the necessary transformations:
- *Measurement date: dc 7e 37 21*
  - *Byte swapping, result: 21 37 7e dc*

mcf88

Author: Colognato Stefano

Created: 29/09/2016
Modified: 30/11/2018
MCF88 DATA FRAME FORMAT 1.17

1.2 TEMPERATURE/PRESSURE/HUMIDITY

HOME

name	size [byte]	hex value	mean
Uplink ID	1 byte	04	Temperature/Pressure/Humidity
Data	10 byte	XX XX	Measure 1, refer to Note1
	10 byte	XX XX	Measure 2, refer to Note1
	10 byte	XX XX	Measure 3, refer to Note1
Batt %	1 byte (optional)	XX	Battery percentage
RFU	4 byte (optional)	XX XX XX XX	Optional RFU byte

Note1:

The 10 bytes for each measurement are divided as follows:

- 4 bytes are for the date and time. The MSB (most significant byte) is on the right so they must be read from the right. The 4 byte in reverse order are as follows:
  - 7 bit for the offset of the year, starting from the year 2000
  - 4 bit per month
  - 5 bit for day of the month
  - 5 bits for hour
  - 6 bits for minutes
  - 5 bits for half the seconds. The seconds range is from 0 to 31, so the result should be multiplied by 2 to find the actual seconds of the measurement.
- 2 bytes for temperature. The temperature is represented by a signed integer with the least significant byte first. The temperature is expressed in hundreds of a °C degree.
- 1 byte for humidity. Relative humidity is an unsigned integer corresponding to twice the percentage of humidity.
- 3 bytes for pressure. Pressure is an unsigned integer with the least significant byte first; it is expressed in Pascal.

Example

Sample payload:
04dc7e3721b40a47608801dd7e3721b10a43608801e07e3721b20a425d8801
Remove the first byte and divide the other 30 into 3 parts by 10 byte that correspond to 3 measurements.
The 3 measurements will be:

- *dc7e3721b40a47608801*
- *dd7e3721b10a43608801*
- *e07e3721b20a425d8801*

Decipher the first measurement dividing it by groups and applying the necessary transformations:

- *Measurement date: dc 7e 37 21*
  - *Byte swapping, result: 21 37 7e dc*
  - *The result in bits will be: 00100001 00110111 01111110 11011100*
    - *The bits are divided as explained above*
      - *Year: 0010000*
        - *Result: 16*
          - *2000+16 = 2016*
      - *Month: 1001*
        - *Result: 9*
      - *Day: 10111*
        - *Result: 23*
      - *Hour: 01111*
        - *Result: 15*
      - *Minutes: 110110*
        - *Result: 54*
      - *Seconds: 11100*
        - *Result: 28*
          - *28*2 = 56*
  - *The date of the measurament will be: 23/09/2016 15:54:56.*

mcf88

Author: Colognato Stefano

Created: 29/09/2016
Modified: 30/11/2018
MCF88 DATA FRAME FORMAT 1.17

1.2 TEMPERATURE/PRESSURE/HUMIDITY

HOME

name	size [byte]	hex value	mean
Uplink ID	1 byte	04	Temperature/Pressure/Humidity
Data	10 byte	XX XX	Measure 1, refer to Note1
	10 byte	XX XX	Measure 2, refer to Note1
	10 byte	XX XX	Measure 3, refer to Note1
Batt %	1 byte (optional)	XX	Battery percentage
RFU	4 byte (optional)	XX XX XX XX	Optional RFU byte

Note1:

The 10 bytes for each measurement are divided as follows:

- 4 bytes are for the date and time. The MSB (most significant byte) is on the right so they must be read from the right. The 4 byte in reverse order are as follows:
  - 7 bit for the offset of the year, starting from the year 2000
  - 4 bit per month
  - 5 bit for day of the month
  - 5 bits for hour
  - 6 bits for minutes
  - 5 bits for half the seconds. The seconds range is from 0 to 31, so the result should be multiplied by 2 to find the actual seconds of the measurement.
- 2 bytes for temperature. The temperature is represented by a signed integer with the least significant byte first. The temperature is expressed in hundreds of a °C degree.
- 1 byte for humidity. Relative humidity is an unsigned integer corresponding to twice the percentage of humidity.
- 3 bytes for pressure. Pressure is an unsigned integer with the least significant byte first; it is expressed in Pascal.

Example

Sample payload:
04dc7e3721b40a47608801dd7e3721b10a43608801e07e3721b20a425d8801
Remove the first byte and divide the other 30 into 3 parts by 10 byte that correspond to 3 measurements.
The 3 measurements will be:

- *dc7e3721b40a47608801*
- *dd7e3721b10a43608801*
- *e07e3721b20a425d8801*

Decipher the first measurement dividing it by groups and applying the necessary transformations:

- *Measurement date: dc 7e 37 21*
  - *Byte swapping, result: 21 37 7e dc*
  - *The result in bits will be: 00100001 00110111 01111110 11011100*
    - *The bits are divided as explained above*
      - *Year: 0010000*
        - *Result: 16*
          - *2000+16 = 2016*
      - *Month: 1001*
        - *Result: 9*
      - *Day: 10111*
        - *Result: 23*
      - *Hour: 01111*
        - *Result: 15*
      - *Minutes: 110110*
        - *Result: 54*
      - *Seconds: 11100*
        - *Result: 28*
          - *28*2 = 56*
  - *The date of the measurament will be: 23/09/2016 15:54:56.*

mcf88

Author: Colognato Stefano

Created: 29/09/2016
Modified: 30/11/2018
MCF88 DATA FRAME FORMAT 1.17

1.2 TEMPERATURE/PRESSURE/HUMIDITY

HOME

name	size [byte]	hex value	mean
Uplink ID	1 byte	04	Temperature/Pressure/Humidity
Data	10 byte	XX XX	Measure 1, refer to Note1
	10 byte	XX XX	Measure 2, refer to Note1
	10 byte	XX XX	Measure 3, refer to Note1
Batt %	1 byte (optional)	XX	Battery percentage
RFU	4 byte (optional)	XX XX XX XX	Optional RFU byte

Note1:

The 10 bytes for each measurement are divided as follows:

- 4 bytes are for the date and time. The MSB (most significant byte) is on the right so they must be read from the right. The 4 byte in reverse order are as follows:
  - 7 bit for the offset of the year, starting from the year 2000
  - 4 bit per month
  - 5 bit for day of the month
  - 5 bits for hour
  - 6 bits for minutes
  - 5 bits for half the seconds. The seconds range is from 0 to 31, so the result should be multiplied by 2 to find the actual seconds of the measurement.
- 2 bytes for temperature. The temperature is represented by a signed integer with the least significant byte first. The temperature is expressed in hundreds of a °C degree.
- 1 byte for humidity. Relative humidity is an unsigned integer corresponding to twice the percentage of humidity.
- 3 bytes for pressure. Pressure is an unsigned integer with the least significant byte first; it is expressed in Pascal.

Example

Sample payload:
04dc7e3721b40a47608801dd7e3721b10a43608801e07e3721b20a425d8801
Remove the first byte and divide the other 30 into 3 parts by 10 byte that correspond to 3 measurements.
The 3 measurements will be:

- *dc7e3721b40a47608801*
- *dd7e3721b10a43608801*
- *e07e3721b20a425d8801*

Decipher the first measurement dividing it by groups and applying the necessary transformations:

- *Measurement date: dc 7e 37 21*
  - *Byte swapping, result: 21 37 7e dc*
  - *The result in bits will be: 00100001 00110111 01111110 11011100*
    - *The bits are divided as explained above*
      - *Year: 0010000*
        - *Result: 16*
          - *2000+16 = 2016*
      - *Month: 1001*
        - *Result: 9*
      - *Day: 10111*
        - *Result: 23*
      - *Hour: 01111*
        - *Result: 15*
      - *Minutes: 110110*
        - *Result: 54*
      - *Seconds: 11100*
        - *Result: 28*
          - *28*2 = 56*
  - *The date of the measurament will be: 23/09/2016 15:54:56.*
- *Temperature: b40a*
  - *Byte swapping, result: 0ab4*
    - *The result (with sign) will be +2740 with two decimal places, then + 27.40 °C.*
- *Humidity: 47*
  - *In decimal is 71, the humidity is 71/2 = 35.5% rH.*
- *Pressure: 608801*
  - *Byte swapping, result: 018860*
    - *In decimal, the result is 100448, with two decimal places the pressure is 1004.48 hPa.*

Data Interoperability and Semantics

Part 1. Encoding base data types

Part 1.3. Computer number 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

C number data types

- char (8 bits) – [0, 255]
- short/int (16 bits) - [-32,767, +32,767] or [0, 65,535]
- long (32 bits) - [–2,147,483,647, +2,147,483,647] or [0, 4,294,967,295]
- long long (64 bits) [–9,223,372,036,854,775,807, +9,223,372,036,854,775,807] or <<positive>>
- float - IEEE 754 single-precision binary floating-point format (32 bits)
- double - IEEE 754 double-precision binary floating-point format (64 bits)

Integer encoding

Unsigned

$$B2U(X) = \sum_{i=0}^{w-1} x_i \cdot 2^i$$

Two's Complement

$$B2T(X) = -x_{w-1} \cdot 2^{w-1} + \sum_{i=0}^{w-2} x_i \cdot 2^i$$

```
short int x =  15213;
short int y = -15213;
```

Sign Bit

- C `short` 2 bytes long

	Decimal	Hex	Binary
x	15213	3B 6D	00111011 01101101
y	-15213	C4 93	11000100 10010011

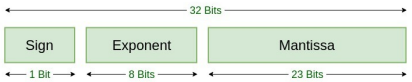
Sign Bit

- For 2's complement, most significant bit indicates sign
  - 0 for nonnegative
  - 1 for negative

https://slideplayer.com/slide/5048952/

IEEE 754 single/double encoding

https://www.geeksforgeeks.org/ieee-standard-754-floating-point-numbers/



Single Precision
IEEE 754 Floating-Point Standard

$$-1^s \times 2^{(\exp-127)} \times 1.frac$$

exceptional cases:

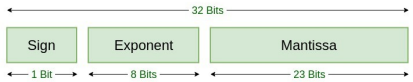
- If $E = 255$ and F is nonzero, then $x = NaN$ ("Not a number").
- If $E = 255$, F is zero, and S is 1, then $x = -$ Infinity.
- If $E = 255$, F is zero, and S is 0, then $x = +$ Infinity.
- If $0 < E < 255$, then $x = (-1)^s \times (1.F) \times 2^{E-127}$, where 1.$F$ represents the binary number created by prefixing F with an implicit leading 1 and a binary point.
- If $E = 0$ and F is nonzero, then $x = (-1)^s \times (0.F) \times 2^{-126}$. This is an "unnormalized" value.
- If $E = 0$, F is zero, and S is 1, then $x = -0$.
- If $E = 0$, F is zero, and S is 0, then $x = 0$.

Simple examples of conversions

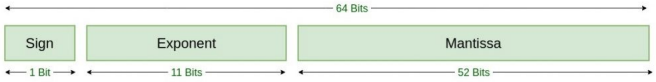
$$0\ 10000000\ 00000000000000000000000 = (-1)^0 \times (1.0_2) \times 2^{128-127} = 2.0$$
$$0\ 10000001\ 10100000000000000000000 = (-1)^0 \times (1.101_2) \times 2^{129-127} = 6.5$$
$$1\ 10000001\ 10100000000000000000000 = (-1)^1 \times (1.101_2) \times 2^{129-127} = -6.5.$$

IEEE 754 single/double encoding

https://www.geeksforgeeks.org/ieee-standard-754-floating-point-numbers/



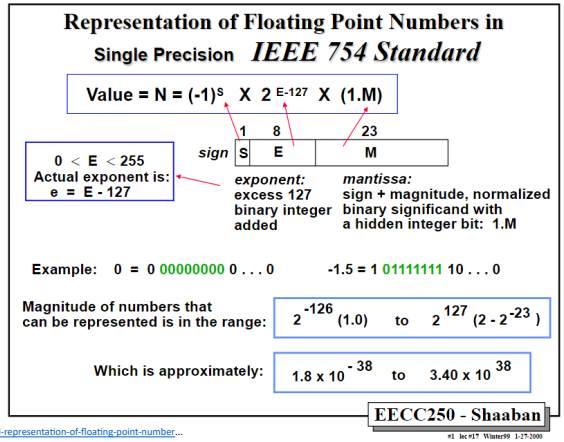
Single Precision
IEEE 754 Floating-Point Standard



Double Precision
IEEE 754 Floating-Point Standard

https://www.geeksforgeeks.org/ieee-standard-754-floating-point-numbers/

IEEE 754 single/double/quadruple encoding



https://studylib.net/doc/17671880/ieee-754-standard-representation-of-floating-point-number...

Data Interoperability and Semantics

Part 1. Encoding base data types

Part 1.4. Character encoding

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

Character encoding

In computing, data storage, and data transmission, character encoding is used to represent a repertoire of characters by some kind of encoding system that assigns a number to each character for digital representation

— *https://en.wikipedia.org/wiki/Character_encoding*

Common character encodings [edit]

- ISO 646
  - ASCII
- EBCDIC
- ISO 8859
  - ISO 8859-1 Western Europe
  - ISO 8859-2 Western and Central Europe
  - ISO 8859-3 Western Europe and South European (Turkish, Maltese plus Esperanto)
  - ISO 8859-4 Western Europe and Baltic countries (Lithuania, Estonia, Latvia and Lapp)
  - ISO 8859-5 Cyrillic alphabet
  - ISO 8859-6 Arabic
  - ISO 8859-7 Greek
  - ISO 8859-8 Hebrew
  - ISO 8859-9 Western Europe with amended Turkish character set
  - ISO 8859-10 Western Europe with rationalised character set for Nordic languages, including complete Icelandic set
  - ISO 8859-11 Thai
  - ISO 8859-13 Baltic languages plus Polish
  - ISO 8859-14 Celtic languages (Irish Gaelic, Scottish, Welsh)
  - ISO 8859-15 Added the Euro sign and other rationalisations to ISO 8859-1
  - ISO 8859-16 Central, Eastern and Southern European languages (Albanian, Bosnian, Croatian, Hungarian, Polish, Romanian, Serbian and Slovenian, but also French, German, Italian and Irish Gaelic)
- CP437, CP720, CP737, CP850, CP852, CP855, CP857, CP858, CP860, CP861, CP862, CP863, CP865, CP866, CP869, CP872
- MS-Windows character sets:
  - Windows-1250 for Central European languages that use Latin script, (Polish, Czech, Slovak, Hungarian, Slovene, Serbian, Croatian, Bosnian, Romanian and Albanian)
  - Windows-1251 for Cyrillic alphabets
  - Windows-1252 for Western languages
  - Windows-1253 for Greek
  - Windows-1254 for Turkish
  - Windows-1255 for Hebrew
  - Windows-1256 for Arabic
  - Windows-1257 for Baltic languages
  - Windows-1258 for Vietnamese
- Mac OS Roman
- KOI8-R, KOI8-U, KOI7
- MIK
- ISCII
- TSCII
- VISCII
- JIS X 0208 is a widely deployed standard for Japanese character encoding that has several encoding forms.
  - Shift JIS (Microsoft Code page 932 is a dialect of Shift_JIS)
  - EUC-JP
  - ISO-2022-JP
- JIS X 0213 is an extended version of JIS X 0208.
  - Shift_JIS-2004
  - EUC-JIS-2004
  - ISO-2022-JP-2004
- Chinese Guobiao
  - GB 2312
  - GBK (Microsoft Code page 936)
  - GB 18030
- Taiwan Big5 (a more famous variant is Microsoft Code page 950)
  - Hong Kong HKSCS
- Korean
  - KS X 1001 is a Korean double-byte character encoding standard
  - EUC-KR
  - ISO-2022-KR
- Unicode (and subsets thereof, such as the 16-bit 'Basic Multilingual Plane')
  - UTF-8
  - UTF-16
  - UTF-32
- ANSEL or ISO/IEC 6937

Let's focus on the main standards

- ASCII
- UTF-8

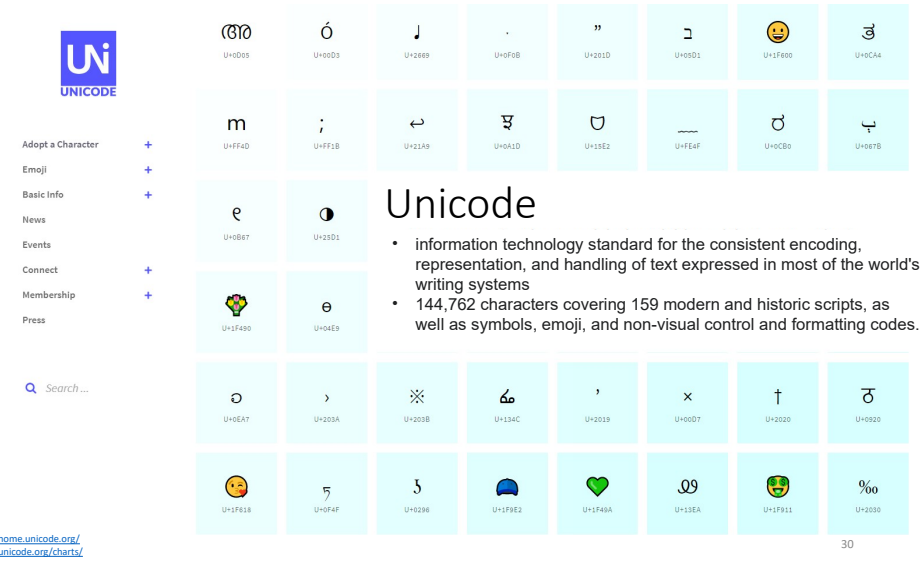
Common character encodings [edit]

- ISO 646
  - ASCII
- EBCDIC
- ISO 8859
  - ISO 8859-1 Western Europe
  - ISO 8859-2 Western and Central Europe
  - ISO 8859-3 Western Europe and South European (Turkish, Maltese plus Esperanto)
  - ISO 8859-4 Western Europe and Baltic countries (Lithuania, Estonia, Latvia and Lapp)
  - ISO 8859-5 Cyrillic alphabet
  - ISO 8859-6 Arabic
  - ISO 8859-7 Greek
  - ISO 8859-8 Hebrew
  - ISO 8859-9 Western Europe with amended Turkish character set
  - ISO 8859-10 Western Europe with rationalised character set for Nordic languages, including complete Icelandic set
  - ISO 8859-11 Thai
  - ISO 8859-13 Baltic languages plus Polish
  - ISO 8859-14 Celtic languages (Irish Gaelic, Scottish, Welsh)
  - ISO 8859-15 Added the Euro sign and other rationalisations to ISO 8859-1
  - ISO 8859-16 Central, Eastern and Southern European languages (Albanian, Bosnian, Croatian, Hungarian, Polish, Romanian, Serbian and Slovenian, but also French, German, Italian and Irish Gaelic)
- CP437, CP720, CP737, CP850, CP852, CP855, CP857, CP858, CP860, CP861, CP862, CP863, CP865, CP866, CP869, CP872
- MS-Windows character sets:
  - Windows-1250 for Central European languages that use Latin script, (Polish, Czech, Slovak, Hungarian, Slovene, Serbian, Croatian, Bosnian, Romanian and Albanian)
  - Windows-1251 for Cyrillic alphabets
  - Windows-1252 for Western languages
  - Windows-1253 for Greek
  - Windows-1254 for Turkish
  - Windows-1255 for Hebrew
  - Windows-1256 for Arabic
  - Windows-1257 for Baltic languages
  - Windows-1258 for Vietnamese
- Mac OS Roman
- KOI8-R, KOI8-U, KOI7
- MIK
- ISCII
- TSCII
- VISCII
- JIS X 0208 is a widely deployed standard for Japanese character encoding that has several encoding forms.
  - Shift JIS (Microsoft Code page 932 is a dialect of Shift_JIS)
  - EUC-JP
  - ISO-2022-JP
- JIS X 0213 is an extended version of JIS X 0208.
  - Shift_JIS-2004
  - EUC-JIS-2004
  - ISO-2022-JP-2004
- Chinese Guobiao
  - GB 2312
  - GBK (Microsoft Code page 936)
  - GB 18030
- Taiwan Big5 (a more famous variant is Microsoft Code page 950)
  - Hong Kong HKSCS
- Korean
  - KS X 1001 is a Korean double-byte character encoding standard
  - EUC-KR
  - ISO-2022-KR
- Unicode (and subsets thereof, such as the 16-bit 'Basic Multilingual Plane')
  - UTF-8
  - UTF-16
  - UTF-32
- ANSEL or ISO/IEC 6937

You will be given this document in appendix of the written exam !

ASCII (7 bits)

Decimal	Hex	Char	Decimal	Hex	Char	Decimal	Hex	Char	Decimal	Hex	Char
0	0	[NULL]	32	20	[SPACE]	64	40	@	96	60	`
1	1	[START OF HEADING]	33	21	!	65	41	A	97	61	a
2	2	[START OF TEXT]	34	22	"	66	42	B	98	62	b
3	3	[END OF TEXT]	35	23	#	67	43	C	99	63	c
4	4	[END OF TRANSMISSION]	36	24	$	68	44	D	100	64	d
5	5	[ENQUIRY]	37	25	%	69	45	E	101	65	e
6	6	[ACKNOWLEDGE]	38	26	&	70	46	F	102	66	f
7	7	[BELL]	39	27	'	71	47	G	103	67	g
8	8	[BACKSPACE]	40	28	(	72	48	H	104	68	h
9	9	[HORIZONTAL TAB]	41	29	)	73	49	I	105	69	i
10	A	[LINE FEED]	42	2A	*	74	4A	J	106	6A	j
11	B	[VERTICAL TAB]	43	2B	+	75	4B	K	107	6B	k
12	C	[FORM FEED]	44	2C	,	76	4C	L	108	6C	l
13	D	[CARRIAGE RETURN]	45	2D	-	77	4D	M	109	6D	m
14	E	[SHIFT OUT]	46	2E	.	78	4E	N	110	6E	n
15	F	[SHIFT IN]	47	2F	/	79	4F	O	111	6F	o
16	10	[DATA LINK ESCAPE]	48	30	0	80	50	P	112	70	p
17	11	[DEVICE CONTROL 1]	49	31	1	81	51	Q	113	71	q
18	12	[DEVICE CONTROL 2]	50	32	2	82	52	R	114	72	r
19	13	[DEVICE CONTROL 3]	51	33	3	83	53	S	115	73	s
20	14	[DEVICE CONTROL 4]	52	34	4	84	54	T	116	74	t
21	15	[NEGATIVE ACKNOWLEDGE]	53	35	5	85	55	U	117	75	u
22	16	[SYNCHRONOUS IDLE]	54	36	6	86	56	V	118	76	v
23	17	[ENG OF TRANS. BLOCK]	55	37	7	87	57	W	119	77	w
24	18	[CANCEL]	56	38	8	88	58	X	120	78	x
25	19	[END OF MEDIUM]	57	39	9	89	59	Y	121	79	y
26	1A	[SUBSTITUTE]	58	3A	:	90	5A	Z	122	7A	z
27	1B	[ESCAPE]	59	3B	;	91	5B	[	123	7B	{
28	1C	[FILE SEPARATOR]	60	3C	<	92	5C	\	124	7C	|
29	1D	[GROUP SEPARATOR]	61	3D	=	93	5D	]	125	7D	}
30	1E	[RECORD SEPARATOR]	62	3E	>	94	5E	^	126	7E	~
31	1F	[UNIT SEPARATOR]	63	3F	?	95	5F	_	127	7F	[DEL]



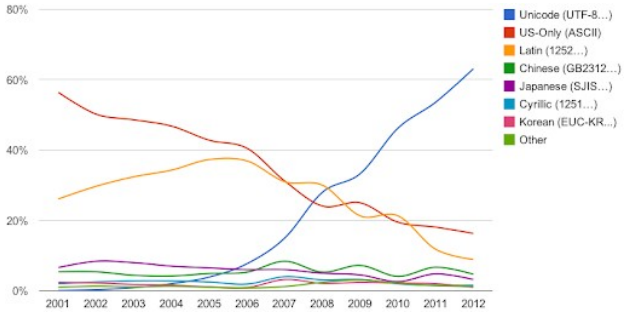
Unicode

- information technology standard for the consistent encoding, representation, and handling of text expressed in most of the world's writing systems
- 144,762 characters covering 159 modern and historic scripts, as well as symbols, emoji, and non-visual control and formatting codes.

https://home.unicode.org/
https://unicode.org/charts/

UTF-8

- Implementation of Unicode on 1-4 bytes (as little as needed)



share of web pages with different encodings

Inserting characters

Inserting Characters

There are many ways character é ("e" with an acute accent, character code 233 (decimal) in Latin-1 and Unicode), can be inserted into a document:

- On Windows, I hold down the Alt key and type 0233 on the numeric keyboard and release the Alt key. I could use the charmap program, too. Or I could copy and paste it (e.g., é). But entering the code directly is risky because, if the character encoding changes, e.g., from Latin-1 to UTF-8, then the meaning of code 233 changes.
- In an HTML document, I can enter these magical incantations, which are displayed correctly regardless of encoding:
  - é (decimal) ⇒ é
  - é (hex) ⇒ é
  - é (mnemonic) ⇒ é

  Note: HTML/XHTML validation programs might not be acquainted with these and complain.
- In Microsoft Word, I type an accent code followed by the accented letter. On Windows, Ctrl+quote, then 'e'. On Mac, Option+quote, then 'e'. Accent codes include: grave=backquote, acute=quote, circumflex=hat, colon=umlaut, comma=cedilla, tilde=tilde, slash=slash, and perhaps others.

http://hcibib.org/multilingual/badchars.htm

Encoding errors

What Could Possibly Go Wrong?

If é is UTF-8 encoded, but displayed without decoding, it looks like this:

Ã©

The first 128 characters in the Latin-1 character set (same as ASCII), are simply represented as themselves in UTF-8. The second half of Latin-1 characters are split. The first half of the non-ASCII Latin-1 characters are represented by themselves, preceded by code 194 decimal or C2 hex, so the UTF-8 encoding for character code 191 (decimal), ¿, is

Â¿

The second half of the non-ASCII Latin-1 characters are represented by a different character, preceded by code 195 decimal or C3 hex. So, when looking at UTF-8 encodings of Latin-1 characters, if you see Â or Ã where you do not expect it, there are probably too many UTF-8 encodings. Multiple extra encodings have a pattern to them:

0 é
1 Ã©
2 ÃƒÂ©
3 ÃƒÃ‚Â©
4 ÃƒÆ'Ã‚Æ'ÃƒÂ¢â€šÃ‚Â©
5 *you get the idea*

Note: If you see boxes in the characters above, it is because the font used is missing that character. There is no way to fix it other than getting a new font or by changing the font. Often, the fonts used in a window title or status bar or JavaScript are more limited than those used elsewhere, so the "alert", "title", and "status" buttons in the Character Conversion Corner can be used to test characters in those contexts.

Too few encodings can have a bad effect that looks different. When é is not UTF-8 encoded, it can appear like this very high numbered character:

Progressive under-encoding can result in a question mark being displayed.

http://hcibib.org/multilingual/badchars.htm

Encoding errors

Diagnostic Reference

You are now ready to diagnose UTF-8 encoding problems (e.g., with é):

Symptom	Diagnosis
é	no problems
Ã©	too much UTF-8 encoding, or viewing UTF-8 encoded text with Latin-1 encoding
ÃƒÂ©	much too much UTF-8 encoding
	too little UTF-8 encoding
?	something bad happened to this character
	wild animals have eaten this character
	if you see a box, the font in use is missing this character. Firefox 3's boxes contain the hexadecimal value for the missing character, but it's still just a missing character.

http://hcibib.org/multilingual/badchars.htm

Data Interoperability and Semantics

Part 1. Encoding base data types

Part 1.5. Base32 and Base64 encoding

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

Binary to text encoding

- Base64

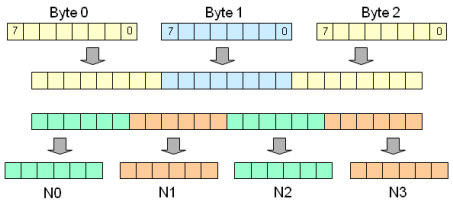


Table 1: The Base 64 Alphabet

Value	Encoding	Value	Encoding	Value	Encoding	Value	Encoding
0	A	17	R	34	i	51	z
1	B	18	S	35	j	52	0
2	C	19	T	36	k	53	1
3	D	20	U	37	l	54	2
4	E	21	V	38	m	55	3
5	F	22	W	39	n	56	4
6	G	23	X	40	o	57	5
7	H	24	Y	41	p	58	6
8	I	25	Z	42	q	59	7
9	J	26	a	43	r	60	8
10	K	27	b	44	s	61	9
11	L	28	c	45	t	62	+
12	M	29	d	46	u	63	/
13	N	30	e	47	v		
14	O	31	f	48	w	(pad)	=
15	P	32	g	49	x		
16	Q	33	h	50	y		

Table 2: The "URL and Filename safe" Base 64 Alphabet

62 - (minus)
63 _
(underline)

https://datatracker.ietf.org/doc/html/rfc4648
https://www.base64decode.org/

Binary to text encoding

- Base64

The following example of Base64 data is from [5], with corrections.

```
   Input data:  0x14fb9c03d97e
   Hex:     1   4    f   b    9   c     | 0   3    d   9    7   e
   8-bit:   00010100 11111011 10011100  | 00000011 11011001 01111110
   6-bit:   000101 001111 101110 011100 | 000000 111101 100101 111110
   Decimal: 5      15     46     28     0      61     37     62
   Output:  F      P      u      c      A      9      l      +

   Input data:  0x14fb9c03d9
   Hex:     1   4    f   b    9   c     | 0   3    d   9
   8-bit:   00010100 11111011 10011100  | 00000011 11011001
                                          pad with 00
   6-bit:   000101 001111 101110 011100 | 000000 111101 100100
   Decimal: 5      15     46     28     0      61     36
                                        pad with =
   Output:  F      P      u      c      A      9      k      =

   Input data:  0x14fb9c03
   Hex:     1   4    f   b    9   c     | 0   3
   8-bit:   00010100 11111011 10011100  | 00000011
                                        pad with 0000
   6-bit:   000101 001111 101110 011100 | 000000 110000
   Decimal: 5      15     46     28     0      48
                                      pad with =      =
   Output:  F      P      u      c      A      w      =      =
```

https://datatracker.ietf.org/doc/html/rfc4648
https://www.base64decode.org/

Binary to text encoding

- Base64
- Base32
- Base16

Table 3: The Base 32 Alphabet

Value	Encoding	Value	Encoding	Value	Encoding	Value	Encoding
0	A	9	J	18	S	27	3
1	B	10	K	19	T	28	4
2	C	11	L	20	U	29	5
3	D	12	M	21	V	30	6
4	E	13	N	22	W	31	7
5	F	14	O	23	X		
6	G	15	P	24	Y	(pad)	=
7	H	16	Q	25	Z		
8	I	17	R	26	2		

Table 5: The Base 16 Alphabet

Value	Encoding	Value	Encoding	Value	Encoding	Value	Encoding
0	0	4	4	8	8	12	C
1	1	5	5	9	9	13	D
2	2	6	6	10	A	14	E
3	3	7	7	11	B	15	F

https://datatracker.ietf.org/doc/html/rfc4648
https://www.base64decode.org/

Data Interoperability and Semantics

Part 1. Encoding base data types

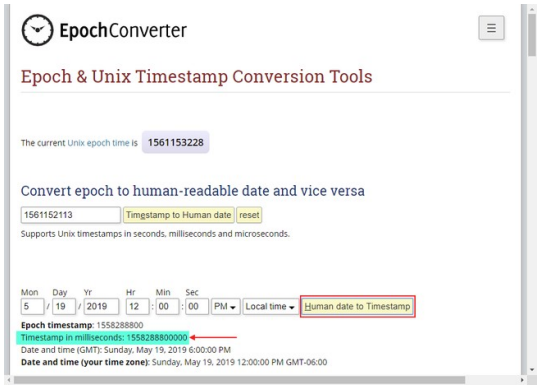
Part 1.6. Date and time

Representing Date and Time is not simple

- Leap year / bissextile years
  - *a calendar year that contains an additional day added to keep the calendar year synchronized with the astronomical year or seasonal year.*
- Leap seconds
  - *a one-second adjustment that is occasionally applied to Coordinated Universal Time (UTC), to accommodate the difference between precise time (International Atomic Time (TAI), as measured by atomic clocks) and imprecise observed solar time (UT1), which varies due to irregularities and long-term slowdown in the Earth's rotation*
- Timezones
  - UTC, GMT, CET, CEST, … https://www.timeanddate.com/time/zones/

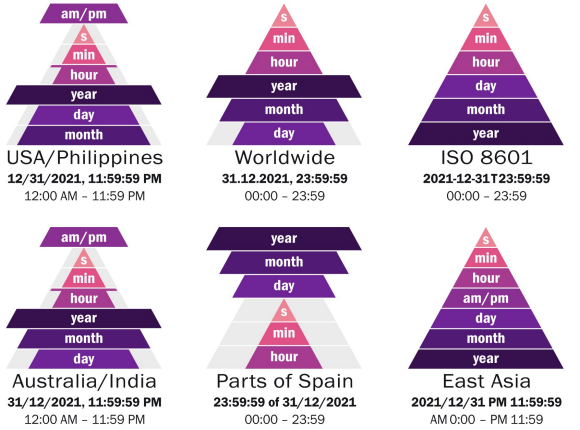
Unix time stamp

The unix time stamp is a way to track time as a running total of seconds. This count starts at the Unix Epoch on January 1st, 1970 at UTC. Therefore, the unix time stamp is merely the number of seconds between a particular date and the Unix Epoch.



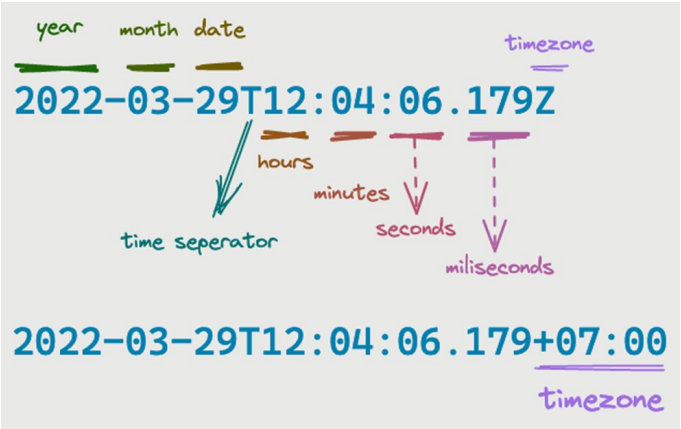
https://www.epochconverter.com/

Too many formats



Sources: CLDR 38.1 / ISO 8601

ISO 8601



https://dev.to/raaynaldo/javascript-datetime-111e

Data Interoperability and Semantics

Part 1. Encoding base data types

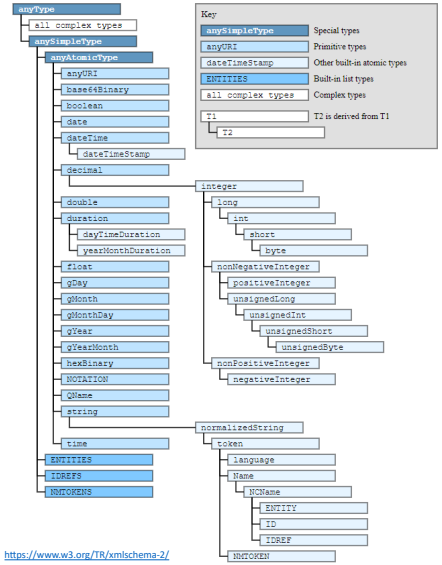
Part 1.7. XML Schema Datatypes

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

XML Schema Datatypes

A **datatype** is denoted by a IRI and has three properties:

- A **value space**, which is a set of values.
- A **lexical space**, which is a set of **literals** used to denote the values.
- A lexical-to-value mapping
- A small collection of *functions, relations, and procedures* associated with the datatype.



https://www.w3.org/TR/xmlschema-2/

Goal: type elements and attributes in XML

```
<ROOT>
  <Customers>
    <Customer CustomerName="Arshad Ali" CustomerID="C001">
      <Orders>
        <Order OrderDate="2012-07-04T00:00:00" OrderID="10248">
          <OrderDetail Quantity="5" ProductID="10" />
          <OrderDetail Quantity="12" ProductID="11" />
          <OrderDetail Quantity="10" ProductID="42" />
        </Order>
      </Orders>
      <Address> Address line 1, 2, 3</Address>
    </Customer>
    <Customer CustomerName="Paul Henriot" CustomerID="C002">
      <Orders>
        <Order OrderDate="2011-07-04T00:00:00" OrderID="10245">
          <OrderDetail Quantity="12" ProductID="11" />
          <OrderDetail Quantity="10" ProductID="42" />
        </Order>
      </Orders>
      <Address> Address line 5, 6, 7</Address>
    </Customer>
    <Customer CustomerName="Carlos Gonzlez" CustomerID="C003">
      <Orders>
        <Order OrderDate="2012-08-16T00:00:00" OrderID="10283">
          <OrderDetail Quantity="3" ProductID="72" />
        </Order>
      </Orders>
      <Address> Address line 1, 4, 5</Address>
    </Customer>
  </Customers>
</ROOT>
```

```
<xsd:attribute name="OrderID" type="xsd:integer"/>

<xsd:attribute name="OrderDate" type="xsd:dateTime"/>

<xsd:attribute name="CustomerID">
    <xsd:simpleType>
        <xsd:restriction base="xsd:string">
            <xsd:pattern value="/^C[0-9]{3}$/"></xsd:pattern>
        </xsd:restriction>
    </xsd:simpleType>
</xsd:attribute>
```


Anatomy of a XSD literal

"24"^^xsd:**integer**

"true"^^xsd:**boolean**

"2001-10-26T21:32:52+02:00"^^xsd:**string**

lexical form — datatype IRI

@prefix xsd: <http://www.w3.org/2001/XMLSchema#>.

xsd:int vs xsd:integer

A **xsd:int** represents a signed 32-bit integer

A **xsd:integer** is an integer unbounded value

https://studylib.net/doc/17671880/ieee-754-standard-representation-of-floating-point-number...

xsd:float vs xsd:double vs xsd:decimal

A **xsd:float** is patterned after the IEEE single-precision 32-bit floating point datatype

```
(\+|-)?([0-9]+(\.[0-9]*)?|\.[0-9]+)([Ee](\+|-)?[0-9]+)? |(\+|-)?INF|NaN
```

A **xsd:double** is patterned after the IEEE double-precision 64-bit floating point datatype

```
(\+|-)?([0-9]+(\.[0-9]*)?|\.[0-9]+)([Ee](\+|-)?[0-9]+)? |(\+|-)?INF|NaN
```

A **xsd:integer** represents a subset of the real numbers, which can be represented by decimal numerals

```
(\+|-)?([0-9]+(\.[0-9]*)?|\.[0-9]+)
```

https://studylib.net/doc/17671880/ieee-754-standard-representation-of-floating-point-number...

Data Interoperability and Semantics

Part 1. Encoding base data types

Part 1.8. Codes: countries, 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

Country codes

ISO 3166-1 – Codes for the representation of names of countries and their subdivisions – Part 1: Country codes

ISO 3166[1]			ISO 3166-1[2]			ISO 3166-2[3]	
Country name[5]	Official state name[6]	Sovereignty[6][7][8]	Alpha-2 code[5]	Alpha-3 code[5]	Numeric code[5]	Subdivision code links[3]	Internet ccTLD[9]
France[I]	The French Republic	UN member state	FR	FRA	250	ISO 3166-2:FR	.fr
United States of America (the)	The United States of America	UN member state	US	USA	840	ISO 3166-2:US	.us
China	The People's Republic of China	UN member state	CN	CHN	156	ISO 3166-2:CN	.cn
Austria	The Republic of Austria	UN member state	AT	AUT	040	ISO 3166-2:AT	.at

Example of country codes
Source: https://en.wikipedia.org/wiki/List_of_ISO_3166_country_codes

Language codes

ISO 639 is a standardized nomenclature used to classify languages. Each language is assigned a two-letter (639-1) and three-letter (639-2 and 639-3) lowercase abbreviation

ISO language name	639-1	639-2/T	639-2/B	639-3	
English	en	eng	eng	eng	
Chinese	zh	zho	**chi**	zho + 16	macrolanguage
Hindi	hi	hin	hin	hin	
Spanish, Castilian	es	spa	spa	spa	
French	fr	fra	**fre**	fra	
Arabic	ar	ara	ara	ara + 29	macrolanguage, Standard Arabic is arb
Bengali	bn	ben	ben	ben	

Example of language names for the most spoken languages
Source: https://en.wikipedia.org/wiki/List_of_ISO_639-1_codes

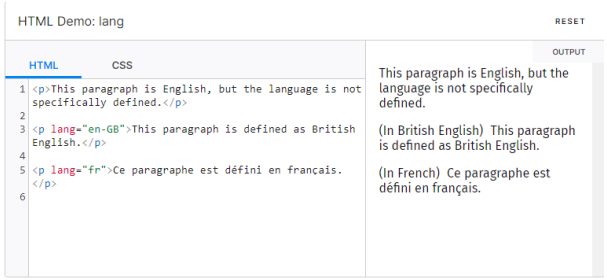
IETF BCP47: Tags for Identifying Languages

Internet Engineering Task Force (IETF) « Best Current Practice » (BCP)

Region subtags

en-GB
es-005
zh-Hant-HK

Read more in the BCP 47 spec:
2.2.4 Region Subtag
4.1 Choice of Language Tag



https://developer.mozilla.org/en-US/docs/Web/HTML/Global_attributes/lang

Currency codes

ISO 4217 defines alpha codes and numeric codes for the representation of currencies and provides information about the relationships between individual currencies and their minor units.

Active ISO 4217 currency codes[1] [hide]

Code	Num	D[a]	Currency	Locations listed for this currency[b]
AED	784	2	United Arab Emirates dirham	United Arab Emirates
EUR	978	2	Euro	Åland Islands (AX), European Union (EU), Andorra (AD), Austria (AT), Belgium (BE), Cyprus (CY), Estonia (EE), Finland (FI), France (FR), French Southern and Antarctic Lands (TF), Germany (DE), Greece (GR), Guadeloupe (GP), Ireland (IE), Italy (IT), Latvia (LV), Lithuania (LT), Luxembourg (LU), Malta (MT), French Guiana (GF), Martinique (MQ), Mayotte (YT), Monaco (MC), Montenegro (ME), Netherlands (NL), Portugal (PT), Réunion (RE), Saint Barthélemy (BL), Saint Martin (MF), Saint Pierre and Miquelon (PM), San Marino (SM), Slovakia (SK), Slovenia (SI), Spain (ES), Vatican City (VA)
KRW	410	0[c]	South Korean won	South Korea
GBP	826	2	Pound sterling	United Kingdom, Isle of Man (IM, see Manx pound), Jersey (JE, see Jersey pound), Guernsey (GG, see Guernsey pound), Tristan da Cunha (SH-TA)
CHF	756	2	Swiss franc	Switzerland, Liechtenstein (LI)

Example of currency codes
Source: https://en.wikipedia.org/wiki/ISO_4217

Data Interoperability and Semantics

Part 1. Encoding base data types

Part 1.9. Quantities and Units of measure

Units of measures: no consensus

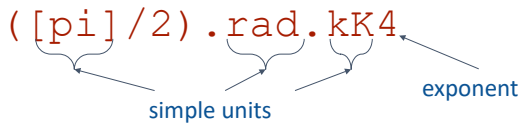
- BIPM (International Bureau of Weights and Measures)
- ISO/IEC 1000 – ISO/IEC 80000
- VIM (international Vocabulary for Measurements)
- UnitsML
- UCUM (Unified Code for Units of Measure)
- UNECE Recommendation 20
- Sweet
- QUDT
- …

https://en.wikipedia.org/wiki/Category:Encodings

UCUM: Unified Code for Units of Measure

- A code system intended to include *all units* of measures being contemporarily used in international science, engineering, and business.
- **Used** by international organizations and standards
- **No ambiguity** possible
- **Clear semantics** of units
- **Con: Problematic** custom license

the Unified Code for Units of Measure



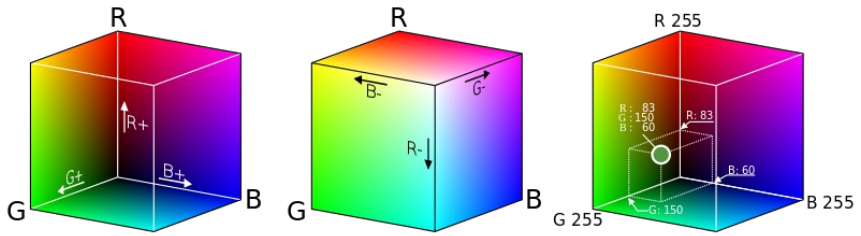
UCUM Code of Unit Concept	EN Unit	EN Symbol	EN Dimension[a]	NCI Concept Code	NCI Term	NCI Abbreviation	SNOMED CT Identifier[b]
[IU]			[arb]	C70497	Anti-Xa Activity International Unit	anti-Xa activity	258997004
Bq	becquerel	Bq	T^{-1}	C42562	Becquerel	Bq	282141004
Bq/g			$M^{-1}T^{-1}$	C70522	Becquerel per gram	Bq/g	
10^9.[CFU]			[arb]	C68897	Billion Colony Forming Units	Billion CFU	
10^9			1	C71189	Billion Organisms		
m3	cubic metre	m^3	L^3	C42570	Cubic Meter	m^3	396154006
Ci/ml			$L^{-3}T^{-1}$	C71172	Curie per Millilitre	Ci/ml	
d	day	d	T	C25301	Day	d	258703001
[drp]			L^3	C48491	Drop Dosing Unit	Gtt	404218003
[IU]/ml			[arb]	C67377	International Unit per Millilitre	IU/mL	259002007
k[USP'U]			[arb]	C71202	Kilo United States Pharmacopoeia Unit	KUSP'U	
kBq/l			$L^{-3}T^{-1}$	C71167	Kilobecquerel per Liter	kBq/L	
mmol/l			$L^{-3}N$	C64387	Millimole per Liter	mmol/L	258813002
[ppm]	part per million	ppm	1	C48523	Part Per Million	ppm	258731005
Pa	pascal	Pa	$L^{-1}MT^{-2}$	C42547	Pascal	P	259016002
%	per cent	%	1	C48570	Percent	%	118582008
%			1	C48571	Percent Volume per Volume	%V/V	419569009
g/ml	per cent (w/v)	%(w/v)	$L^{-3}M$	C48527	Percent Weight Volume	%M/V	396169007
%			1	C48528	Percent Weight Weight	%W/W	118582008
[PFU]			[arb]	C73575	Plaque Forming Unit Equivalent 1000 Mouse LD50	PFU Equivalent 1000 Mouse LD50	
[lb_av]	pound	lb	M	C48531	Pound	LB	258693003
/min	revolution per minute	r.p.m., rev/min, r/min	T^{-1}	C70469	Revolution per Minute	rpm	286549009
[tb'U]			[arb]	C65132	Tuberculin Unit		415758003
[arb'U]{ELISA}				C68875	Enzyme-Linked Immunosorbent Assay Unit	EL. U	

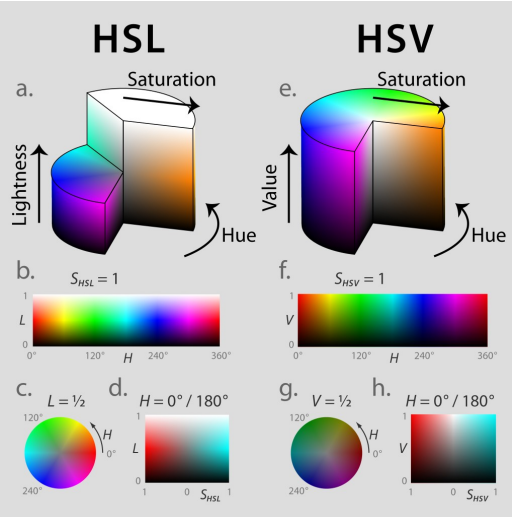
Data Interoperability and Semantics

Part 1. Encoding base data types

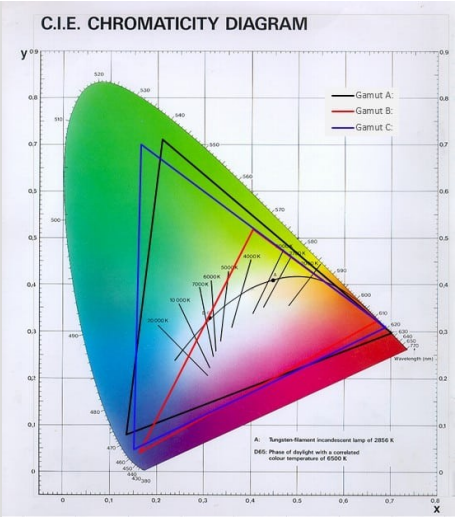
Part 1.10. Colors

RGB color cube





HSL (for hue, saturation, lightness)
HSV (for hue, saturation, value; also known as HSB, for hue, saturation, brightness)
https://en.wikipedia.org/wiki/HSL_and_HSV



HSL HSV

```
{
  "1": {
    "state": {
      "on": true,
      "bri": 254,
      "hue": 14314,
      "sat": 172,
      "effect": "none",
      "xy": [
        0.4791,
        0.4139
      ],
      "ct": 405,
      "alert": "none",
      "colormode": "ct",
      "reachable": true
    },
    "type": "Extended color light",
    "name": "Hue color light 1",
    "modelid": "LCT001",
    "manufacturername": "Philips",
    "uniqueid": "00:17:88:01:00:ff:9a:28-0b",
    "swversion": "5.127.1.26581"
  }
}
```

C.I.E. CHROMATICITY DIAGRAM
Gamut A:
Gamut B:
Gamut C:

HSV ⟷ RGB conversion

When $0 \le H < 360$, $0 \le S \le 1$ and $0 \le V \le 1$:

$$C = V \times S$$

$$X = C \times (1 - |(H / 60°) \bmod 2 - 1|)$$

$$m = V - C$$

$$(R', G', B') = \begin{cases} (C, X, 0) & , 0° \le H < 60° \\ (X, C, 0) & , 60° \le H < 120° \\ (0, C, X) & , 120° \le H < 180° \\ (0, X, C) & , 180° \le H < 240° \\ (X, 0, C) & , 240° \le H < 300° \\ (C, 0, X) & , 300° \le H < 360° \end{cases}$$

$$(R,G,B) = ((R'+m)\times 255, (G'+m)\times 255, (B'+m)\times 255)$$

R′ = R / 255
G′ = G / 255
B′ = B / 255

Cmax = max(R′, G′, B′)
Cmin = min(R′, G′, B)
Δ = Cmax – Cmin

$$H = \begin{cases} 0°, \Delta = 0 \\ 60° \times \left(\frac{G'-B'}{\Delta} \ (mod\ 6)\right), Cmax = R' \\ 60° \times \left(\frac{B'-R'}{\Delta} + 2\right), Cmax = G' \\ 60° \times \left(\frac{R'-G'}{\Delta} + 4\right), Cmax = B' \end{cases}$$

$$S = \begin{cases} 0, Cmax = 0 \\ \frac{\Delta}{Cmax}, Cmax \neq 0 \end{cases}$$

V = Cmax

Data Interoperability and Semantics

</ Part 1. Encoding base data types >

Data Interoperability and Semantics

< Part 2. Data Formats >

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

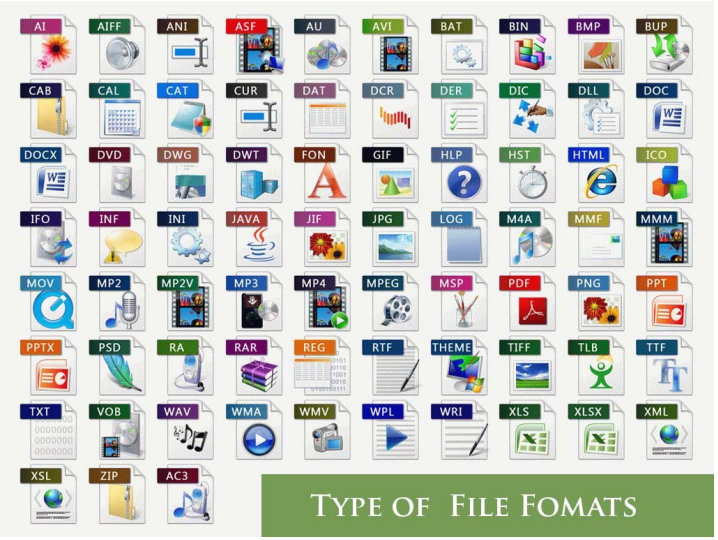
Part 2.1 Generalities

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]

1 Archive and compressed
1.1 Physical recordable media archiving
2 Computer-aided design
2.1 Computer-aided design (CAD)
2.2 Electronic design automation (EDA)
2.3 Test technology
3 Database
4 Big Data (Distributed)
5 Desktop publishing
6 Document
7 Financial records
7.1 Financial data transfer formats
8 Font file
9 Geographic information system
10 Graphical information organizers
11 Graphics
11.1 Color palettes
11.2 Color management
11.3 Raster graphics
11.4 Vector graphics
11.5 3D graphics
12 Links and shortcuts
13 Mathematical
14 Object code, executable files, shared and dynamically linked libraries
15 Page description language
16 Personal information manager
17 Presentation
18 Project management software
19 Reference management software
20 Scientific data (data exchange)
20.1 Multi-domain
20.2 Meteorology
20.3 Chemistry
20.4 Mathematics
20.5 Biology
20.6 Biomedical imaging
20.7 Biomedical signals (time series)
20.8 Other biomedical formats
20.9 Biometric formats
21 Programming languages and scripts
22 Security
22.1 Certificates and keys
22.1.1 X.509
22.2 Encrypted files
22.3 Password files
23 Signal data (non-audio)
24 Sound and music
24.1 Lossless audio
24.1.1 Uncompressed
24.1.2 Compressed
24.2 Lossy audio
24.3 Tracker modules and related
24.4 Sheet music files
24.5 Other file formats pertaining to audio
25 Playlist formats
26 Audio editing and music production
27 Recorded television formats
28 Source code for computer programs
29 Spreadsheet
30 Tabulated data
31 Video
31.1 Video editing, production
32 Video game data
33 Video game storage media
34 Virtual machines
34.1 Microsoft Virtual PC, Virtual Server
34.2 EMC VMware ESX, GSX, Workstation, Player
34.3 VirtualBox
34.4 Parallels Workstation
34.5 QEMU
35 Web page
36 Markup languages and other web standards-based formats
37 Other
37.1 Cursors
38 Generalized files
38.1 General data formats
38.1.1 Text-based
38.2 Generic file extensions
38.2.1 Binary files
38.2.2 Text files
38.3 Partial files
38.3.1 Differences and patches
38.3.2 Incomplete transfers
38.4 Temporary files
39 See also
40 References
41 External links

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

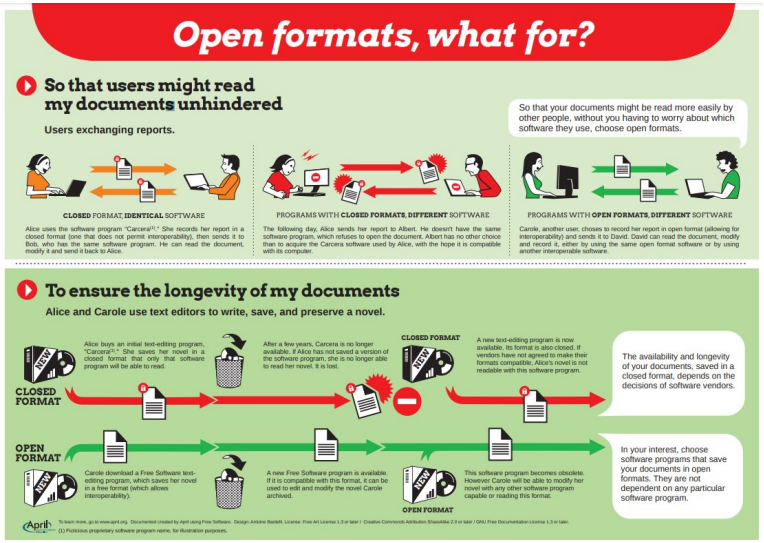
Open file format

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. »

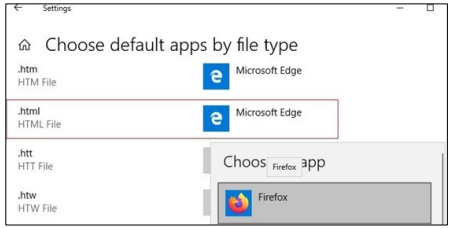


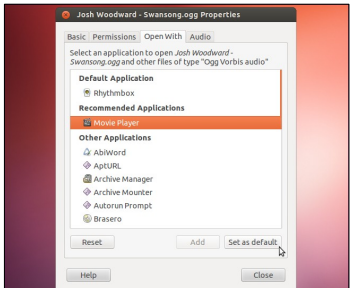
https://www.april.org/en/open-formats

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␣	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>␊	0	deb	linux deb file
37 7A BC AF 27 1C	7z¼¯'␜	0	7z	7-Zip File Format
1F 8B	␋␈	0	gz tar.gz	GZIP compressed file[1]
FD 37 7A 58 5A 00	ý7zXZ␀	0	xz tar.xz	XZ compression utility using LZMA2 compression
4E 45 53 1A	NES␚	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 ?

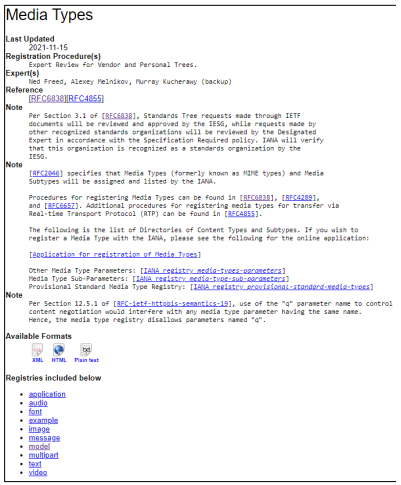
Example: the gzip file format starts with:

- 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 ?

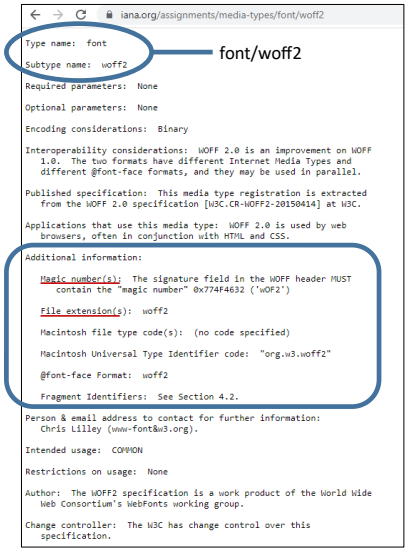
MIME type = Media types

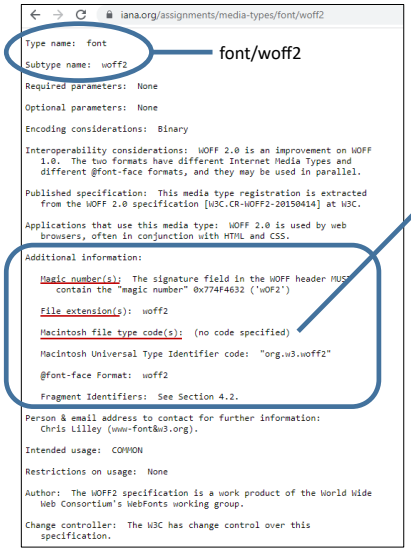
- managed by Internet Assigned Numbers Authority (IANA)
- example: text/html, image/gif, font/woff2, ...
- used by many internet protocols
- demo with your browser: with the Network tab of the developer tools open, see HTTP response headers of requests when accessing https://fonts.google.com/


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], [RFC4288], and [RFC4855]. 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



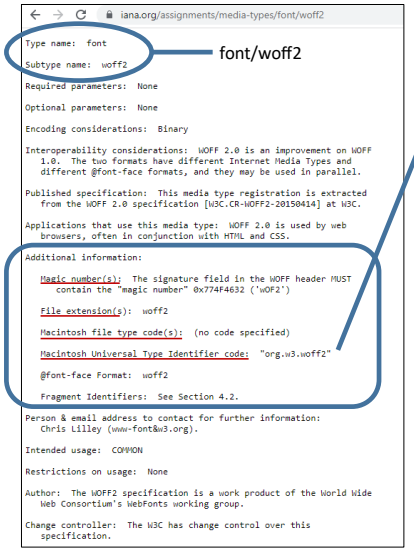


Macintosh file type code(s)

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

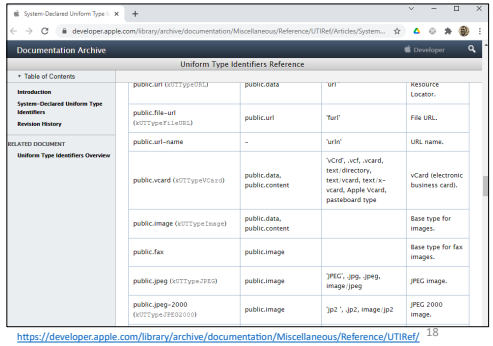
PICT	picture	Stores a PICT image contained in the file
PREF	preference	Stores the environment settings for an application
snd	sound	Stores a sound used in the file
STR	string	Stores a string or hexadecimal data used in the file
STR#	string list	Stores multiple strings used in the file
styl	style	Defines 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



Macintosh Universal Type Identifier code

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



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

Data Interoperability and Semantics

Part 2. Data Formats

Part 2.2. Delimiter-separated values

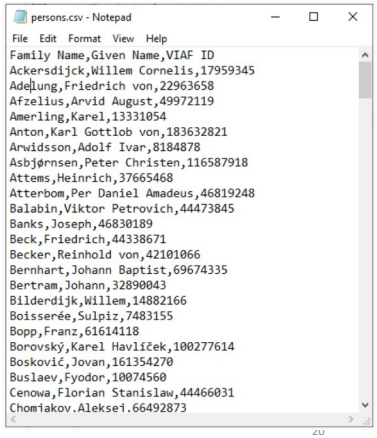
Delimiter-separated values

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
```

[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

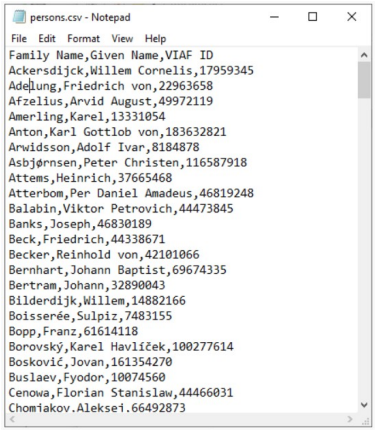
```
field_name,field_name,field_name CRLF
aaa,bbb,ccc CRLF
zzz,yyy,xxx CRLF
```

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

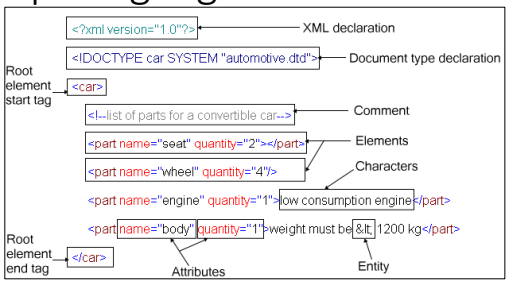
Part 2. Data Formats

Part 2.3. Extensible Markup Language (XML)

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 ':'

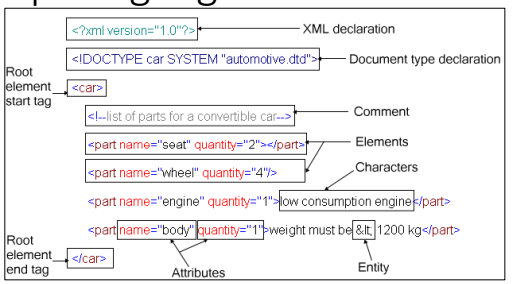
Schemas and validation
- *valid vs invalid*
- DTD, or XML Schema

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

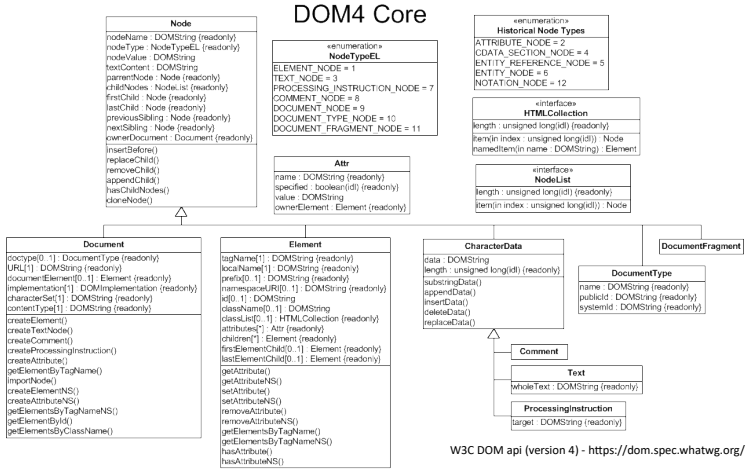
XML (file format)

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

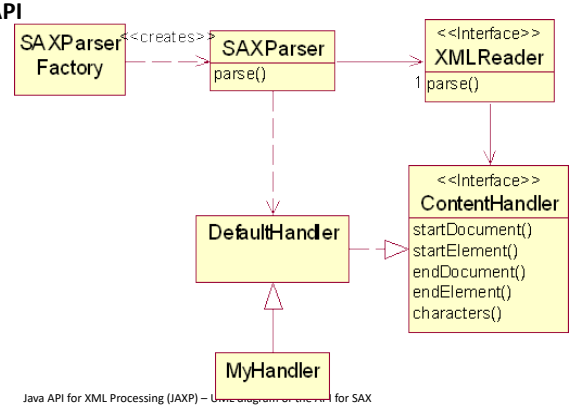
XML manipulation: Document Object Model (DOM)



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

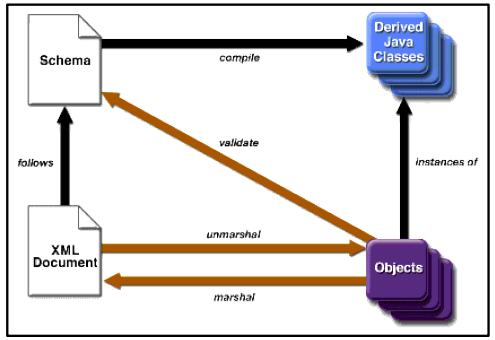
SAX (Simple API for XML)

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

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

```
employee.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"

//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"

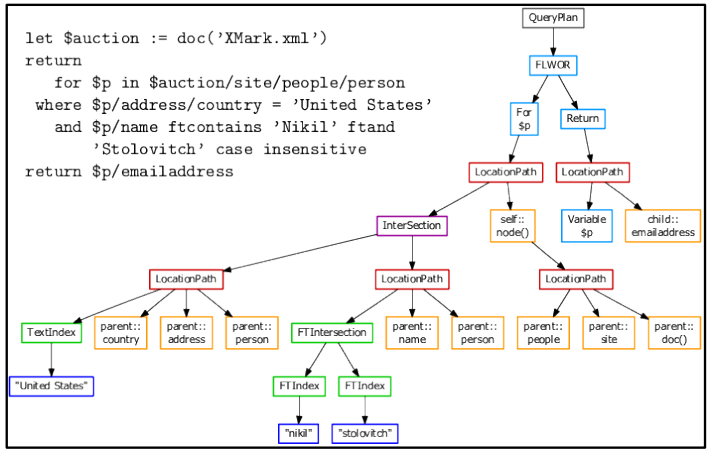
//a[string-length(@href) > 70]

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

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

XQuery: Declarative query language

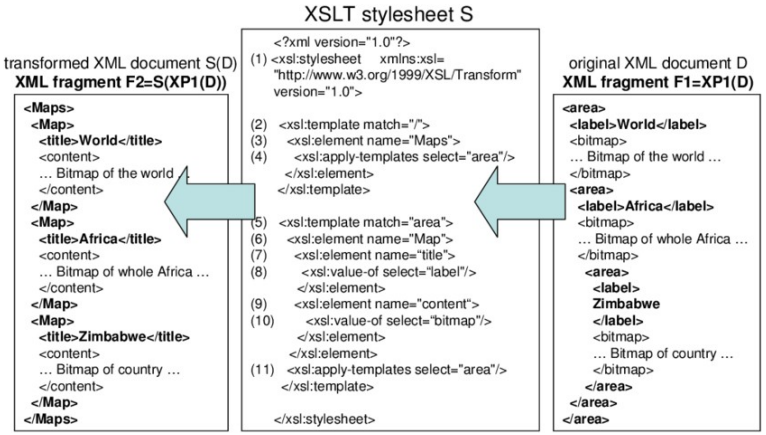
```
let $auction := doc('XMark.xml')
return
   for $p in $auction/site/people/person
 where $p/address/country = 'United States'
   and $p/name ftcontains 'Nikil' ftand
       'Stolovitch' case insensitive
return $p/emailaddress
```



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

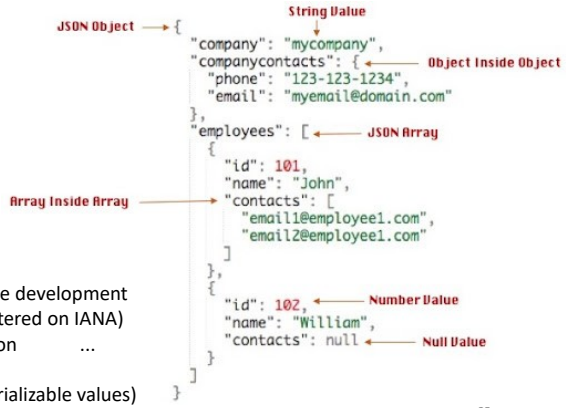
- since early 2000s. Now used a lot in software development
- many mediatypes based on JSON (118 registered on IANA)
  application/geo+json application/ld+json ...
- simple, human-readable
- attribute–value pairs and arrays (or other serializable values)

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

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
    }
  }
}
```

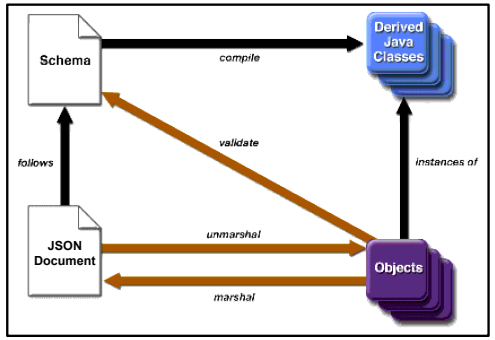
XPath	JSONPath	Result
/store/book/author	$.store.book[*].author	the authors of all books in the store
//author	$..author	all authors
/store/*	$.store.*	all things in store, which are some books and a red bicycle.
/store//price	$.store..price	the price of everything in the store.
//book[3]	$..book[2]	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

Java ⇄ {JSON}

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`

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

```
# 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
```

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/

toml-lang/**toml**
Tom's Obvious, Minimal Language

- 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

python-poetry/**poetry**
Python dependency management and packaging made easy.

```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)

YAML Ain't Markup Language (YAML)

YAML

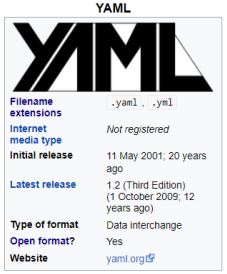
Filename extensions	.yaml , .yml
Internet media type	*Not registered*
Initial release	11 May 2001; 20 years ago
Latest release	1.2 (Third Edition) (1 October 2009; 12 years ago)
Type of format	Data interchange
Open format?	Yes
Website	yaml.org

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

```yaml
--- !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)



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

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

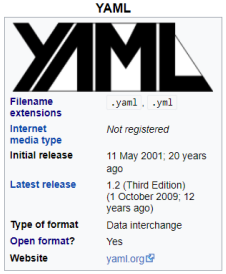
or

```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)



- strings

```yaml
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
```

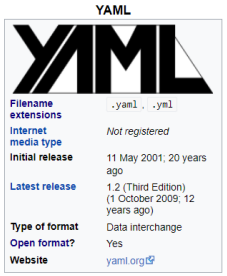
or

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

  Blank lines denote
  paragraph breaks
```

- datatypes

```yaml
---
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)



- 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)

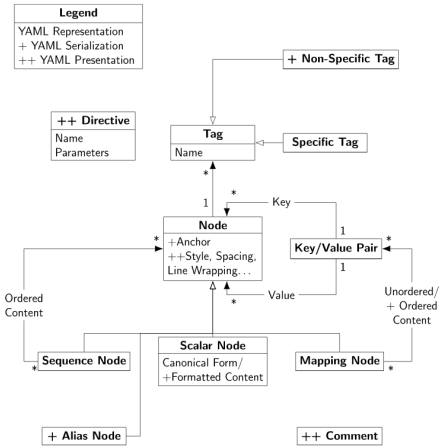
https://yaml.org/refcard.html

```
%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 }
```

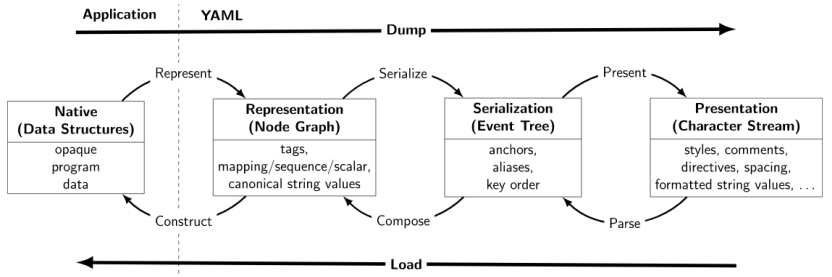
YAML Ain’t Markup Language (YAML)

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

Lightweight markup languages

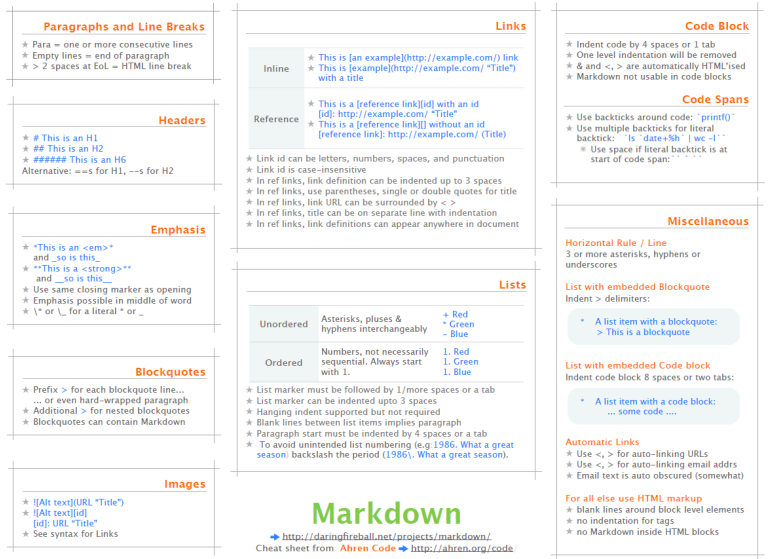
Comparing language features

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 bold text	emphasized text italic text	<code>code</code> <tt>monospace text</tt>	semantic presentational
AsciiDoc	*bold text*	'italic text' _italic text_	+monospace text+ `monospace text`	Can double operators to apply formatting where there is no word boundary example **b**old t**ex**t yields bold 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** __bold text__	*italic text* _italic text_	`monospace text`	semantic HTML tags
MarkedText[9]	**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* **bold text**	_emphasis_ __italic text__	@monospace text@	semantic HTML tags presentational HTML tags
Texy!	**bold text**	*italic text* //italic text//	`monospace text`	semantic HTML tags by default, optional support for presentational tags
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



Data Interoperability and Semantics

Part 2. Data Formats

Part 2.8. Compressed formats

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...

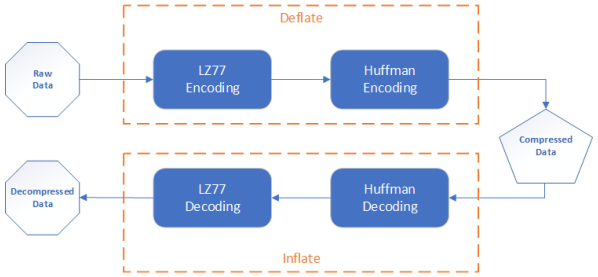
CC

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

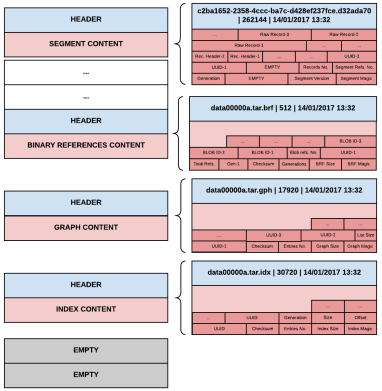
Filename extension	`.gz`
Internet media type	`application/gzip`[2]
Uniform Type Identifier (UTI)	org.gnu.gnu-zip-archive
Magic number	`1f 8b`
Developed by	Jean-loup Gailly and Mark Adler
Type of format	Data compression
Open format?	Yes
Website	gzip.org (obsolete)

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

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

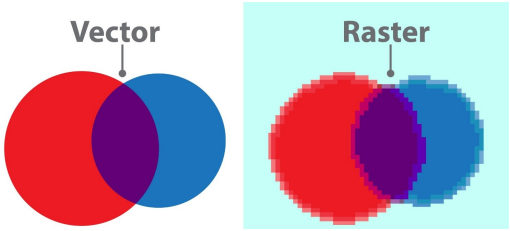
Authority control: National libraries — France (data)

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

65

Image file formats

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

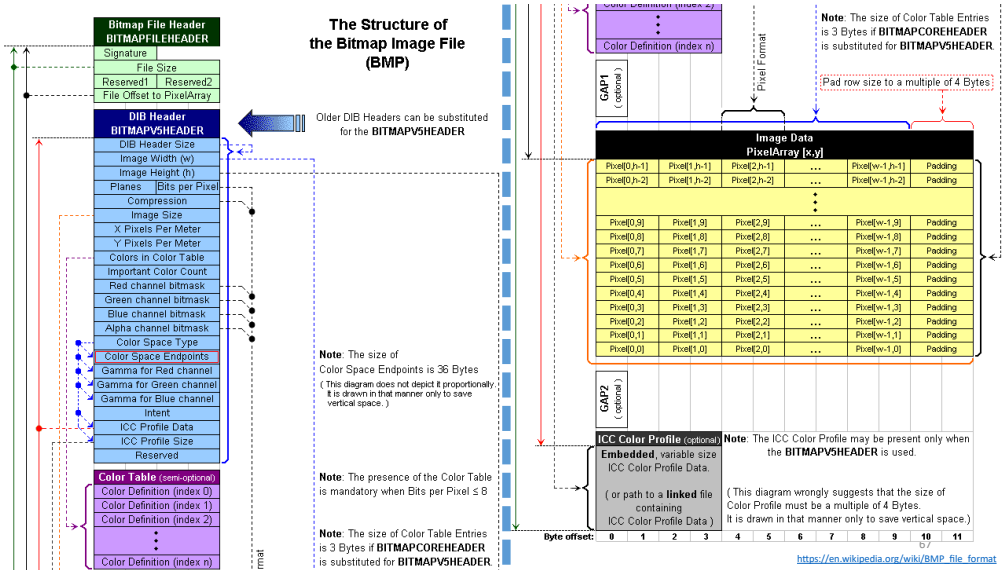


Graphics file formats [hide]

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

66



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

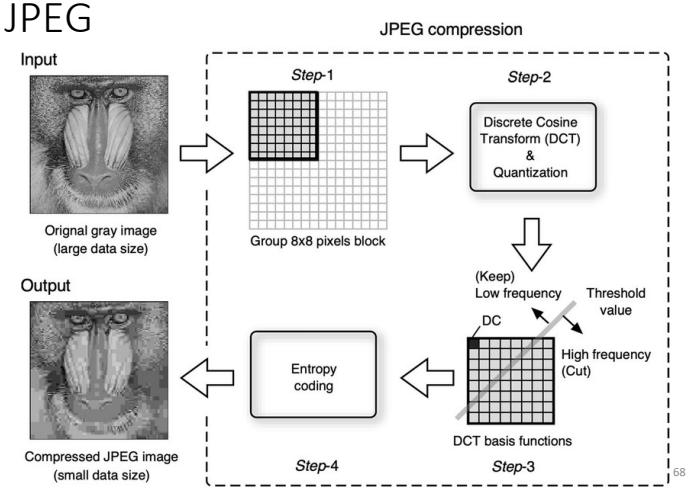
67

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/

JPEG



68

Portable Network Graphics

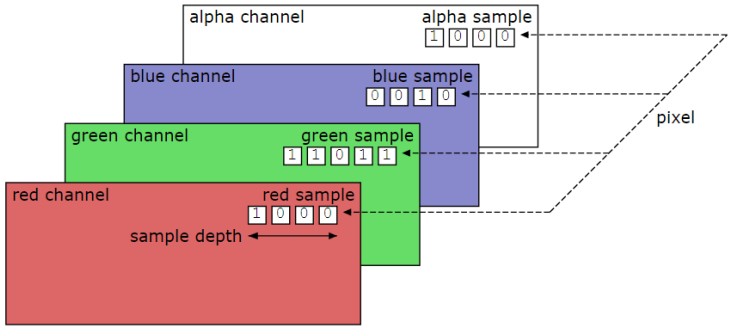
A PNG image with an 8-bit transparency channel, overlaid onto a checkered background, typically used in graphics software to indicate transparency

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

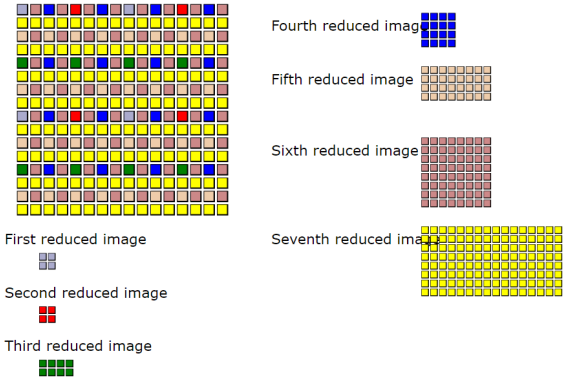
A PNG image with an 8-bit transparency channel, overlaid onto a checkered background, typically used in graphics software to indicate transparency

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



WebP

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: 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.

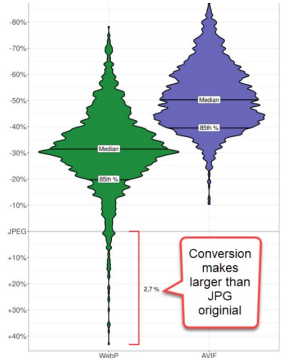
Image	File Name	Original Size	Compressed JPG	WebP Format
	jpg-to-webp-1.jpg	480 KB	407 KB	43 KB
	jpg-to-webp-2.jpg	659 KB	578 KB	113 KB
	jpg-to-webp-3.jpg	787 KB	715 KB	127 KB
	jpg-to-webp-4.jpg	617 KB	543 KB	61 KB

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.

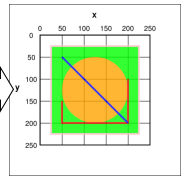


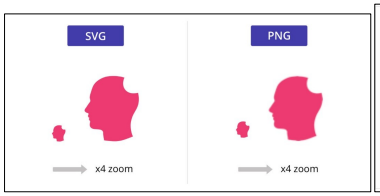
https://imageengine.io/blog/how-efficient-is-avif/

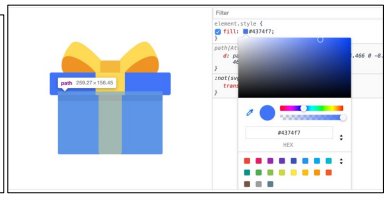
Scalable Vector Graphics

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

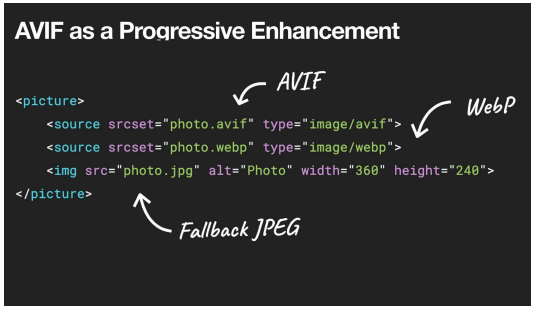
https://developer.mozilla.org/en-US/docs/Web/Media/Formats

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 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)

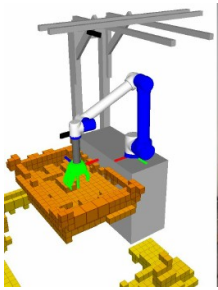
Data Interoperability and Semantics

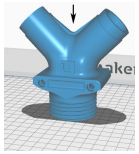
Part 2. Data Formats

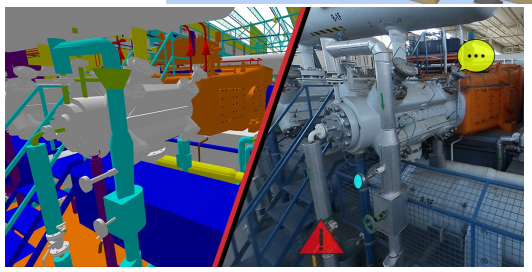
Part 2.10 3D models

(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

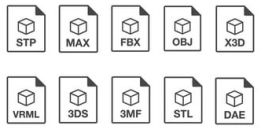




3D models

Many file formats



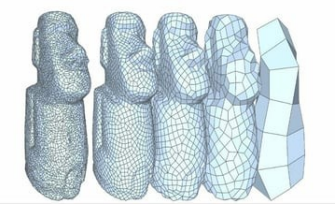
Key features of a 3D file:

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

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

Shape Geometry

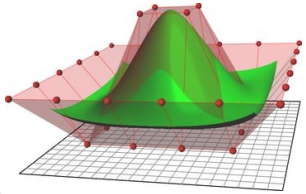
a) Approximate Mesh Encoding
« tesselations »: decompose a surface in polygons (usually triangles)
Good for 3D printing
Files may be huge.



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

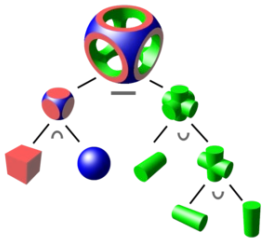
https://all3dp.com/2/most-common-3d-file-formats-model/

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



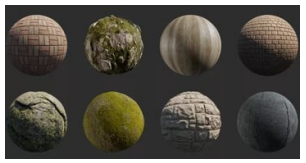
https://en.wikipedia.org/wiki/Non-uniform_rational_B-spline

c) Constructive Solid Geometry
Primitive shapes that are combined using Boolean operations
Good for CAD



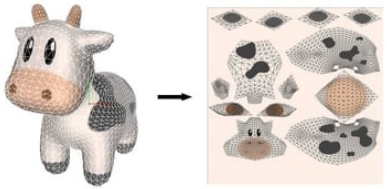
https://en.wikipedia.org/wiki/Constructive_solid_geometry

Surface Textures



https://www.creativebloq.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.



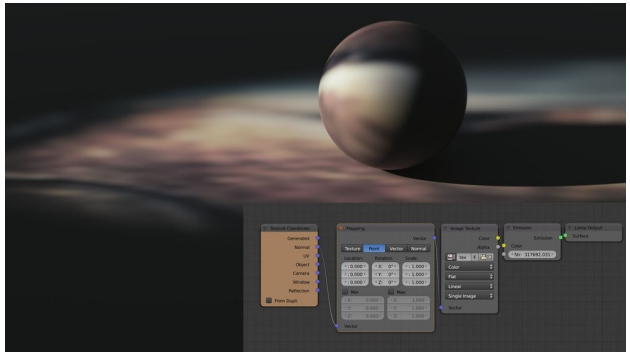
https://metalbyexample.com/textures-and-samplers/

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

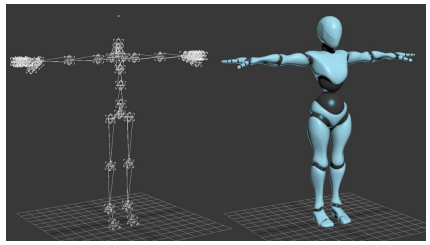
- layout of the 3D model in terms of cameras, light sources, and other nearby 3D models



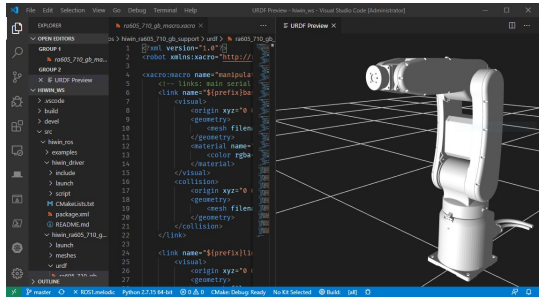
https://www.creativeshrimp.com/light-texture-tutorial-1-lighting-book.html

Animation

- skeletal animation: skeleton + joints



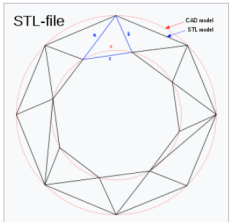
http://www.downloads.redway3d.com/downloads/public/documentation/wf_skeletal_animation.html



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”)

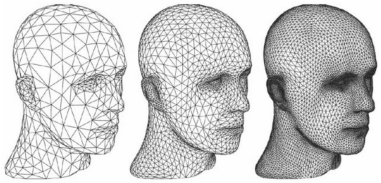
STL



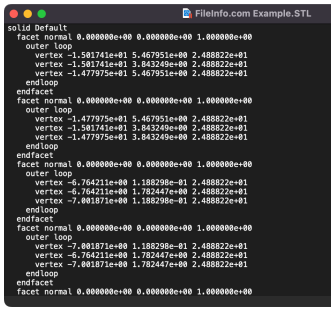
A CAD representation of a torus (shown as two concentric red circles) and an STL approximation of the same shape (composed of triangular planes)

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)



- 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

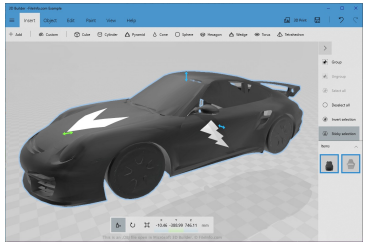
OBJ geometry format

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

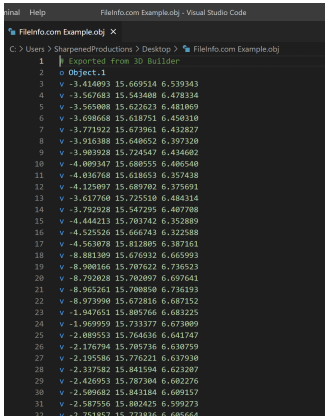
MTL material format

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)

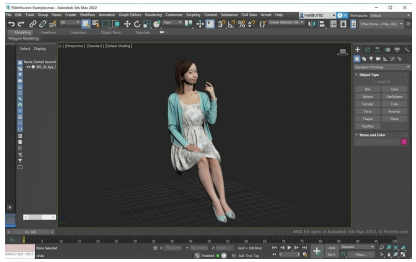
https://people.sc.fsu.edu/~jburkardt/data/mtl/mtl.html

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

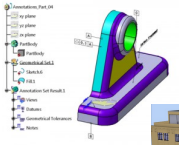
Reference: http://paulbourke.net/dataformats/3ds/
Tree structure made of chunks
In binary :
- 6 byte Chunk header:
  - 1-2 = chunk ID
  - 3-6 = little-endian length of the chunk
- Next bytes: chunk’s data, including sub-chunks

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

Data Interoperability and Semantics

< Part 3. Data schemas and semantics >

Data Interoperability and Semantics
Outline

- < Part 3. Data schemas and semantics >
  - Part 3.1. Data schemas
    - Part 3.1.1. XML Schema
    - Part 3.1.2. JSON Schema
  - Part 3.2. Semantics
    - Part 3.2.1. Heterogeneities and data conflicts
    - Part 3.2.2. Controlled vocabularies and ontologies
    - Part 3.2.3. Resource Description Framework
    - Part 3.2.4. RDFa: Rich structured data markup for web documents
    - Part 3.2.5. JSON-LD: JSON for Linking Data

Data Interoperability and Semantics

< Part 3. Data schemas and semantics >

Part 3.1. Data schemas

Data Interoperability and Semantics

< Part 3. Data schemas and semantics >

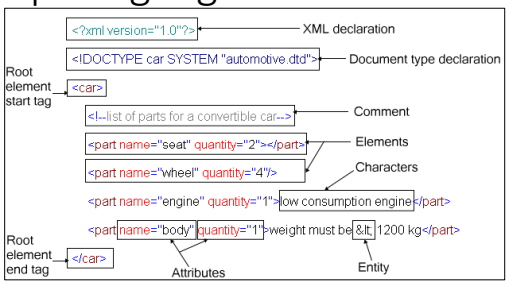
Part 3.1. Data schemas

Part 3.1.1. XML Schema

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

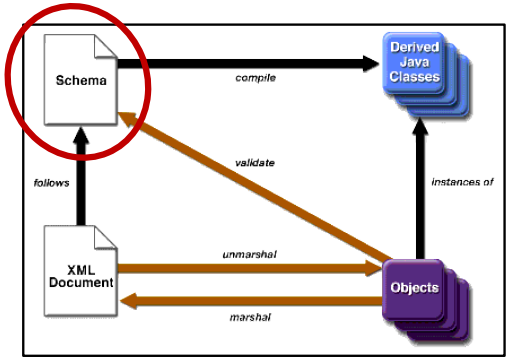
Characters and escaping
- unicode implementations: `<?xml version="1.0" encoding="UTF-8"?>`
- escaping characters: `<` '<' - `&` '&' - `❤` '❤' - etc.

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>`

XML – OOP data binding



XML-OOP data binding

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

```
Employee.java

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

```
employee.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/

XML Schema

- Document Type Definition (DTD, 2008)

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN"
"http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd">
```

link to some .dtd document

```
[...]
<!ELEMENT html (head, body)>
<!ELEMENT p (#PCDATA | p | ul | dl | table | h1|h2|h3)*>
[...]
<!ATTLIST img
    src     CDATA          #REQUIRED
    id      ID             #IMPLIED
    sort    CDATA          #FIXED "true"
    print   (yes | no) "yes"
>
[...]
```

```
<?xml version="1.0" encoding="utf-8"?>
<!DOCTYPE html>
<!-- the XHTML document body starts here-->
<html xmlns="http://www.w3.org/1999/xhtml">
 ...
</html>
```

HTML 5 : no more link to a DTD

XML Schema

- Regular Language for XML Next Generation (RELAX NG, 2001) https://relaxng.org/

```
<element name="addressBook" xmlns="http://relaxng.org/ns/structure/1.0">
  <zeroOrMore>
    <element name="card">
      <element name="name">
        <text/>
      </element>
      <element name="email">
        <text/>
      </element>
    </element>
  </zeroOrMore>
</element>
```

RELAX NG document

```
<addressBook>
  <card>
    <name>John Smith</name>
    <email>js@example.com</email>
  </card>
  <card>
    <name>Fred Bloggs</name>
    <email>fb@example.net</email>
  </card>
</addressBook>
```

Valid XML document

XML Schema

- XML Schema (W3C, 2003)

  The only one you should use (if you need to)

```
<?xml version="1.0"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema">

<xs:element name="note">
  <xs:complexType>
    <xs:sequence>
      <xs:element name="to" type="xs:string"/>
      <xs:element name="from" type="xs:string"/>
      <xs:element name="heading" type="xs:string"/>
      <xs:element name="body" type="xs:string"/>
    </xs:sequence>
  </xs:complexType>
</xs:element>

</xs:schema>
```

XML Schema document

```
<?xml version="1.0"?>

<note
xmlns="https://www.w3schools.com"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="https://www.w3schools.com/xml note.xsd">
  <to>Tove</to>
  <from>Jani</from>
  <heading>Reminder</heading>
  <body>Don't forget me this weekend!</body>
</note>
```

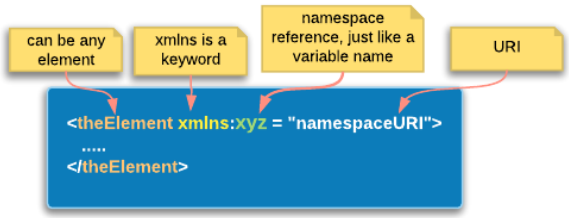
Valid XML document

XML Schema

- Tutorials
  - https://www.tutorialspoint.com/xml/xml_schemas.htm
  - https://www.w3schools.com/xml/schema_intro.asp
  - …
- What to do with XML Schemas
  - validate documents
  - generate forms
  - generate classes (any OOP language)
  - …

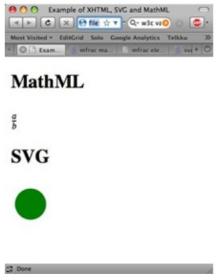
Multiple XML Schemas ?

- How to combine different XML schemas in a single file?
- Using XML namespaces



Example: A document with XHTML + MathML + SVG

```
<?xml version="1.0"?>
<!DOCTYPE html PUBLIC
   "-//W3C//DTD XHTML 1.1 plus MathML 2.0 plus SVG 1.1//EN"
   "http://www.w3.org/2002/04/xhtml-math-svg/xhtml-math-svg-flat.dtd">
<html xmlns     = "http://www.w3.org/1999/xhtml"
    xmlns:svg   = "http://www.w3.org/2000/svg">

 <head>
  <title>Example of XHTML, SVG and MathML</title>
 </head>
 <body>

  <h2>MathML</h2>
  <p>
   <math xmlns="http://www.w3.org/1998/Math/MathML">
    <mfrac>
     <mi>a</mi>
     <mi>b</mi>
    </mfrac>
   </math>
  </p>

  <h2>SVG</h2>

  <p>
   <svg:svg width="50px" height="50px">
    <svg:circle cx="25px" cy="25px" r="20px" fill="green"/>
   </svg:svg>
  </p>

 </body>
</html>
```



Data Interoperability and Semantics

< Part 3. Data schemas and semantics >

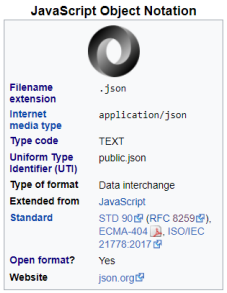
Part 3.1. Data schemas

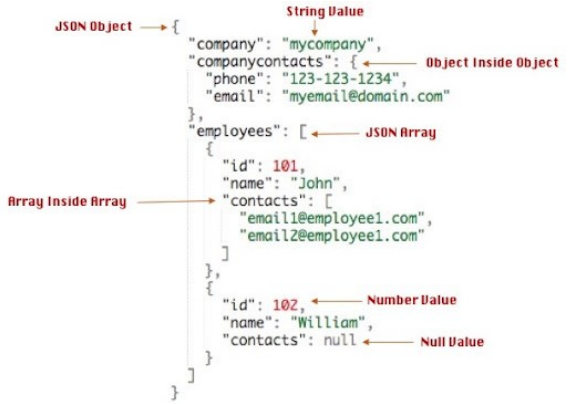
Part 3.1.2. JSON Schema

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

reminder

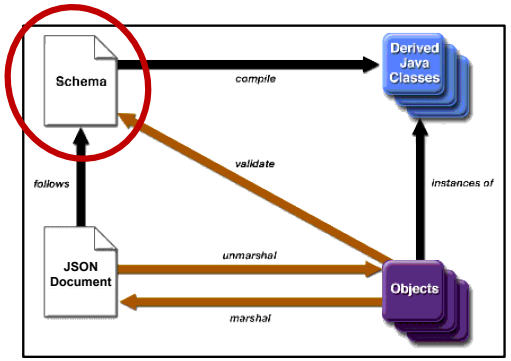
JavaScript Object Notation





JSON Schema

JSON – OOP data binding



JSON-OOP data binding

JSON Schema Specification (latest)

https://json-schema.org/draft/2020-12/json-schema-validation.html

Table of Contents

1. Introduction
2. Conventions and Terminology
3. Overview
4. Interoperability Considerations
   4.1. Validation of String Instances
   4.2. Validation of Numeric Instances
   4.3. Regular Expressions
5. Meta-Schema
6. A Vocabulary for Structural Validation
   6.1. Validation Keywords for Any Instance Type
      6.1.1. type
      6.1.2. enum
      6.1.3. const
   6.2. Validation Keywords for Numeric Instances (number and integer)
      6.2.1. multipleOf
      6.2.2. maximum
      6.2.3. exclusiveMaximum
      6.2.4. minimum
      6.2.5. exclusiveMinimum
   6.3. Validation Keywords for Strings
      6.3.1. maxLength
      6.3.2. minLength
      6.3.3. pattern
   6.4. Validation Keywords for Arrays
      6.4.1. maxItems
      6.4.2. minItems
      6.4.3. uniqueItems
      6.4.4. maxContains
      6.4.5. minContains
   6.5. Validation Keywords for Objects
      6.5.1. maxProperties
      6.5.2. minProperties
      6.5.3. required
      6.5.4. dependentRequired
7. Vocabularies for Semantic Content With "format"
   7.1. Foreword
   7.2. Implementation Requirements
      7.2.1. Format-Annotation Vocabulary
      7.2.2. Format-Assertion Vocabulary
      7.2.3. Custom format attributes
   7.3. Defined Formats
      7.3.1. Dates, Times, and Duration
      7.3.2. Email Addresses
      7.3.3. Hostnames
      7.3.4. IP Addresses
      7.3.5. Resource Identifiers
      7.3.6. uri-template
      7.3.7. JSON Pointers
      7.3.8. regex
8. A Vocabulary for the Contents of String-Encoded Data
   8.1. Foreword
   8.2. Implementation Requirements
   8.3. contentEncoding
   8.4. contentMediaType
   8.5. contentSchema
   8.6. Example
9. A Vocabulary for Basic Meta-Data Annotations
   9.1. "title" and "description"
   9.2. "default"
   9.3. "deprecated"
   9.4. "readOnly" and "writeOnly"
   9.5. "examples"
10. Security Considerations
11. References
   11.1. Normative References
   11.2. Informative References

Appendix A. Keywords Moved from Validation to Core
Appendix B. Acknowledgments
Appendix C. ChangeLog
Authors' Addresses

JSON Schema

Examples:

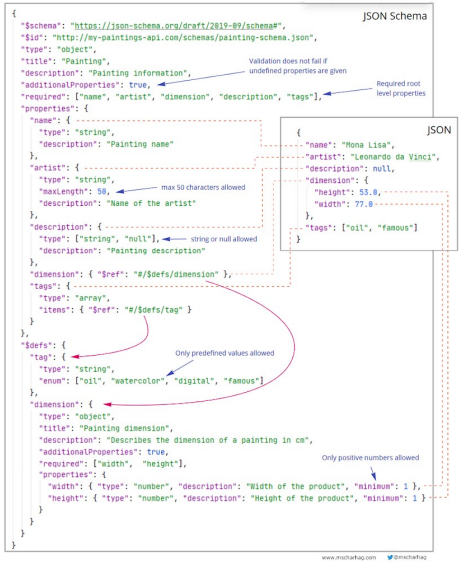
- https://json-schema.org/learn/miscellaneous-examples.html
- https://www.fiware.org/developers/data-models/
- https://oneiota.org/

FIWARE/**data-models**

Code and specifications to support harmonized data models

openconnectivityfoundation/**IoTDataModels**

Repository for data models found in www.oneIoTa.org



Data Interoperability and Semantics

< Part 3. Data schemas and semantics >

Part 3.2. Semantics

Data Interoperability and Semantics

< Part 3. Data schemas and semantics >

Part 3.2. Semantics

Part 3.2.1. Heterogeneities and data conflicts

DEMO: Compare JSON documents

https://samples.openweathermap.org/data/2.5/weather?id=2172797&appid=b6907d289e10d714a6e88b30761fae22

and

https://www.prevision-meteo.ch/services/json/lausanne

and

https://github.com/smart-data-models/dataModel.Weather/blob/master/WeatherForecast/examples/example-normalized.json

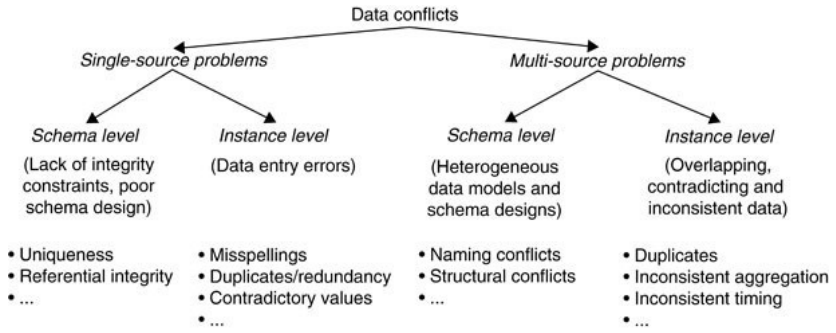
DEMO: Compare JSON documents

Different modeling choices were made, which make these two services completely non-interoperable:

- the lat/long coordinates: string vs number
- the UNIX timestamps vs dates and times
- the choice of keys and the semantics (meaning) of the values
- the units of temperature, pressure, wind speed, …
- the semantics of wind direction
- the value for "icon": "03n" - (if we follow our nose on the website, we may figure out it refers to http://openweathermap.org/img/w/03n.png)
- the country codes ISO 3166-1 ALPHA-2 and ISO 3166-1 ALPHA-3 (example of Australia and Austria)

Data conflicts

*Data conflicts are **deviations between data intended to capture the same state of a real-world entity**. Data with conflicts are often called "dirty" data and can mislead analysis performed on it.*



Do HH. (2009) Data Conflicts. In: LIU L., ÖZSU M.T. (eds) Encyclopedia of Database Systems. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-39940-9_97

Data Interoperability and Semantics

< Part 3. Data schemas and semantics >

Part 3.2. Semantics

Part 3.2.2. Controlled vocabularies, thesauri, taxonomies

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

Controlled vocabularies

… an established list of standardized terminology for use in indexing and retrieval of information

— *OECD*

… an organized arrangement of words and phrases used to index content and/or retrieve content through browsing or searching

— *Getty institute*

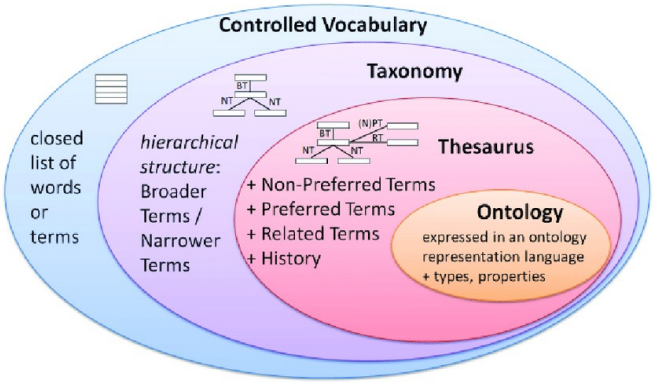
… an standardized – yet dynamic – set of terms and phrases authorized for use in an indexing system to describe a subject area or information domain

— *SCIP*

vocabulary for which the entries, i.e. **definition/term** pairs, are controlled by a **Source Authority** based on a **rulebase** and **process** for addition/deletion of entries

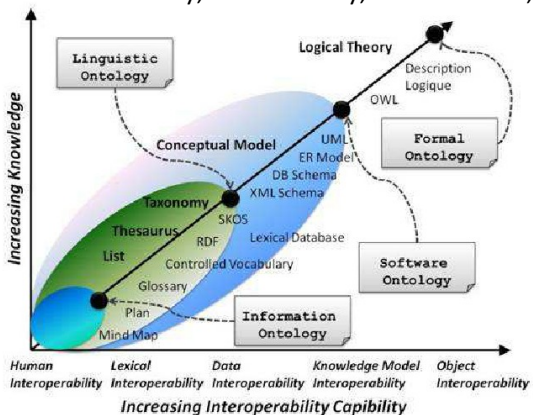
— *ISO/IEC 15944-5:2008(en)*
Information technology — Business Operational View

Controlled vocabulary, taxonomy, thesaurus, ontology



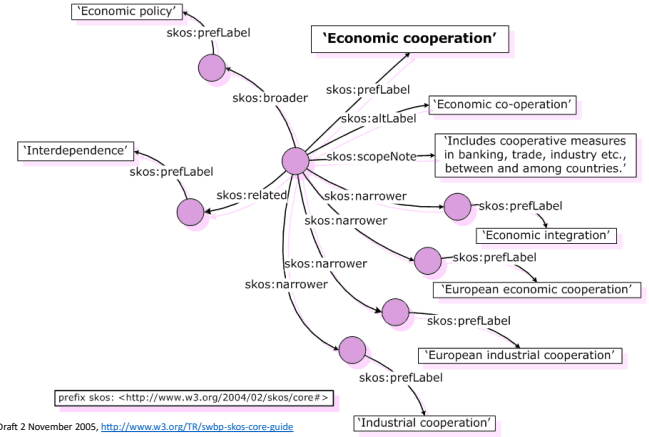
Kopácsi, Sándor & Hudak, Rastislav & Ganguly, Raman. (2017). Implementation of a Classification Server to Support Metadata Organization for Long Term Preservation Systems. Mitteilungen der Vereinigung Österreichischer Bibliothekarinnen und Bibliothekare. 70. 225. 10.31263/voebm.v70i2.1897.

Controlled vocabulary, taxonomy, thesaurus, ontology



Roussey, Catherine & Pinet, François & Kang, Myoung-Ah & Corcho, Oscar & Falquet, Gilles & Métral, Claudine & Teller, Jacques & Tweed, Christopher. (2011). Ontologies for Interoperability. 10.1007/978-0-85729-724-2_3.

SKOS Simple Knowledge Organization System



SKOS Core Guide, W3C Working Draft 2 November 2005, http://www.w3.org/TR/swbp-skos-core-guide

Examples of thesauri, taxonomies, …

UNESCO Thesaurus

The UNESCO Thesaurus is **a controlled and structured list of terms used in subject analysis and retrieval of documents and publications** in the fields of education, culture, natural sciences, social and human sciences, communication and information.

https://skos.um.es › unescothes

UNESCO Thesaurus - SKOS

GEMET General Multilingual Environmental Thesaurus

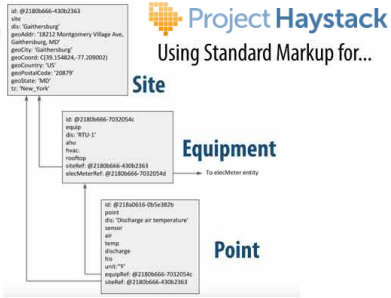
https://www.eionet.europa.eu/gemet/en/webservices/

nasa/**dictionaries**

A collection of NASA "dictionaries", including thesauri, taxonomies and ontologies.

2 Contributors · 2 Issues · 51 Stars · 31 Forks

https://github.com/nasa/dictionaries



https://project-haystack.org/

List at https://www.w3.org/2001/sw/wiki/SKOS/Datasets

Data Interoperability and Semantics

< Part 3. Data schemas and semantics >

Part 3.2. Semantics

Part 3.2.3. Resource Description Framework

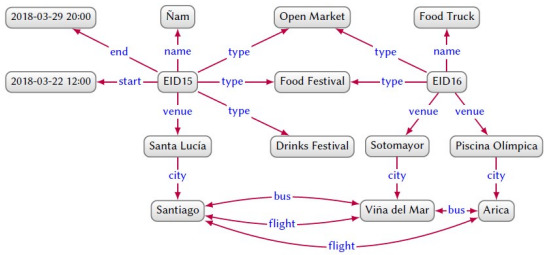
ICM – Toolbox Engineering and Interoperability of Software Systems – Course unit on Data Interoperability and Semantics
M1 Cyber Physical and Social Systems – Course unit on Data Interoperability and Semantics
Maxime Lefrançois https://maxime-lefrancois.info
Course unit URL: https://ci.mines-stetienne.fr/cps2/course/data

RDF- Resource Description Framework

RDF

Resource Description Framework

Abbreviation	RDF
Status	Published, W3C Recommendation
Year started	1996; 25 years ago
First published	February 28, 1996; 25 years ago
Latest version	1.1 (Recommendation) February 25, 2014; 7 years ago
Organization	World Wide Web Consortium (W3C)
Committee	RDF Working Group
Editors	Richard Cyganiak, David Wood, Markus Lanthaler
Base standards	URI
Related standards	RDFS, OWL, RIF, RDFa
Domain	Semantic Web
Website	www.w3.org/TR/rdf11-concepts/



An RDF Graph

Vertices:

Resources: Things. Uniquely identified by URIs

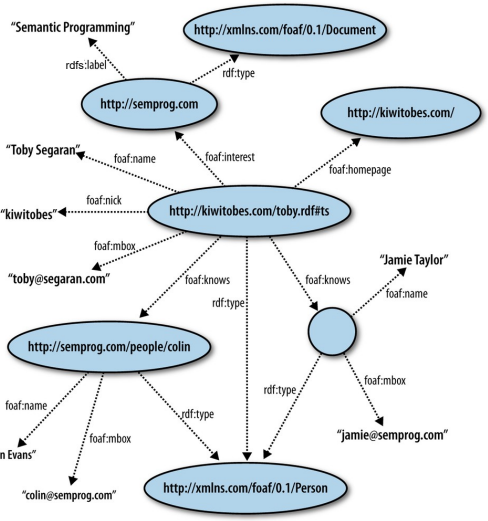
Literals: Unicode string + Datatype identified by URI. *out-degree=0*

Edges:

Relationships: connections between vertices. Uniquely identified by URIs

CURIE = Compact URIs

```
@prefix rdf:
   <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
@prefix rdfs:
   <http://www.w3.org/2000/01/rdf-schema#>.
@prefix foaf:
   <http://xmlns.com/foaf/0.1/>.
@prefix schema:
   <http://schema.org/>.
...
```



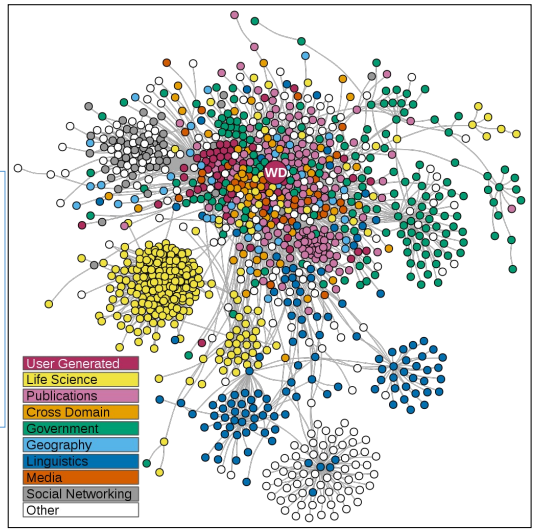
The Linked Data

It's possible to use URIs defined in an external graph

- to use a reference identification system
- to augment a graph
  anyone can say anything about anything

One can specify vertice co-reference with some special edges: owl:sameAs from W3C OWL vocabulary

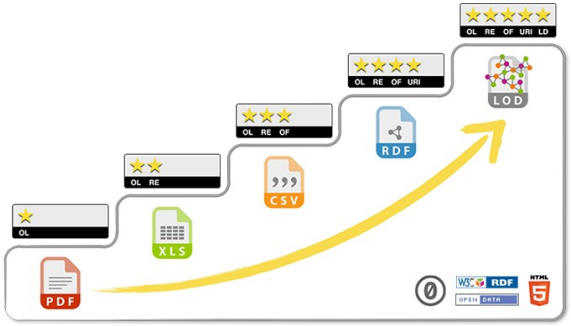
chile:Santiago → owl:sameAs → geo:SantiagoDeChile



Linked Open Data cloud visualisation: each node is a RDF dataset. Links indicate the existence of external identification links. Source: wikipedia

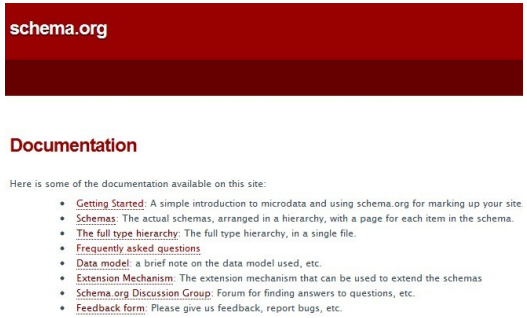
5 ★ OPEN DATA

Tim Berners-Lee, the inventor of the Web and Linked Data initiator, suggested a 5-star deployment scheme for Open Data.



https://5stardata.info/en/

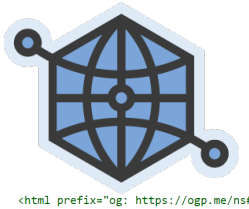
Examples of RDF vocabularies & ontologies



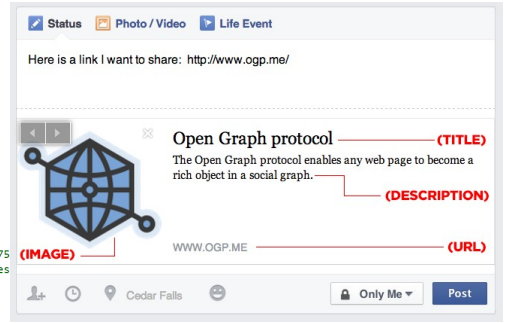
Over 10 million sites use Schema.org to markup their web pages and email messages. Many applications from Google, Microsoft, Pinterest, Yandex and others already use these vocabularies to power rich, extensible experiences.

Founded by Google, Microsoft, Yahoo and Yandex, 2011
Test with https://validator.schema.org/ or https://developers.google.com/search/docs/advanced/structured-data

Examples of RDF vocabularies & ontologies



```
<html prefix="og: https://ogp.me/ns#">
<head>
<title>The Rock (1996)</title>
<meta property="og:title" content="The Rock" />
<meta property="og:type" content="video.movie" />
<meta property="og:url" content="https://www.imdb.com/title/tt01175
<meta property="og:image" content="https://ia.media-imdb.com/images
...
</head>
...
</html>
```



Open Graph Protocol https://ogp.me/ (Facebook, 2010), test with https://developers.facebook.com/tools/debug/

Data Interoperability and Semantics

< Part 3. Data schemas and semantics >

Part 3.2. Semantics

Part 3.2.4. RDFa: Rich structured data markup for web documents

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

RDFa Resource Description Framework in Attributes

See RDFa 1.1 Primer - Rich Structured Data Markup for Web Documents - https://www.w3.org/TR/rdfa-primer/

```
<html prefix="og: https://ogp.me/ns#">
<head>
<title>The Rock (1996)</title>
<meta property="og:title" content="The Rock" />
<meta property="og:type" content="video.movie" />
<meta property="og:url" content="https://www.imdb.com/title/tt0117500/" />
<meta property="og:image" content="https://ia.media-imdb.com/images/rock.jpg" />
...
</head>
...
</html>
```

```
<div typeof="foaf:Person" xmlns:foaf="http://xmlns.com/
foaf/0.1/">
   <p property="foaf:name">
     Alice Birpemswick
   </p>
   <p>
     Email: <a rel="foaf:mbox"
href="mailto:alice@example.com">alice@example.com</a>
   </p>
   <p>
     Phone: <a rel="foaf:phone" href="tel:
+1-617-555-7332">+1 617.555.7332</a>
   </p>
</div>
```

RDFa Resource Description Framework in Attributes

See RDFa 1.1 Primer - Rich Structured Data Markup for Web Documents - https://www.w3.org/TR/rdfa-primer/

The Goal

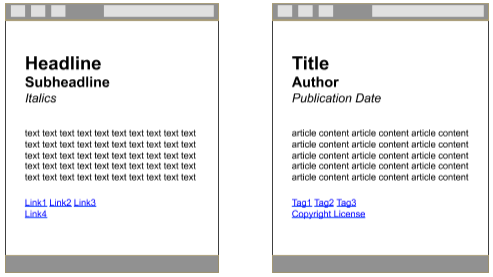
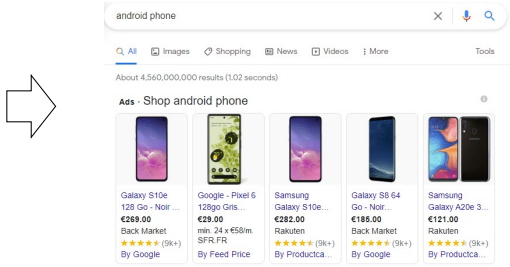


Figure 1: On the left, what browsers see. On the right, what humans see. Can we bridge the gap so that browsers see more of what we see?

RDFa Resource Description Framework in Attributes

See RDFa 1.1 Primer - Rich Structured Data Markup for Web Documents - https://www.w3.org/TR/rdfa-primer/

The Result



Google can extract structured data from e.g., e-commerce websites, and "understand" its meaning

Data Interoperability and Semantics

< Part 3. Data schemas and semantics >

Part 3.2. Semantics

Part 3.2.5. JSON-LD: JSON for Linking Data

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 for Linked Data

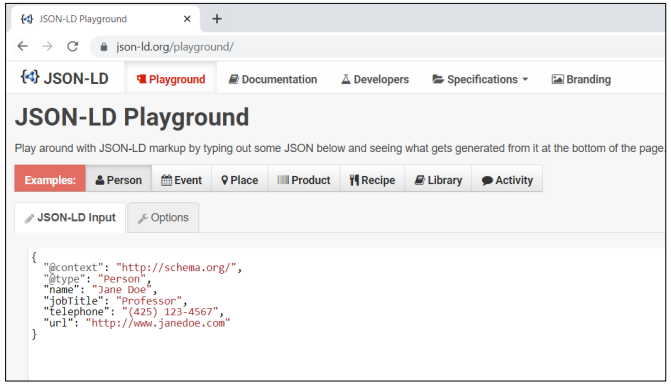
See JSON-LD 1.1 - A JSON-based Serialization for Linked Data - https://www.w3.org/TR/json-ld11/

JSON-LD

Filename extension	.jsonld
Internet media type	application/ld+json
Type of format	Semantic Web
Container for	Linked Data
Extended from	JSON
Standard	JSON-LD 1.1 / JSON-LD 1.1 API
Open format?	Yes

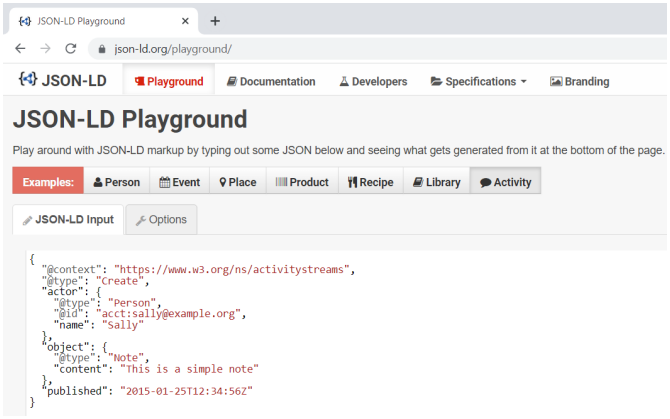
https://json-ld.org/learn.html

JavaScript Object Notation for Linked Data



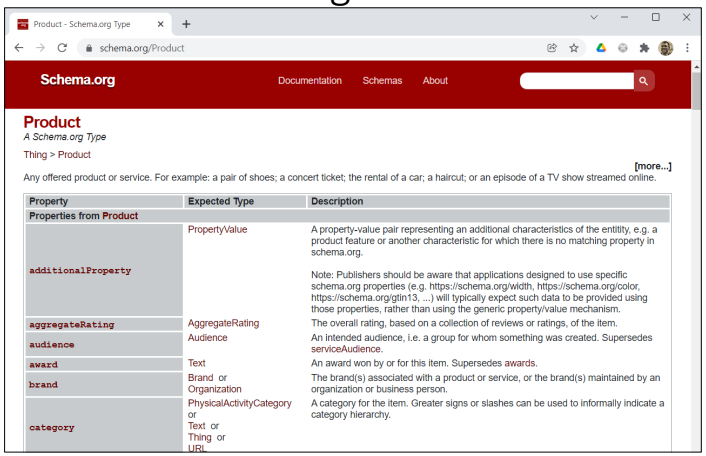
https://json-ld.org/playground/

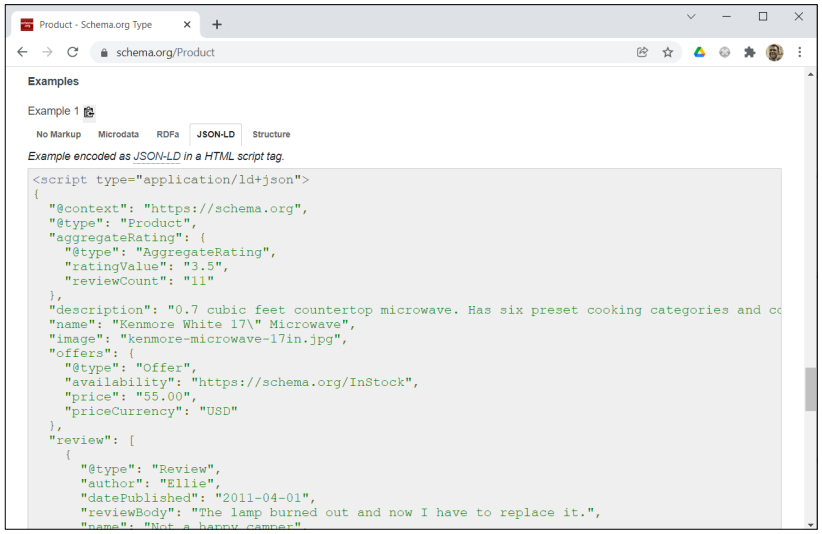
JavaScript Object Notation for Linked Data



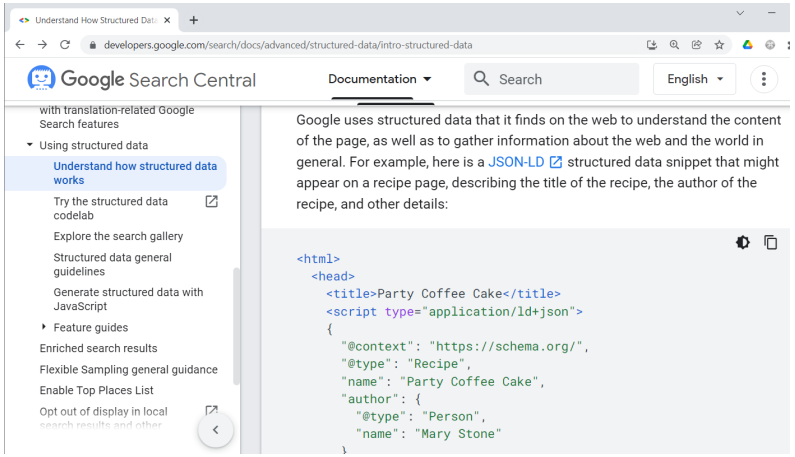
https://json-ld.org/playground/

JSON-LD and schema.org

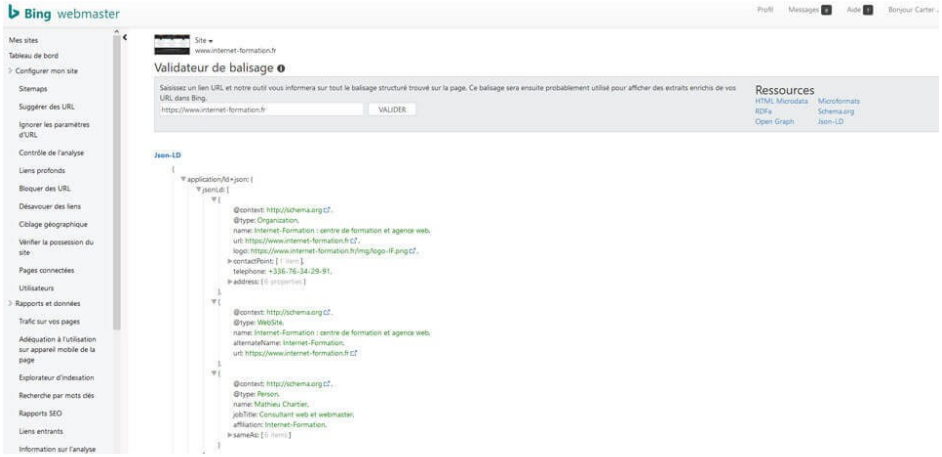




JSON-LD and Google



JSON-LD and Bing



Data Interoperability and Semantics

</ Part 3. Data schemas and semantics >

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

< Part 4. The value of data >

Data Interoperability and Semantics
Outline

- < Part 4. The value of data >
  - Part 4.1. Value as one of the V's of Big Data
  - Part 4.2. Data, Information, Knowledge, Metadata, …
  - Part 4.3. Interoperability unlocks the value of data
  - Part 4.4. Open data generates economic and societal value
  - Part 4.5. Machine-actionability of data increases its value

Data Interoperability and Semantics

Part 4. The value of data

Part 4.1. Value as one of the V's of Big Data

Orders of magnitude of the total amount of data created, captured, copied, and consumed globally ?

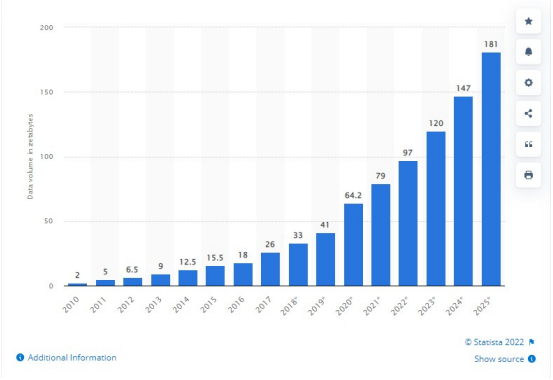
Multiple-byte units V·T·E

Decimal			Binary		
Value	Metric		Value	IEC	Legacy
1000	kB	kilobyte	1024	KiB kibibyte	KB kilobyte
1000^2	MB	megabyte	1024^2	MiB mebibyte	MB megabyte
1000^3	GB	gigabyte	1024^3	GiB gibibyte	GB gigabyte
1000^4	TB	terabyte	1024^4	TiB tebibyte	TB terabyte
1000^5	PB	petabyte	1024^5	PiB pebibyte	–
1000^6	EB	exabyte	1024^6	EiB exbibyte	–
1000^7	ZB	zettabyte	1024^7	ZiB zebibyte	–
1000^8	YB	yottabyte	1024^8	YiB yobibyte	–

Orders of magnitude of data

Projected figures 2025 for the total amount of data created, captured, copied, and consumed globally

530%
increase of global data volume
From 33 zettabytes in 2018 to 175 zettabytes



https://www.statista.com/statistics/871513/worldwide-data-created/

https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/european-data-strategy_en#projected-figures-2025

Projected figures 2025

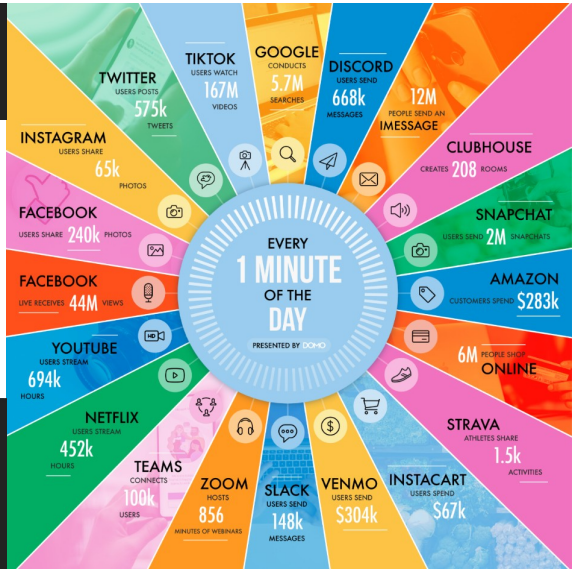
530%
increase of global data volume
From 33 zettabytes in 2018 to 175 zettabytes

€829 billion
value of data economy in the EU27
From €301 billion (2.4% of EU GDP) in 2018

10.9 million
data professionals in the EU27
From 5.7 million in 2018

65%
Percentage of EU population with basic digital skills
From 57% in 2018

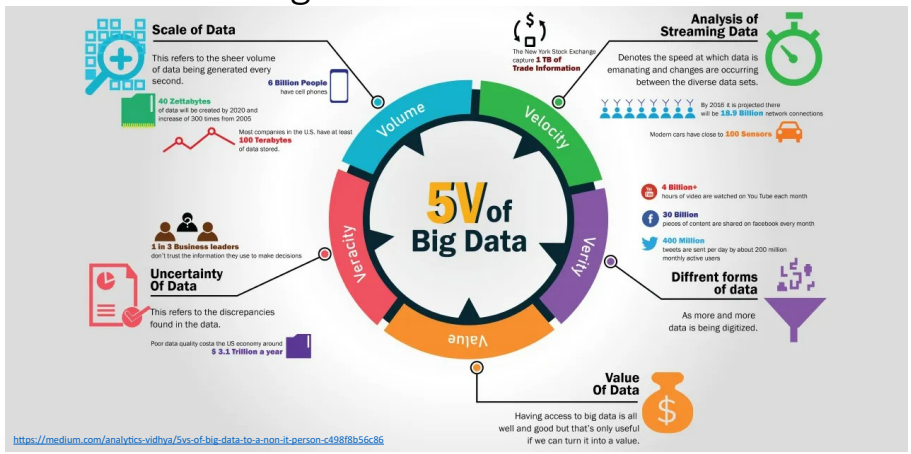
https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/european-data-strategy_en#projected-figures-2025



This happens every minute of every day, according to DOMO
(Source: Data Never Sleeps 9.0)

https://www.domo.com/learn/infographic/data-never-sleeps-9

The 5 V's of Big Data ?



https://medium.com/analytics-vidhya/5vs-of-big-data-to-a-non-it-person-c498f8b56c86

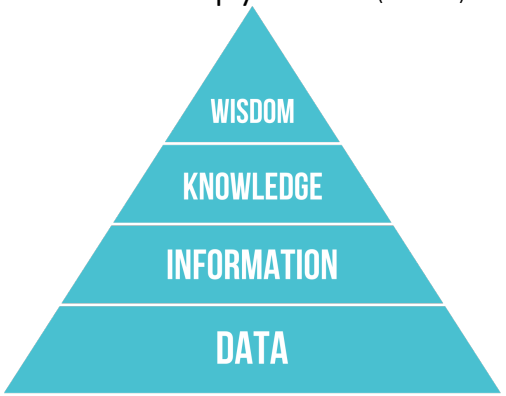
Data Interoperability and Semantics

Part 4. The value of data

Part 4.2. Data, Information, Knowledge, Metadata, …

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

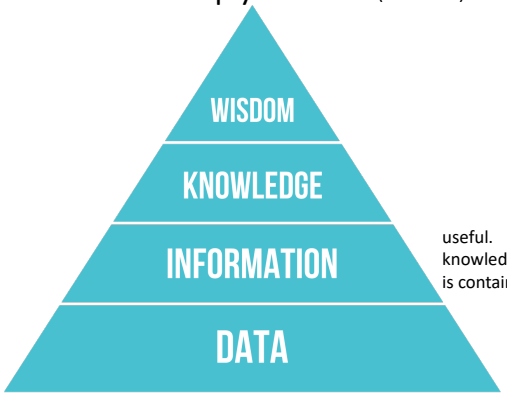
The DIKW pyramid (~1982)



useless.
symbols or signs, representing stimuli or signals, that are "of no use until...in a usable (that is, relevant) form

https://en.wikipedia.org/wiki/DIKW_pyramid , and chosen definitions

The DIKW pyramid (~1982)

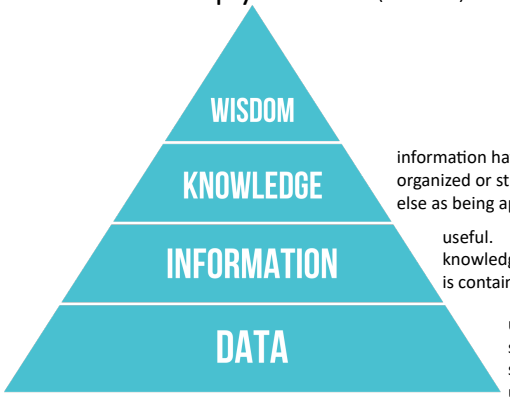


useful.
knowledge by description ("information is contained in descriptions")

useless.
symbols or signs, representing stimuli or signals, that are "of no use until...in a usable (that is, relevant) form

https://en.wikipedia.org/wiki/DIKW_pyramid , and chosen definitions

The DIKW pyramid (~1982)



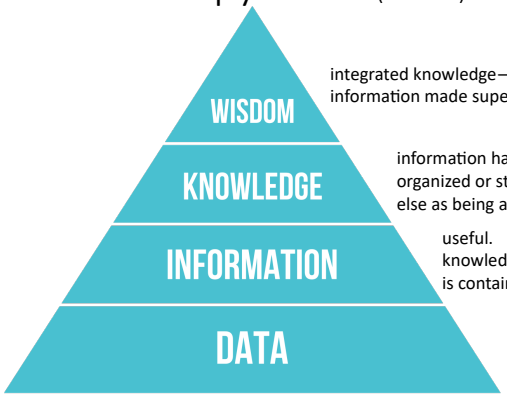
information having been processed, organized or structured in some way, or else as being applied or put into action

useful.
knowledge by description ("information is contained in descriptions")

useless.
symbols or signs, representing stimuli or signals, that are "of no use until...in a usable (that is, relevant) form

https://en.wikipedia.org/wiki/DIKW_pyramid , and chosen definitions

The DIKW pyramid (~1982)



integrated knowledge—
information made super-useful

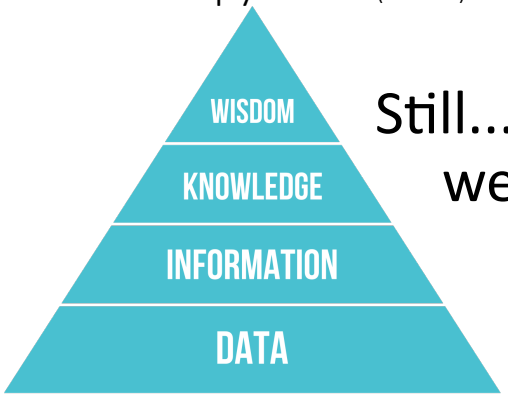
information having been processed, organized or structured in some way, or else as being applied or put into action

useful.
knowledge by description ("information is contained in descriptions")

useless.
symbols or signs, representing stimuli or signals, that are "of no use until...in a usable (that is, relevant) form

https://en.wikipedia.org/wiki/DIKW_pyramid , and chosen definitions

The DIKW pyramid (~1982)



Still...
we often use
« Data »
to generalize

Definitions on Data

— ISO/IEC20546:2019 (Big data – Overview and vocabulary)

Dataset
Identifiable collection of data available for access or download in one or more formats

Data
Re-interpretable representation of information in a formalized manner suitable for communication, interpretation, or processing
Note 1 to entry: Data can be processed by humans or by automatic means.

Metadata
Data about data or data elements, possibly including their data descriptions and data about data ownership, access paths, access rights and data volatility

Information
Data that are processed, organised and correlated to produce meaning.
Note 1 to entry: Information concerns facts, concepts, objects, events, ideas, processes, etc.

Metadata

"data that provides information about other data"

- **Descriptive metadata** – the descriptive information about a resource.
  - For discovery and identification.
  - Ex: title, abstract, author, keywords.
- **Structural metadata** – containers of data, how compound objects are put together
  - Ex, how pages are ordered to form chapters.
  - Ex: types, versions, relationships, and other characteristics of digital materials.
- **Administrative metadata** – the information to help manage a resource
  - Ex: resource type, permissions, time, when and how it was created.
- **Reference metadata** – contents and quality
  - For quality assessment of the data
  - Ex: conceptual metadata, quality metadata, methodological metadata.
- **Statistical metadata**, also called process data,
  - May describe processes that collect, process, or produce statistical data.
  - Number of rows, columns, etc.
- **Legal metadata** –
  - Ex: license, creator, copyright holder

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

Data Interoperability and Semantics

Part 4. The value of data

Part 4.3. Interoperability unlocks the value of data

Data Analytics vs Data Processing

— ISO/IEC20546:2019 (Big data – Overview and vocabulary)

Data Analytics
Composite concept consisting of data acquisition, data collection, data validation, data processing, including data quantification, data visualisation and data interpretation

Data Processing
Systematic performance of operations upon data
Note 1 to entry: Example: Arithmetic or logic operations upon data, merging or sorting of data, or operations on text, such as editing, sorting, merging, storing, retrieving, displaying, or printing.

Data silos problem illustrated

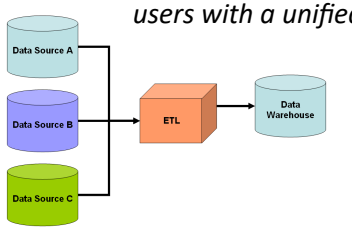
Repositories of fixed data that are isolated, incompatible, or not integrated

Example of silos https://factory.dev/pimcore-knowledge-base/whitepaper/data-silos

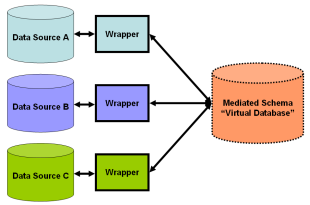
Data integration

*combining **heterogeneous data** and providing users with a unified view of them.*

Simple schematic for a data warehouse.
The Extract, transform, load (ETL) process extracts information from the source databases, transforms it and then loads it into the data warehouse.



Simple schematic for a data-integration solution.
A system designer constructs a mediated schema against which users can run queries. The virtual database interfaces with the source databases via wrapper code if required.



Sub-areas:
- Data warehousing
- Data migration
- Enterprise application/information integration
- Master data management

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

Interoperability vs Portability

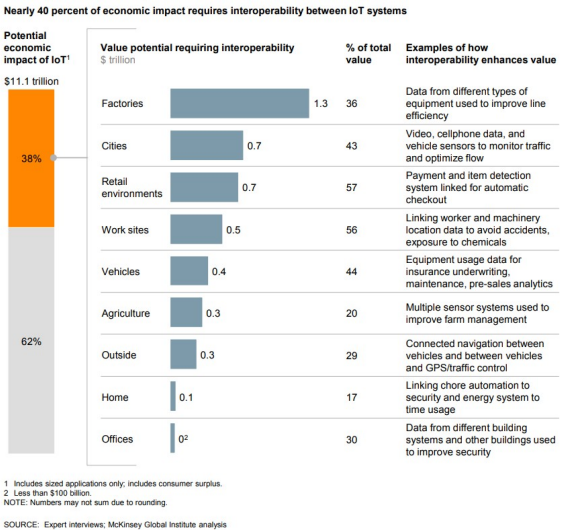
— IEEE Standard Computer Dictionary
— ISO/IEC19941:2017 (Cloud computing – interoperability and portability)

Interoperability
Ability of two or more systems or components to exchange information and to mutually use the information that has been exchanged

Portability
Ability to easily transfer data from one system to another without being required to re-enter data

Interoperability as an enabler for the potential value of data

ex:
Internet of Things



https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/the-internet-of-things-the-value-of-digitizing-the-physical-world

Interoperability at different levels: from the hardware to policies

- **Policies**: Laws, regulations, norms, ...
- **Behaviour**: Processus, services, operations, ...
- **Semantic**: Knowledge domain vocabularies, taxonomies, ontologies, ...
- **Syntactic**: Encoding, Formats, Schemas, ...
- **Transport:** see also Open Systems Interconnections (OSI) network layers
- **OS:** Kernel, File systems,...
- **Hardware**: Instruction Set Architectures, ...

ISO/IEC definitions of interoperability

Transport interoperability
interoperability where information exchange uses an established communication infrastructure between the participating systems
— ISO/IEC 22123-1:2021, Cloud computing — Part 4: Vocabulary

Syntactic interoperability
interoperability such that the formats of the exchanged information can be understood by the participating systems
— ISO/IEC 22123-1:2021, Cloud computing — Part 4: Vocabulary

Semantic data interoperability
interoperability so that the meaning of the data model within the context of a subject area is understood by the participating systems
— ISO/IEC 22123-1:2021, Cloud computing — Part 4: Vocabulary

Behavioural interoperability
interoperability so that the actual result of the exchange achieves the expected outcome
— ISO/IEC 22123-1:2021, Cloud computing — Part 4: Vocabulary

Policy interoperability
interoperability while complying with the legal, organizational, and policy frameworks applicable to the participating systems
— ISO/IEC 22123-1:2021, Cloud computing — Part 4: Vocabulary

Data Interoperability – a definition

Data interoperability addresses the ability of systems and services that create, exchange and consume data to have clear, shared expectations for the contents, context and meaning of that data.

— https://datainteroperability.org/

Syntactic interoperability
interoperability such that the formats of the exchanged information can be understood by the participating systems

— ISO/IEC 22123-1:2021, Cloud computing — Part 4: Vocabulary

Semantic data interoperability
interoperability so that the meaning of the data model within the context of a subject area is understood by the participating systems

— ISO/IEC 22123-1:2021, Cloud computing — Part 4: Vocabulary

Norm-based interoperability

- Hardware and/or software standards
- Standard = Detailed set of technical requirements intended to establish a certain uniformity (in a field of hardware or software development)
- Standard ---- Norm: no clear boundary
  - docx: Microsoft
  - HTML, XML: World Wide Web consortium (W3C)
  - SIM cards: European Telecommunication Standard Institute (ETSI)
  - Postscript became standardized after release
  - De facto standards (not produced by a standard development organization). example: csv, json, rar, pdf, java, flash

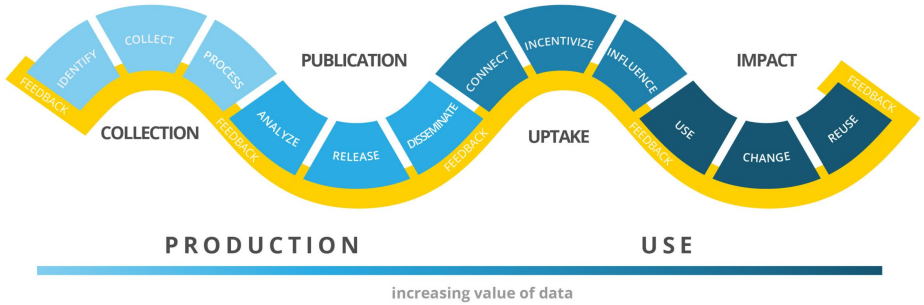
Data Interoperability and Semantics

Part 4. The value of data

Part 4.4. Sharing data generates economic and societal value

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

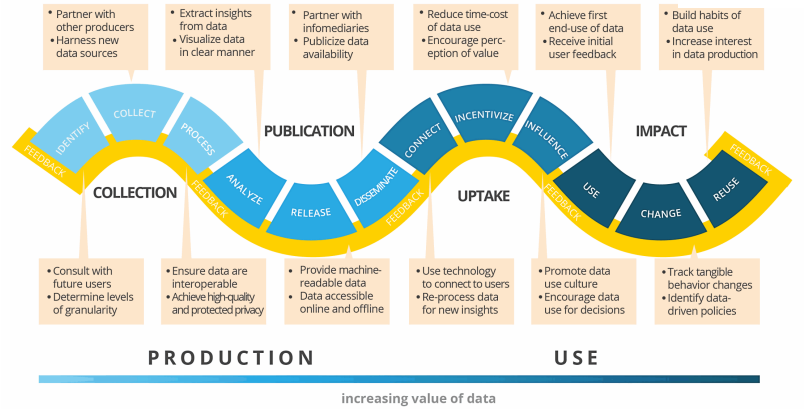
The Data Value chain



Open Data Watch. "The data value chain: Moving from production to impact." *Data2X.* *https://opendatawatch.com/publications/the-data-value-chain-moving-from-production-to-impact* (2018).

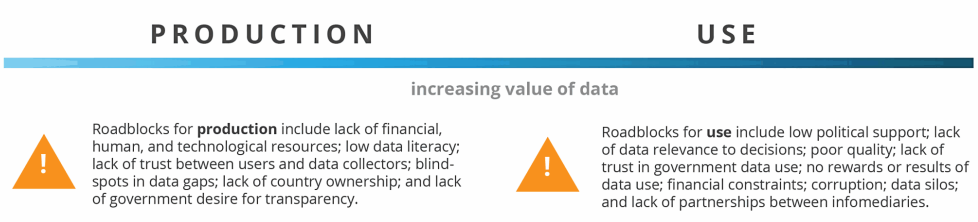
MARKERS — Potential achievements within each process of the value chain mark progress towards data impacts.

The Data Value chain



Open Data Watch. "The data value chain: Moving from production to impact." *Data2X*. https://opendatawatch.com/publications/the-data-value-chain-moving-from-production-to-impact (2018).

The Data Value roadblocks



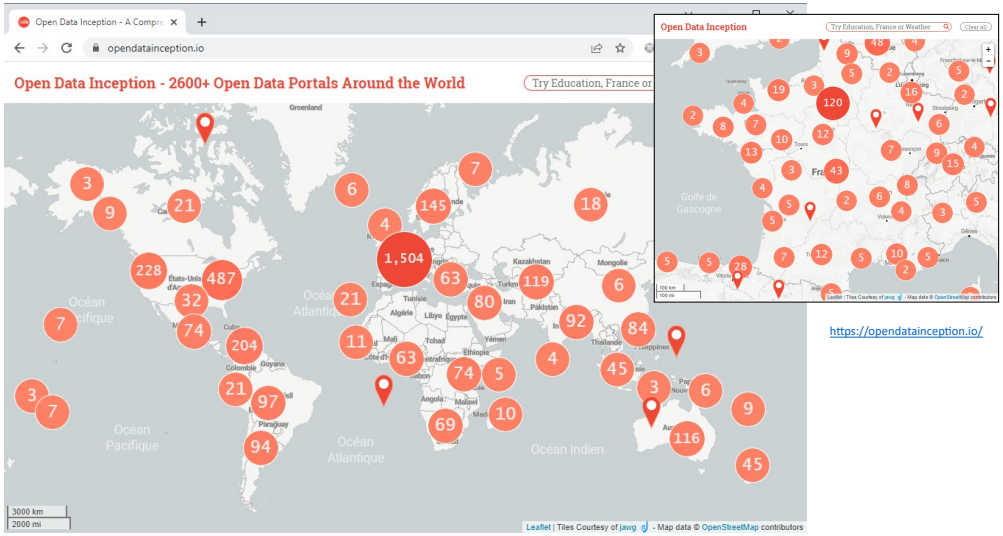
Open Data Watch. "The data value chain: Moving from production to impact." *Data2X*. https://opendatawatch.com/publications/the-data-value-chain-moving-from-production-to-impact (2018).

Open Data in the US

- Data.gov 2009
- Legal framework:
  - The U.S. Open Government Directive of December 8, 2009, required that all agencies post **at least three high-value data sets online** and register them on Data.gov **within 45 days**
  - OPEN Government Data Act, as part of the Foundations for Evidence Based Policymaking Act (2019)

Open Data in France

- France at the forefront of Open Data in Europe:
  - Légifrance 1999
- Legal framework:
  - "The society has the right of requesting account from any public agent of its administration." (Declaration of rights of man and of the citizen of 1789)
  - Law on the liberty of access to administrative documents (1978)
  - Euopean directove 2003 + French Law 2005 + Decree 2011
  - Bill on a Digital Republic (2016)
  - The law on Energy Transition (2015)
- 2014: Chief Data Officer in the French public administration



https://opendatainception.io/

Open Data in Europe

EUROPEAN DATA PORTAL

The European Data Portal:
Opening up Europe's public data

data.europa.eu
The official portal for European data

Open Data, Open Content, and Open Knowledge

The Open Definition

The Open Definition sets out principles that define "openness" in relation to **data and content**.

It makes **precise** the meaning of "open" in the terms **"open data"** and **"open content"** and thereby ensures **quality** and encourages **compatibility** between different pools of open material.

It can be summed up in the statement that:

> "Open means **anyone** can **freely access, use, modify, and share** for **any purpose** (subject, at most, to requirements that preserve provenance and openness)."

Put most succinctly:

> "Open data and content can be **freely used, modified, and shared** by **anyone** for **any purpose**"

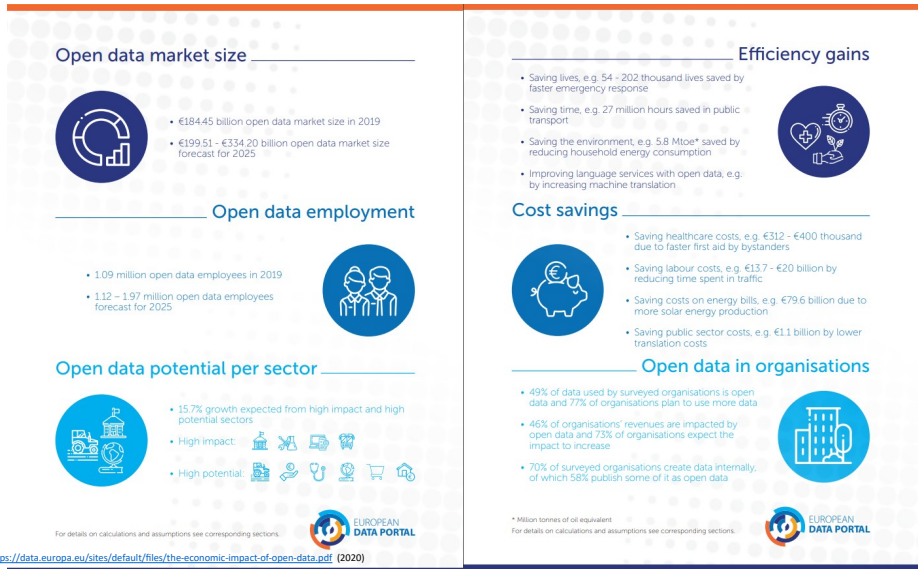
http://opendefinition.org/

Open Work

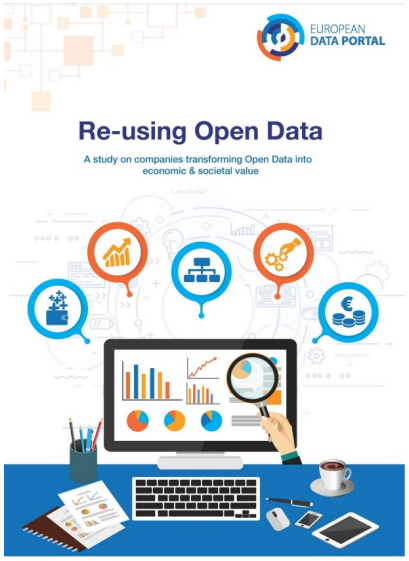
1. Open License or Status
   The **work** *must* be in the **public domain** or provided under an open **license**
2. Access
   The **work** *must* be provided as a whole and at no more than a reasonable one-time reproduction cost, and *should* be downloadable via the Internet without charge.
3. Machine Readability
   The **work** *must* be provided in a form readily processable by a computer and where the individual elements of the work can be easily accessed and modified.
4. Open Format
   The **work** *must* be provided in an open format.

http://opendefinition.org/

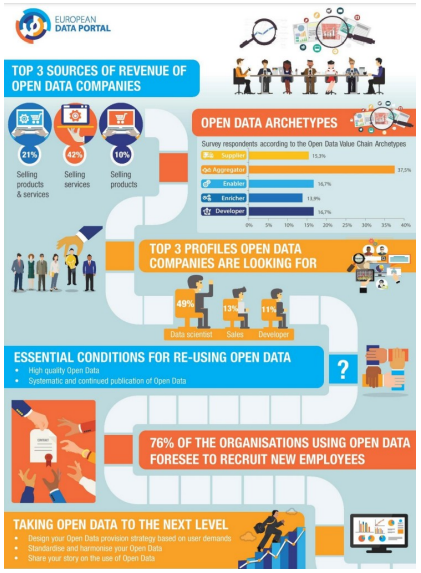
Economic impact of Open Data

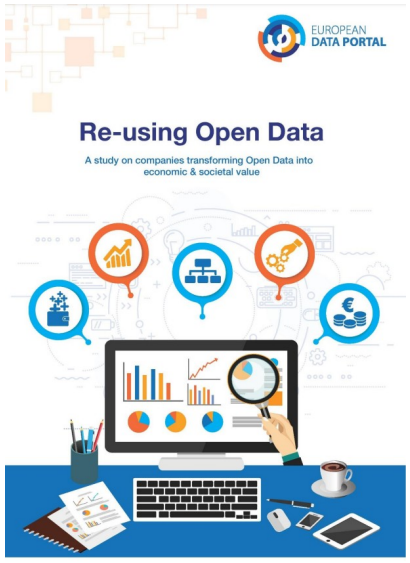


https://data.europa.eu/sites/default/files/the-economic-impact-of-open-data.pdf (2020)

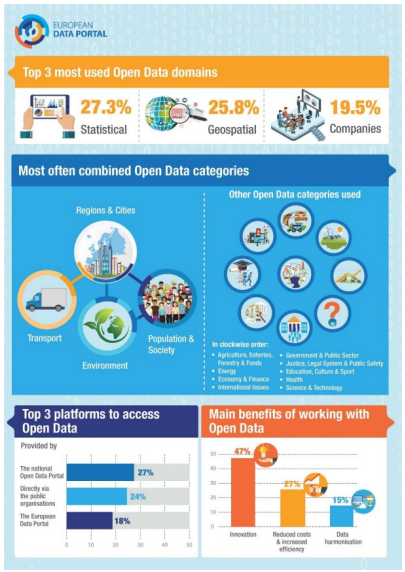


https://data.europa.eu/sites/default/files/re-using_open_data.pdf (2020)





https://data.europa.eu/sites/default/files/re-using_open_data.pdf (2020)



Data Interoperability and Semantics

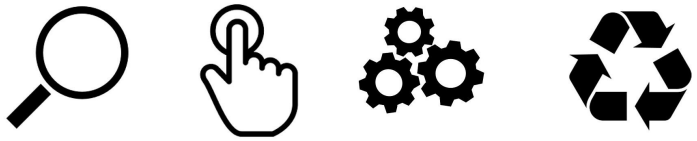
Part 4. The value of data

Part 4.5. Machine-actionability of data increases its value

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

FAIR Principles

Findable **A**ccessible **I**nteroperable **R**eusable



The FAIR principles emphasize machine-actionability (i.e., the capacity of computational systems to find, access, interoperate, and reuse data with none or minimal human intervention) because humans increasingly rely on computational support to deal with data as a result of the increase in volume, complexity, and creation speed of data.

https://www.go-fair.org/fair-principles/

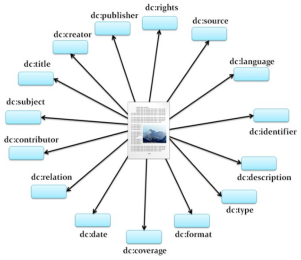
FAIR Guiding Principles for scientific data management and stewardship (2016)

Source : Australian National Data Service (ANDS)

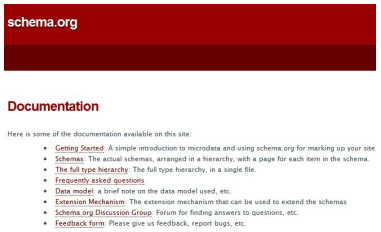
https://www.go-fair.org/fair-principles/

FAIR Guiding Principles for scientific data management and stewardship (2016)

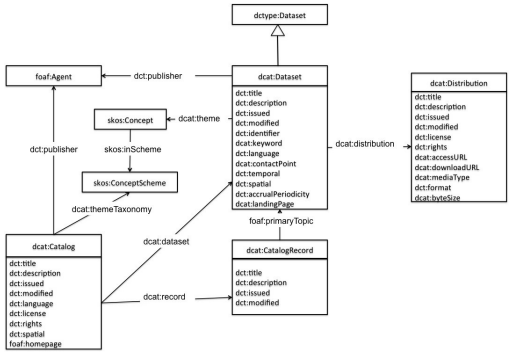
Example of standards for Metadata



Dublin Core



schema.org



Data Catalog Vocabulary (DCAT) - Version 2
W3C Recommendation 04 February 2020

- Demo Dublin core or schema.org in web pages
- Demo DCAT for open data portal catalogs

Data Interoperability and Semantics

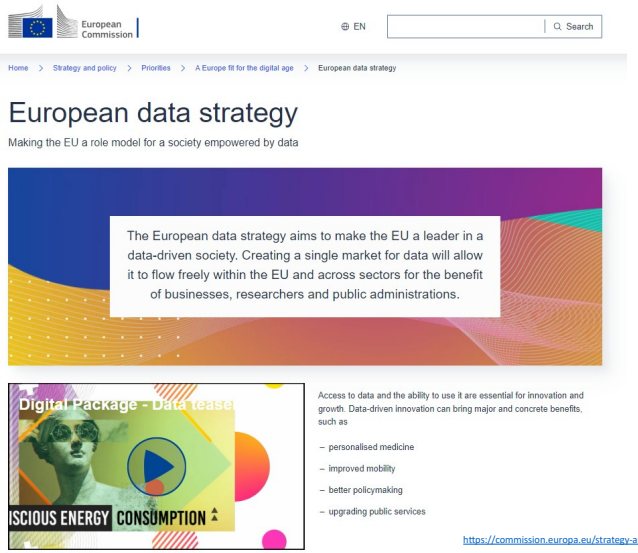
< Part 4. The value of data >

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

< Part 5. The European Data Strategy >

A guided tour !



1)

https://ec.europa.eu/commission/presscorner/detail/en/fs_20_283

2)

https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/european-data-strategy

Data Interoperability and Semantics
Outline

- < Part 5. The European Data Strategy >
  - Part 5.1. The European legislation on open data
  - Part 5.2. The General Data Protection Regulation (GDPR)
  - Part 5.3. The Data Governance Act (DGA)
  - Part 5.4. The Data Act
  - Part 5.5. Data Spaces

Data Interoperability and Semantics

Part 5. The value of data

Part 5.1. The European legislation on open data

European Commission

Home > Strategy > Priorities 2019-2024 > A Europe fit for the digital age > European data strategy

European legislation on open data

(adopted 05/2022 – applicable 09/2023)

The Directive on open data and the re-use of public sector information provides common rules for a European market for government-held data.

The **"Open Data Directive"** (EU) 2019/1024 entered into force on 16 July 2019

It replaced the **Public Sector Information (PSI) Directive** of 2003.

EU countries had to transpose Directive (EU) 2019/1024 by 16 July 2021.

The **Commission Implementing Regulation (EU) 2023/138**
adopted a list of specific high-value datasets by way of an implementing act.

1)

https://digital-strategy.ec.europa.eu/en/policies/legislation-open-data

European Commission

Home > Strategy > Priorities 2019-2024 > A Europe fit for the digital age > European data strategy

European legislation on open data

(adopted 05/2022 – applicable 09/2023)

20.1.2023 EN Official Journal of the European Union L 19/43

COMMISSION IMPLEMENTING REGULATION (EU) 2023/138

of 21 December 2022

laying down a list of specific high-value datasets and the arrangements for their publication and re-use

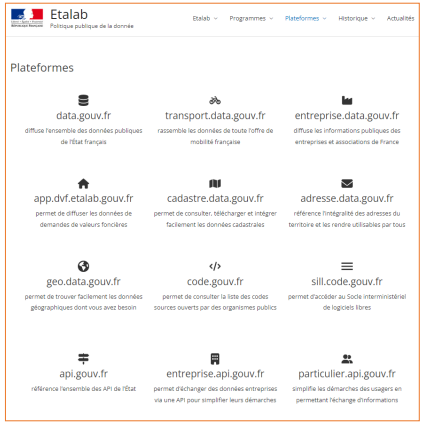
(Text with EEA relevance)

The **Commission Implementing Regulation (EU) 2023/138**
adopted a list of specific high-value datasets by way of an implementing act.

2)

http://data.europa.eu/eli/reg_impl/2023/138/oj

Actual impact (FR)



https://www.etalab.gouv.fr/

https://meteo.data.gouv.fr/

https://transport.data.gouv.fr/

Data Interoperability and Semantics

Part 5. The value of data

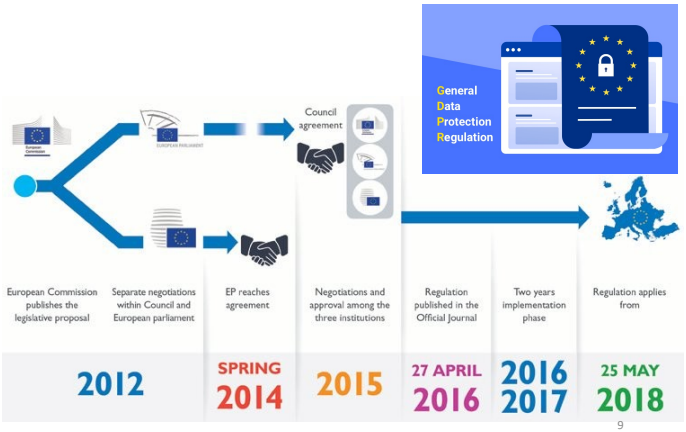
Part 5.2. The General Data Protection Regulation (GDPR)

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

General Data Protection Regulation (GDPR) (2016)

https://gdpr.eu/

- Requiring the consent of subjects for data processing
- Anonymizing collected data to protect privacy
- Providing data breach notifications
- Safely handling the transfer of data across borders
- Requiring certain companies to appoint a data protection officer to oversee GDPR compliance



General Data Protection Regulation (GDPR) (2016)

https://gdpr.eu/

Data Interoperability and Semantics

Part 5. The value of data

Part 5.3. The Data Governance Act (DGA)

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

European Commission

Home > Strategy > Priorities 2019-2024 > A Europe fit for the digital age > European data strategy

The European Data Governance Act

(adopted 05/2022 – applicable 09/2023)

1)

https://digital-strategy.ec.europa.eu/en/library/data-governance-act

European Commission

Home > Strategy > Priorities 2019-2024 > A Europe fit for the digital age > European data strategy

The European Data Governance Act

(adopted 05/2022 – applicable 09/2023)

3.6.2022 EN Official Journal of the European Union L 152/1

REGULATION (EU) 2022/868 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL

of 30 May 2022

on European data governance and amending Regulation (EU) 2018/1724 (Data Governance Act)

(Text with EEA relevance)

ttp://data.europa.eu/eli/reg/2022/868/oj

Setting up a new European way of data governance will facilitate data sharing across sectors and Member States. It will create wealth for society, and provide control to citizens and trust to companies.

2)

https://digital-strategy.ec.europa.eu/en/policies/data-governance-act-explained

Data Interoperability and Semantics

Part 5. The value of data

Part 5.4. The Data Act

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

European Commission

Home > Strategy > Priorities 2019-2024 > A Europe fit for the digital age > European data strategy

The European Data Act

(adopted 01/2024 – applicable 09/2025)

1)

https://ec.europa.eu/newsroom/dae/redirection/document/83517

European Commission

Home > Strategy > Priorities 2019-2024 > A Europe fit for the digital age > European data strategy

The European Data Act

(adopted 01/2024 – applicable 09/2025)

Official Journal of the European Union EN Series L

2023/2854 22.12.2023

REGULATION (EU) 2023/2854 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL

of 13 December 2023

on harmonised rules on fair access to and use of data and amending Regulation (EU) 2017/2394 and Directive (EU) 2020/1828 (Data Act)

(Text with EEA relevance)

ttp://data.europa.eu/eli/reg/2023/2854/oj

The Data Act will make **more data available** for use. It will set up rules on who can use and access what data for which purposes across all economic sectors in the EU.

2)

https://digital-strategy.ec.europa.eu/en/factpages/data-act-explained

European Commission

Home > Strategy > Priorities 2019-2024 > A Europe fit for the digital age > European data strategy

The European Data Act

(adopted 01/2024 – applicable 09/2025)

Chapter VIII – Interoperability

Article 33 – Essential requirements regarding interoperability of data, of data sharing mechanisms and services, as well as of common European data spaces

- Participants in data spaces that offer data or data services to other participants shall comply with the following essential requirements to facilitate the interoperability of data, of data sharing mechanisms and services, as well as of common European data spaces which are purpose- or sector-specific or cross-sectoral interoperable frameworks for common standards and practices to share:

- the dataset content, use restrictions, licences, data collection methodology, data quality and uncertainty shall be sufficiently described, where applicable, in a machine-readable format, to allow the recipient to find, access and use the data
- the data structures, data formats, vocabularies, classification schemes, taxonomies and code lists, where available, shall be described in a publicly available and consistent manner
- the technical means to access the data, such as application programming interfaces, and their terms of use and quality of service shall be sufficiently described to enable automatic access and transmission of data between parties, including continuously, in bulk download or in real-time in a machine-readable format where that is technically feasible and does not hamper the good functioning of the connected product
- where applicable, the means to enable the interoperability of tools for automating the execution of data sharing agreements, such as smart contracts shall be provided

17

Data Interoperability and Semantics

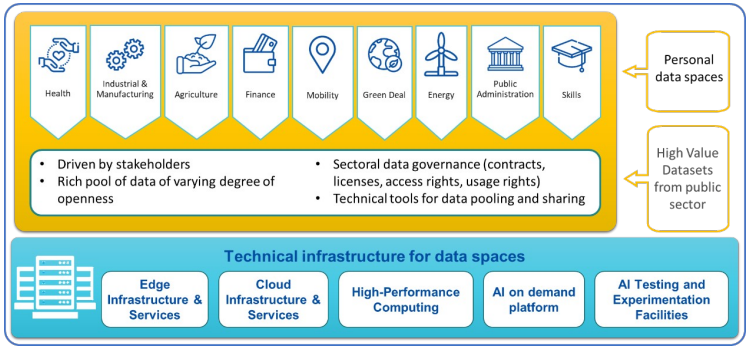
Part 5. The value of data

Part 5.5. Data Spaces

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

Common European data spaces

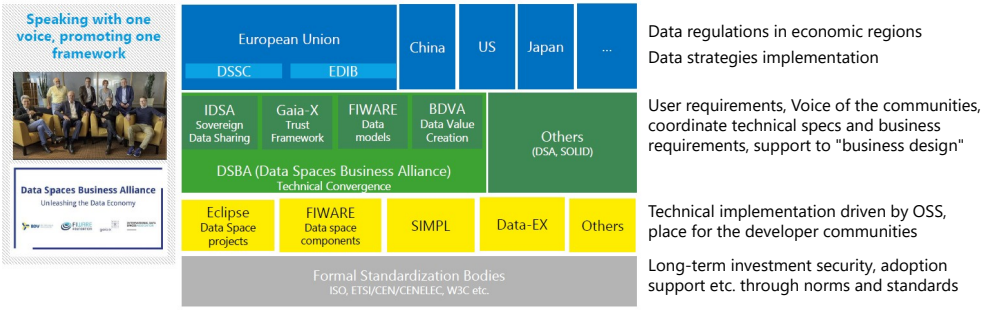
``purpose or sector specific or cross-sectoral interoperable frameworks for common standards and practices to share or jointly process data for, inter alia, the development of new products and services, scientific research or civil society initiatives.''
[source: Data Act]



https://digital-strategy.ec.europa.eu/en/library/building-data-economy-brochure

19

Regulatory, business, and technical foundation for Data Spaces within the Edge-Cloud-Continuum



Data regulations in economic regions
Data strategies implementation

User requirements, Voice of the communities, coordinate technical specs and business requirements, support to "business design"

Technical implementation driven by OSS, place for the developer communities

Long-term investment security, adoption support etc. through norms and standards

https://data-spaces-business-alliance.eu/download/33968/

Adapted from a presentation by L. Nagel at the Data Spaces Symposium 2024

20

The Data Spaces Business Alliance

Unleashing the European Data Economy

https://data-spaces-business-alliance.eu/

Accelerating Business transformation in the Data Economy

The Data Spaces Business Alliance (DSBA) accelerates business transformation in the data economy. It's the first initiative of its kind, uniting industry players to realize a data-driven future in which organizations and individuals can unlock the full value of their data.

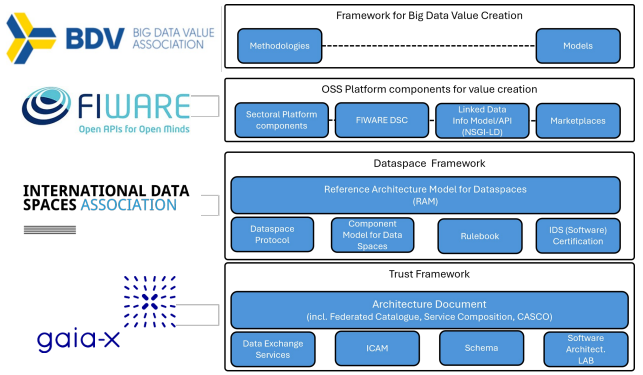
Data spaces are key to achieving sovereign, interoperable and trustworthy data-sharing across businesses and societies – a key step to the data economy of the future. The Alliance embraces this reality, converging the best skills, assets, and experience in Europe into a one-stop-shop for data spaces, from inception to deployment.

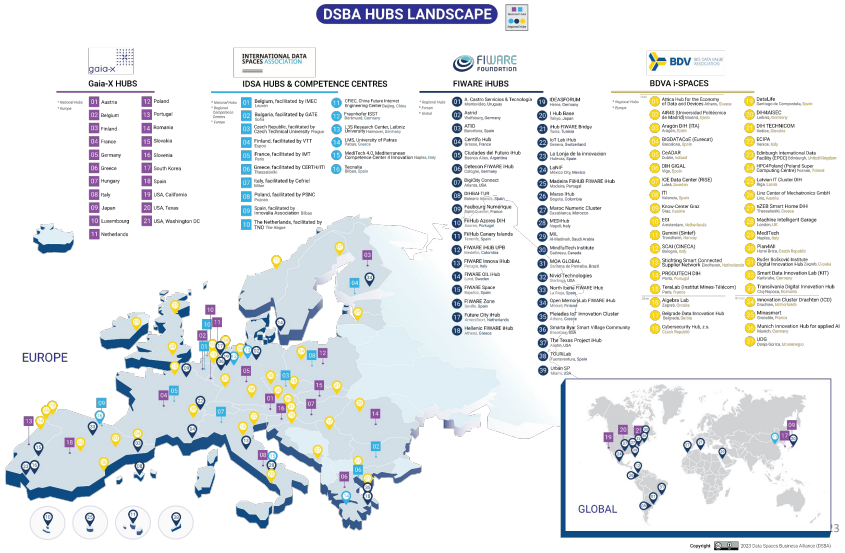
The Data Spaces Business Alliance are Gaia-X European Association for Data and Cloud AISBL, the Big Data Value Association (BDVA), FIWARE Foundation, and the International Data Spaces Association (IDSA). Together they represent 1,000+ leading key industry players, associations, research organizations, innovators, and policymakers worldwide. With this cross-industry expertise, resources and know-how, the Alliance drives awareness, evangelizes technology, shapes standards, and enables integration across industries.

The Data Spaces Business Alliance

Unleashing the European Data Economy

https://data-spaces-business-alliance.eu/





Our members are the backbone of IDSA

INTERNATIONAL DATA SPACES ASSOCIATION

860+ People contributing

31 Countries

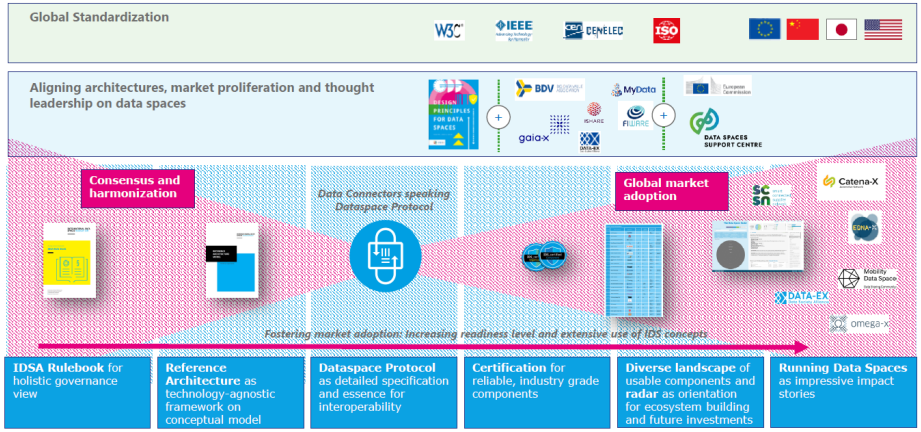
153

Adapted from a presentation by L. Nagel at the Data Spaces Symposium 2024

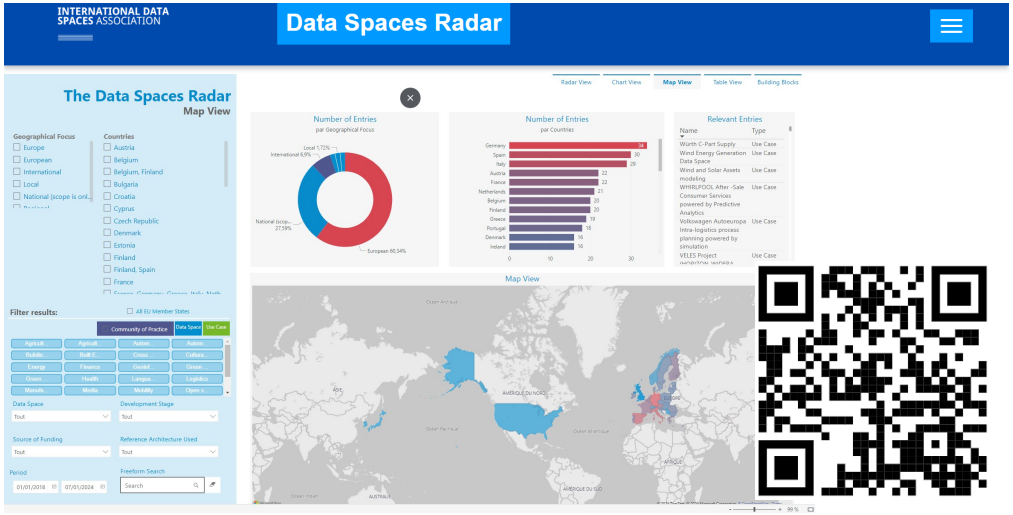
A holistic approach to bring data spaces to global scale

INTERNATIONAL DATA SPACES ASSOCIATION

IDSA on its way to a global standard – with the dataspace protocol in its core



Adapted from a presentation by L. Nagel at the Data Spaces Symposium 2024



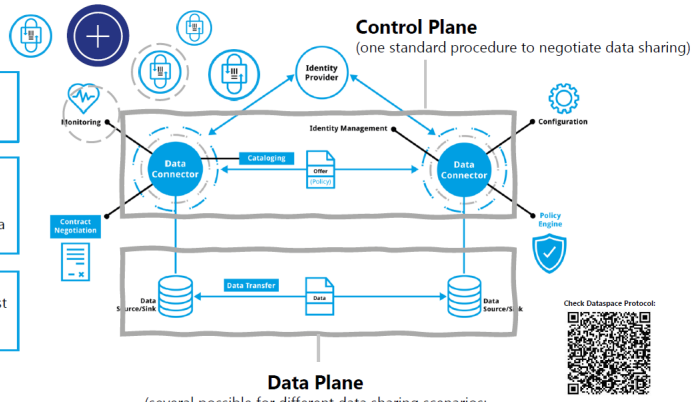
https://www.dataspaces-radar.org/

Dataspace Protocol V1.0 → ISO Standard

INTERNATIONAL DATA SPACES ASSOCIATION

Enables standardized data exchange across different data space instances.

- Ensures standardized data exchange mechanism between different frameworks, products, or services.
- Rrovides the needed schemas and protocols for cataloging data, negotiating contracts and usage agreements, and accessing data within a data space.
- Organizations using this protocol can align with industry standards, foster best practices, and unlock new data-driven business models and opportunities.



16

Adapted from a presentation by L. Nagel at the Data Spaces Symposium 2024

Dataspace Protocol V1.0 → ISO Standard

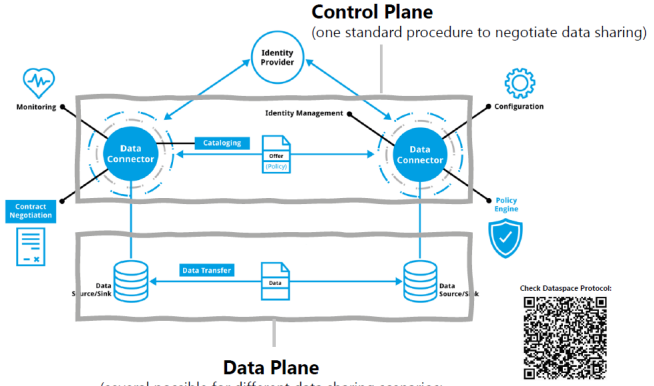
INTERNATIONAL DATA SPACES ASSOCIATION

Enables standardized data exchange across different data space instances.

Control Plane decides who can access the data and how.

Data Plane is where the action (data sharing) happens.

Conceptually divided, can be combined practically



31

Adapted from a presentation by L. Nagel at the Data Spaces Symposium 2024

Make the connection and enable data economy

The key to data spaces is the data connector

INTERNATIONAL DATA SPACES ASSOCIATION

- **Connects participants in a data space** – to share, utilize, benefit from data.
- Ensures **trust through IDS Certification** and **cyber security** assessment.
- Connects to **trust frameworks** and **identity management**
- Includes **identity & policy management**, ensures **data usage control**.
- Guarantees **interoperability**.
- Understands and enforces **data usage policies**.
- **Master** for other connectors of diverse feature sets.

IDSA Story *39*

29

Adapted from a presentation by L. Nagel at the Data Spaces Symposium 2024

Standardized Data Exchange

What does this mean? How does Dataspace Protocol ensure that?

INTERNATIONAL DATA SPACES ASSOCIATION

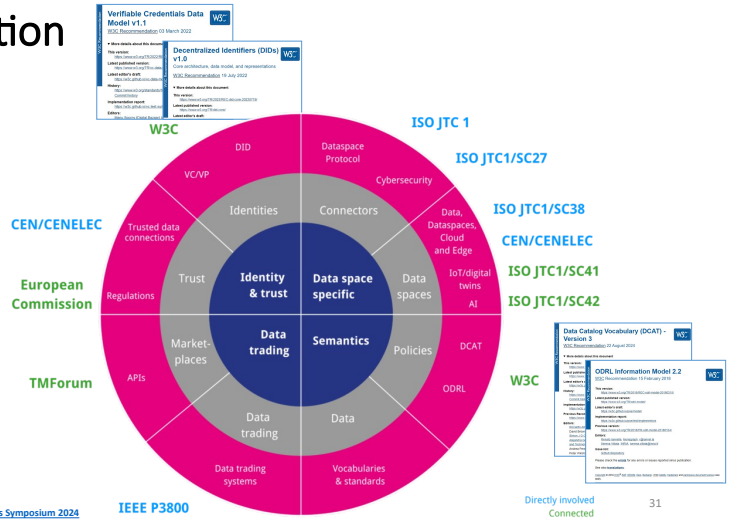
	Catalog	*Contract Negotiation*	*Transfer Process*
What happens?	A public transportation authority decides to share its transit schedules with app developers.	An app developer wants to use these schedules to create a route planning application.	Once the agreement is in place, the actual data transfer begins.
Problem	Inconsistent data formats for schedules across different platforms.	Need for clear terms regarding the use and distribution of the transit data.	Ensuring secure, efficient, and reliable transfer of transit data.
Role of DSP	Standardizes the format for publishing transit schedules.	Facilitates agreement on data usage terms and conditions.	Manages the secure and efficient transfer of the agreed-upon data.
Specification Example	Data provider publishes schedules using *'DCAT Catalogs'* and sets access rules with *'ODRL Policies'*.	Developer and authority negotiate using *'Contract Offer'* messages, leading to a *'Contract Agreement'*.	Data transfer is executed through *'Connector-to-Connector Communication'* and *'Data Transfer Requests'*.

43

30

Adapted from a presentation by L. Nagel at the Data Spaces Symposium 2024

Standardization activities



31

Adapted from a presentation by L. Nagel at the Data Spaces Symposium 2024

Data Interoperability and Semantics

< Part 5. The European Data Strategy >

A guided tour !

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