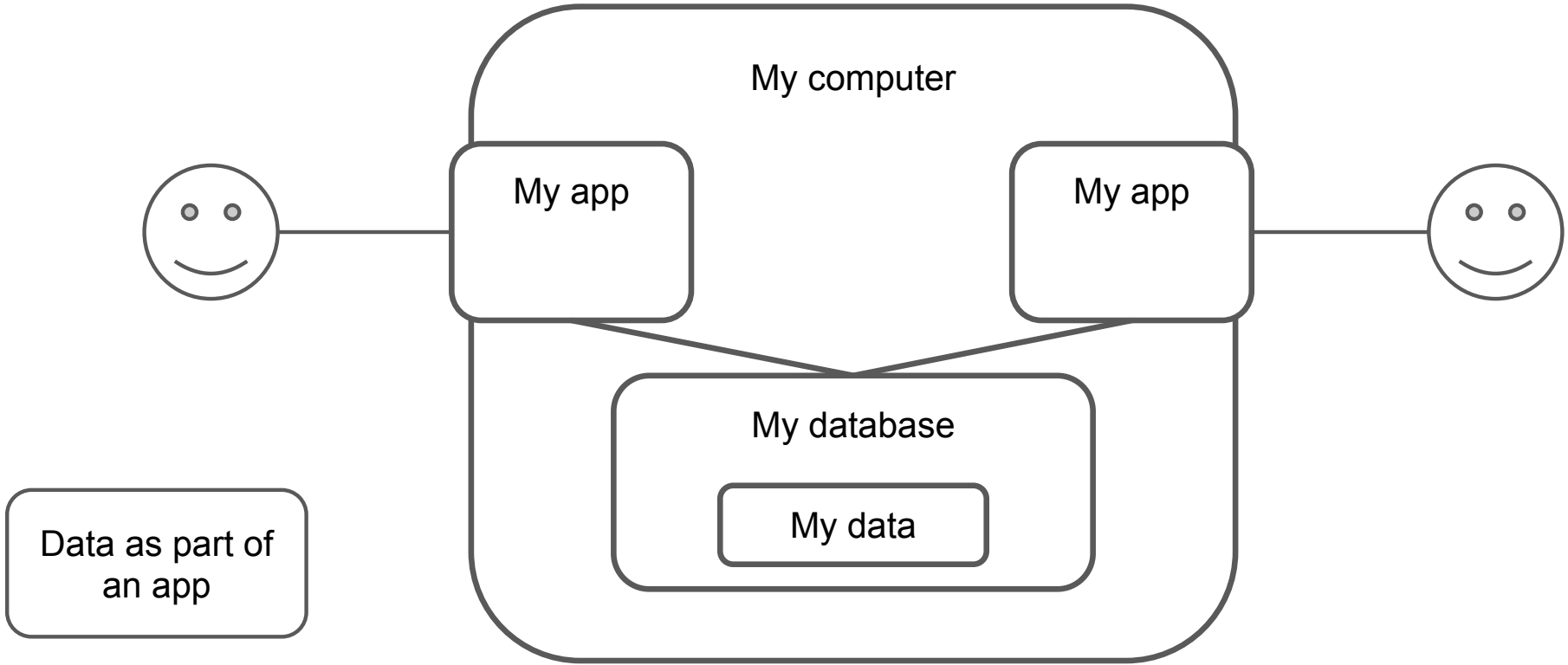
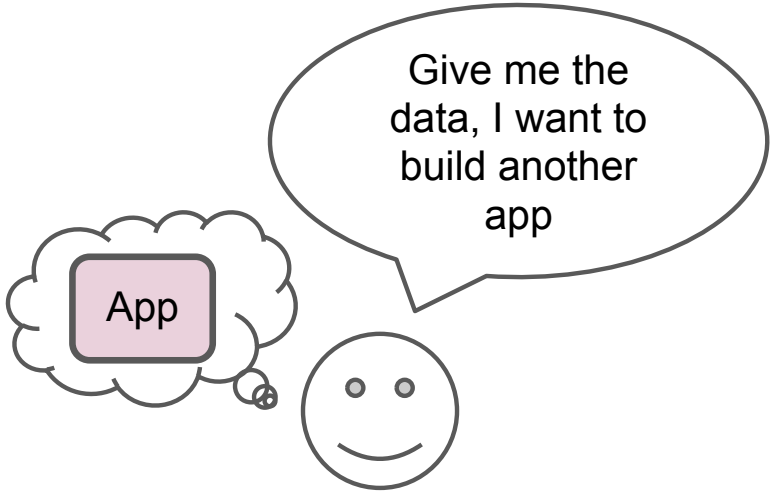
Data Formats Introduction

Jakub Klímek

Some developers think mainly in terms of apps...

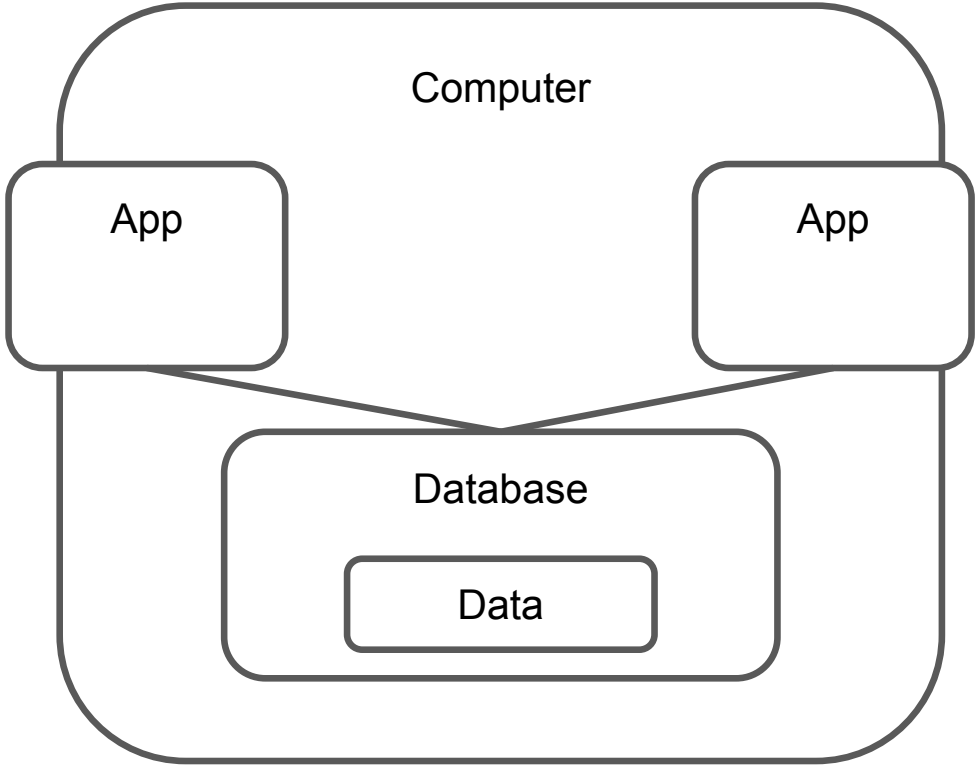


In public administration…

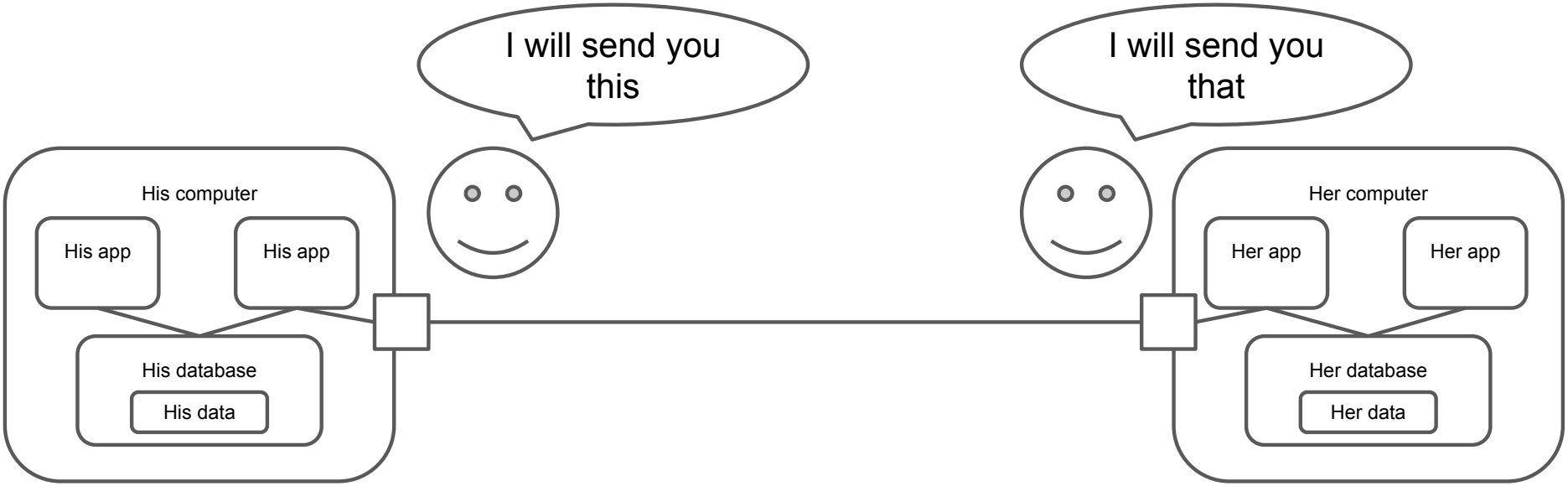


Data as part of an app

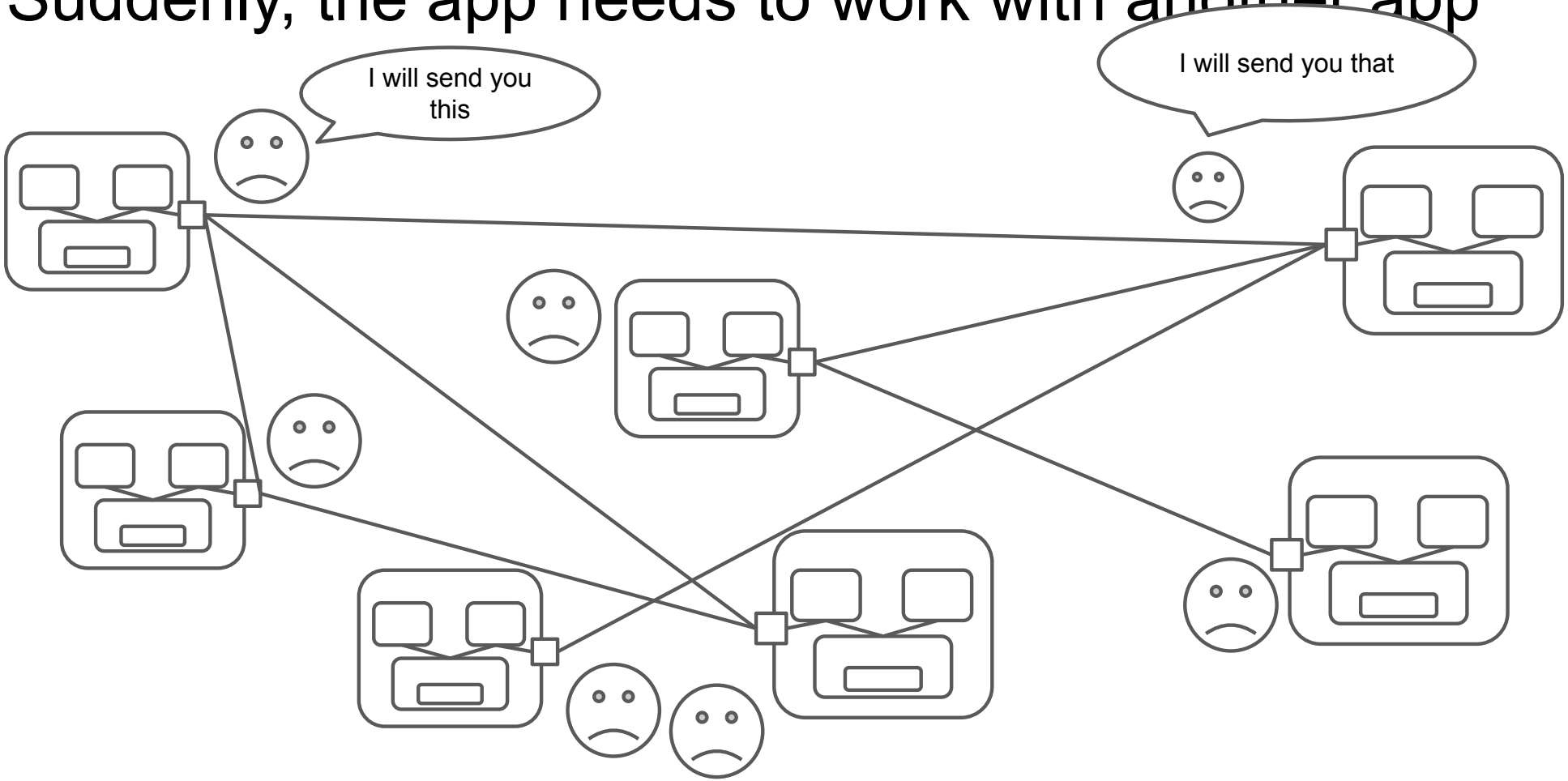
Possible vendor-lock



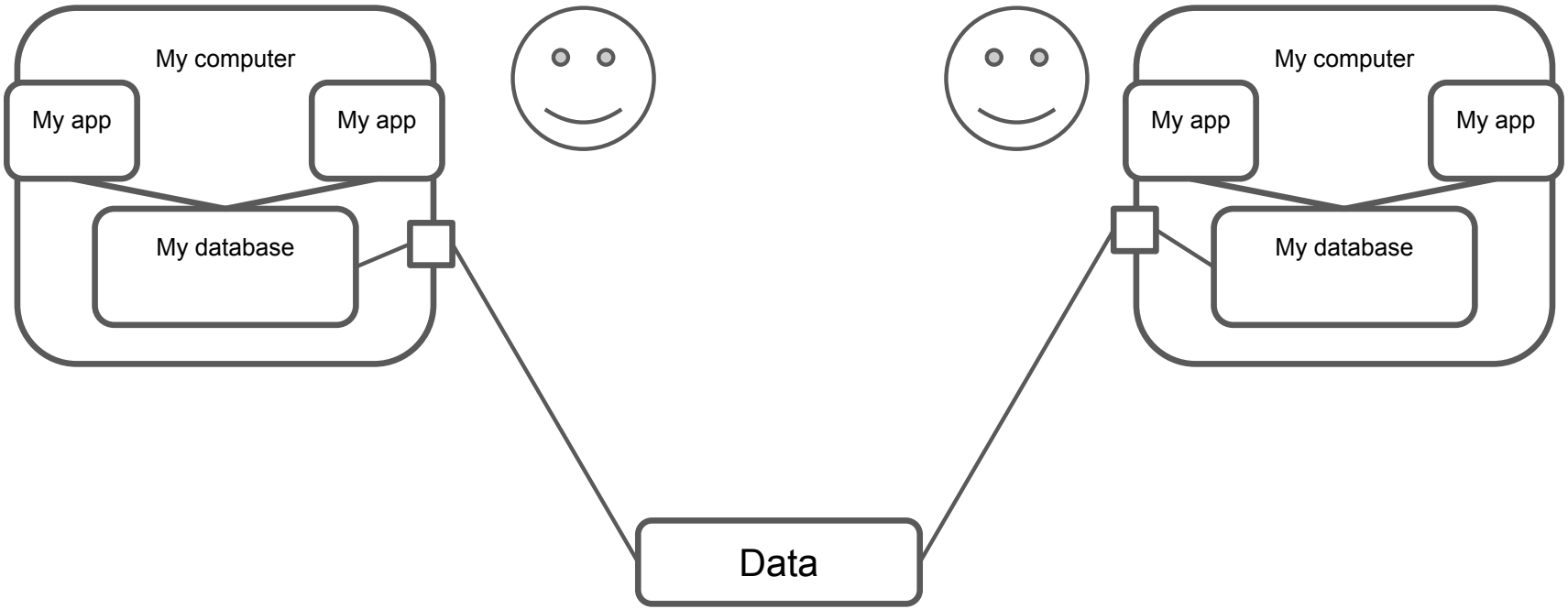
Suddenly, the app needs to work with another app



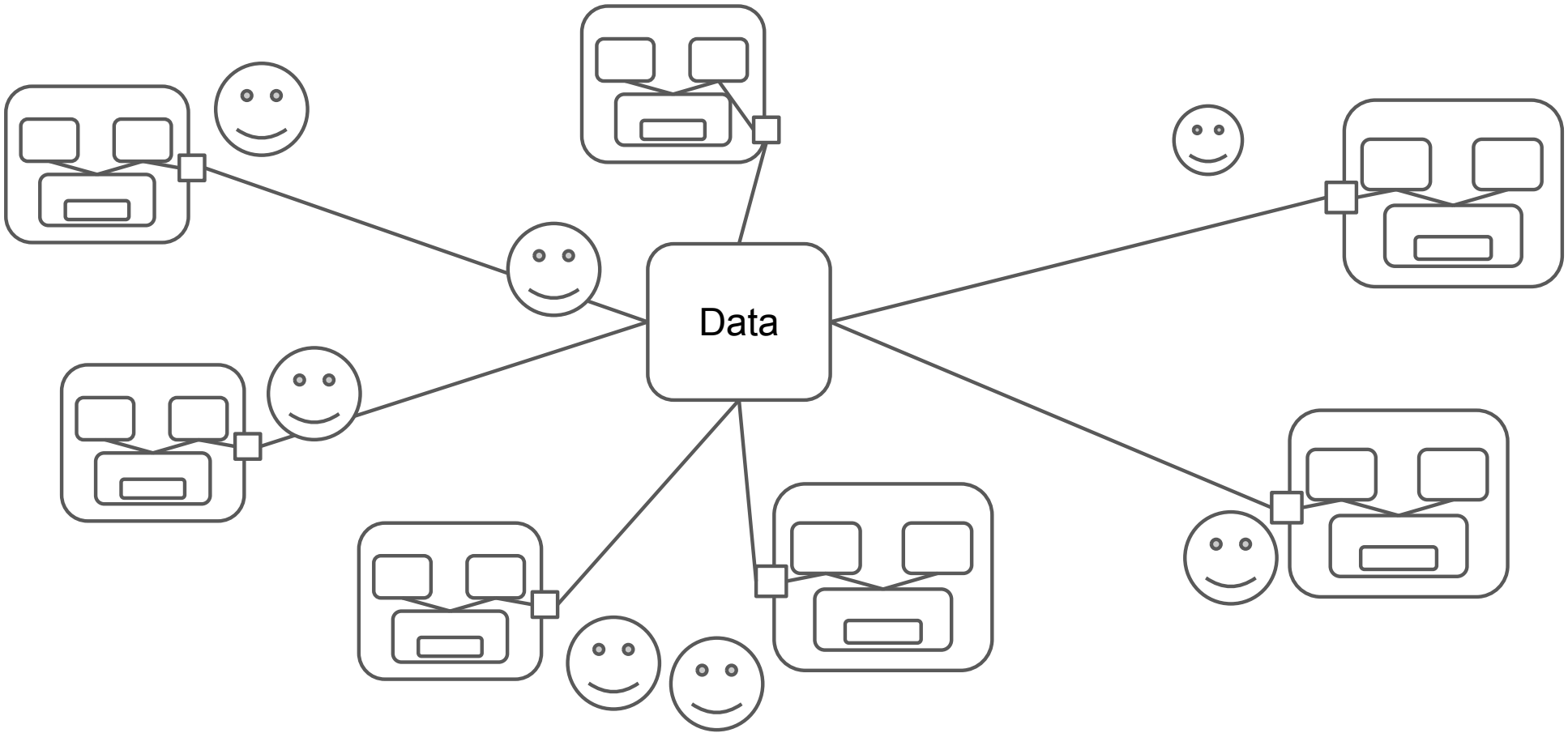
Suddenly, the app needs to work with another app



Data independent of applications



Data independent of applications



OK, data is independent. More problems?

Use of improper formats for a given use case

- e.g. tabular data in hierarchical data format (XML, JSON)

Not following specifications

- errors or unnecessary work when data is shared
- whose fault is it?
  - the producer?
  - the consumer?
- ultimately, who will pay for mitigating the issue?

Point of this course

1. overview of data formats, specifications, tools and use cases
2. ability to choose a proper data format for a given use case
3. thinking about data independently of applications and at various levels of abstraction
    - conceptual level - what is the data about
    - logical level - how is data structured using given technology/format
    - physical level - how do the files look like in storage

Conceptual view of data

Conceptual domain model

Independent of any particular technology or representation. Answers:

What real world entities are described?

What are their properties?

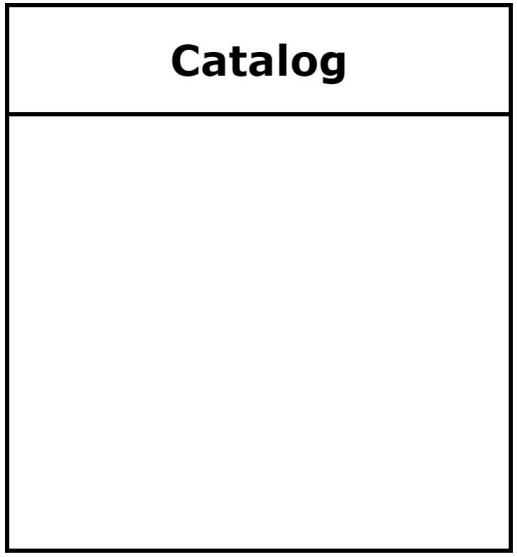
How are they connected?

Conceptual model can be discussed with non-IT personnel.

Conceptual domain model - UML Class diagrams

Class: `Catalog`

This is saying:
“There are things in the real-world of a type Catalog.”

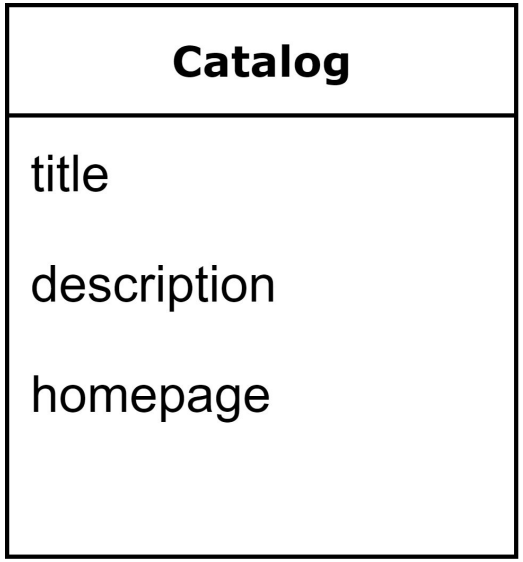


Conceptual domain model - UML Class diagrams

Class: `Catalog`

Attributes: `title`, `description`, `homepage`

This is saying:
“Each instance of a catalog has a title, description and a homepage.”

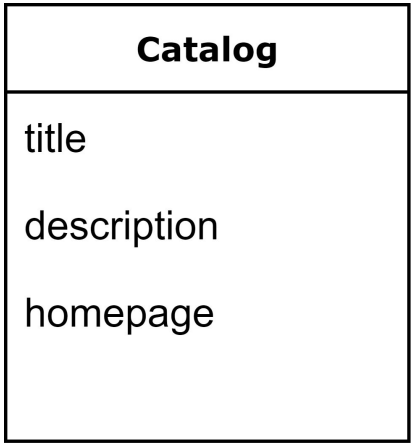


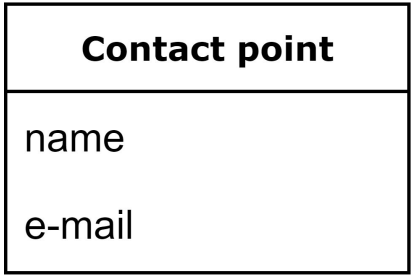
Conceptual domain model - UML Class diagrams

Class: `Catalog`
Attributes: `title`, `description`, `homepage`

Class: `Contact point`
Attributes: `name`, `e-mail`

This is saying:
“There are contact points, each has a name and an e-mail”





Conceptual domain model - UML Class diagrams

Class: `Catalog`
Attributes: `title`, `description`, `homepage`

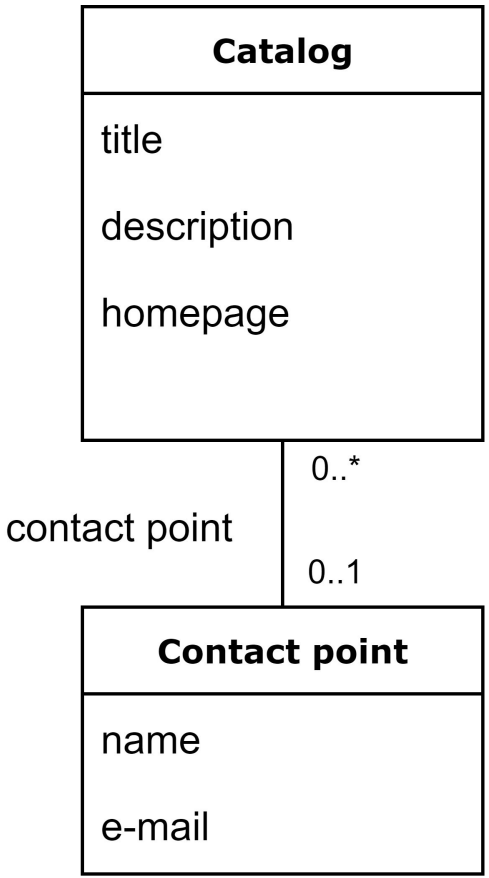
Class: `Contact point`
Attributes: `name`, `e-mail`

Association: `contact point`

- Association end 1: `Catalog`, 0..*
- Association end 2: `Contact point`, 0..1

This is saying:

- An instance of a Catalog may or may not be connected to up to 1 instance of Contact point
- An instance of a Contact point may or may not be connected to an arbitrary number of instances of Catalog



Conceptual domain model - UML Class diagrams

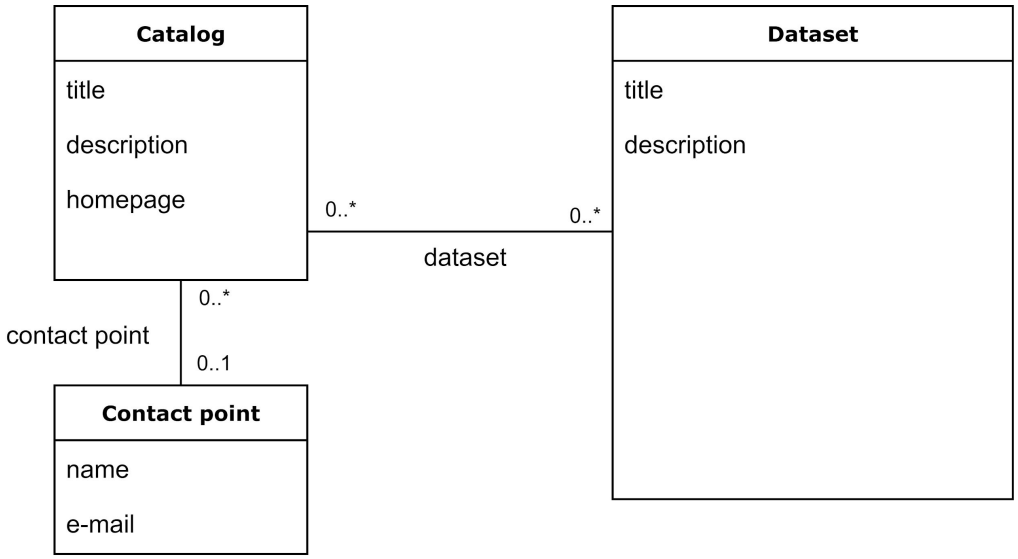
Class: `Dataset`
Attributes: `title`, `description`

Association: `dataset`

- Association end 1: `Catalog`, 0..*
- Association end 2: `Dataset`, 0..*

This is saying:

- An instance of a Catalog may or may not be connected to an arbitrary number of instances of Dataset
- An instance of a Dataset may or may not be connected to an arbitrary number of instances of Catalog



Conceptual domain model - UML Class diagrams

Attributes of `Dataset`:

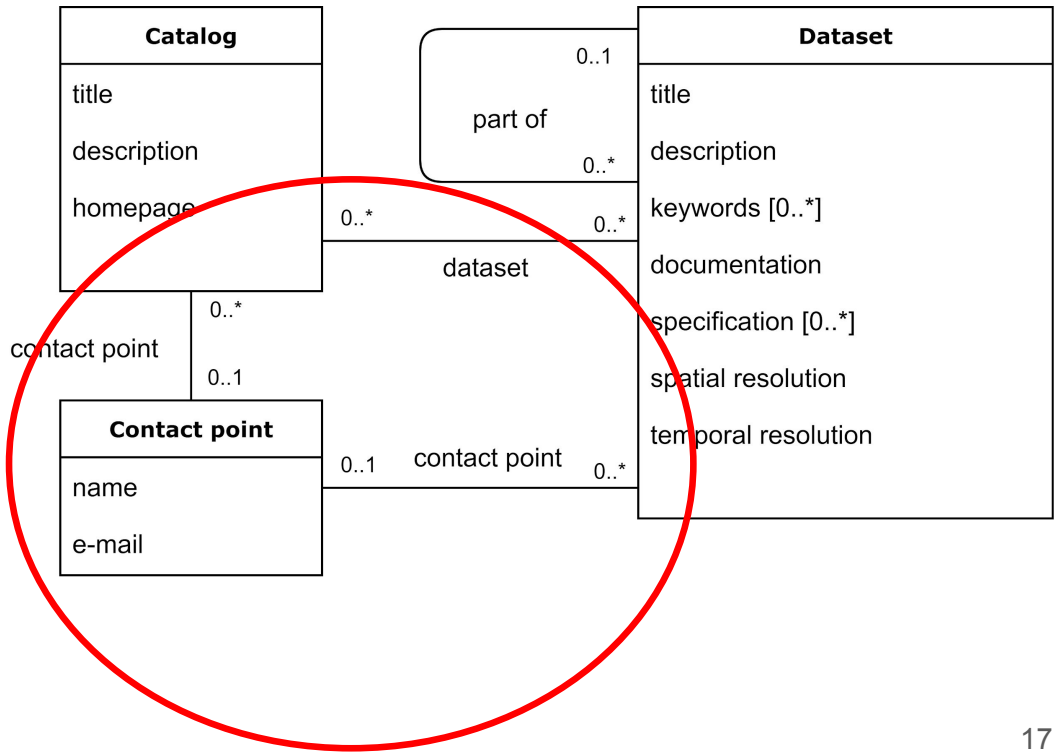
- `keywords [0..*]`, `specification [0..*]`
- `documentation`
- `spatial resolution`, `temporal resolution`

Association: `part of`

- Association end 1: `Dataset`, 0..1
- Association end 2: `Dataset`, 0..*

Association: `contact point`

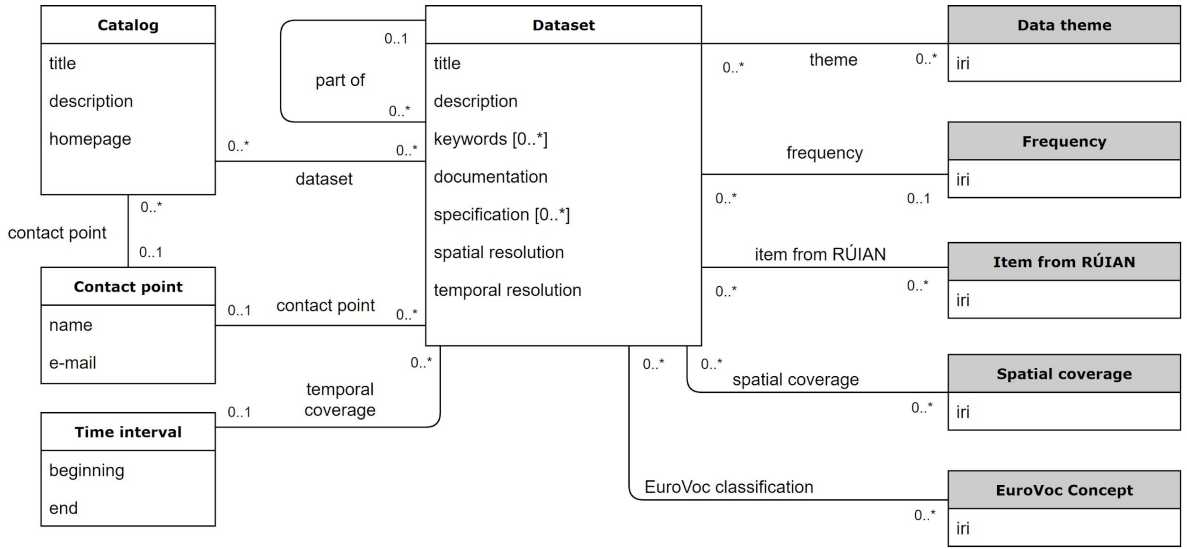
- Association end 1: `Contact point`, 0..1
- Association end 2: `Dataset`, 0..*



Conceptual domain model - UML Class diagrams

Associations:

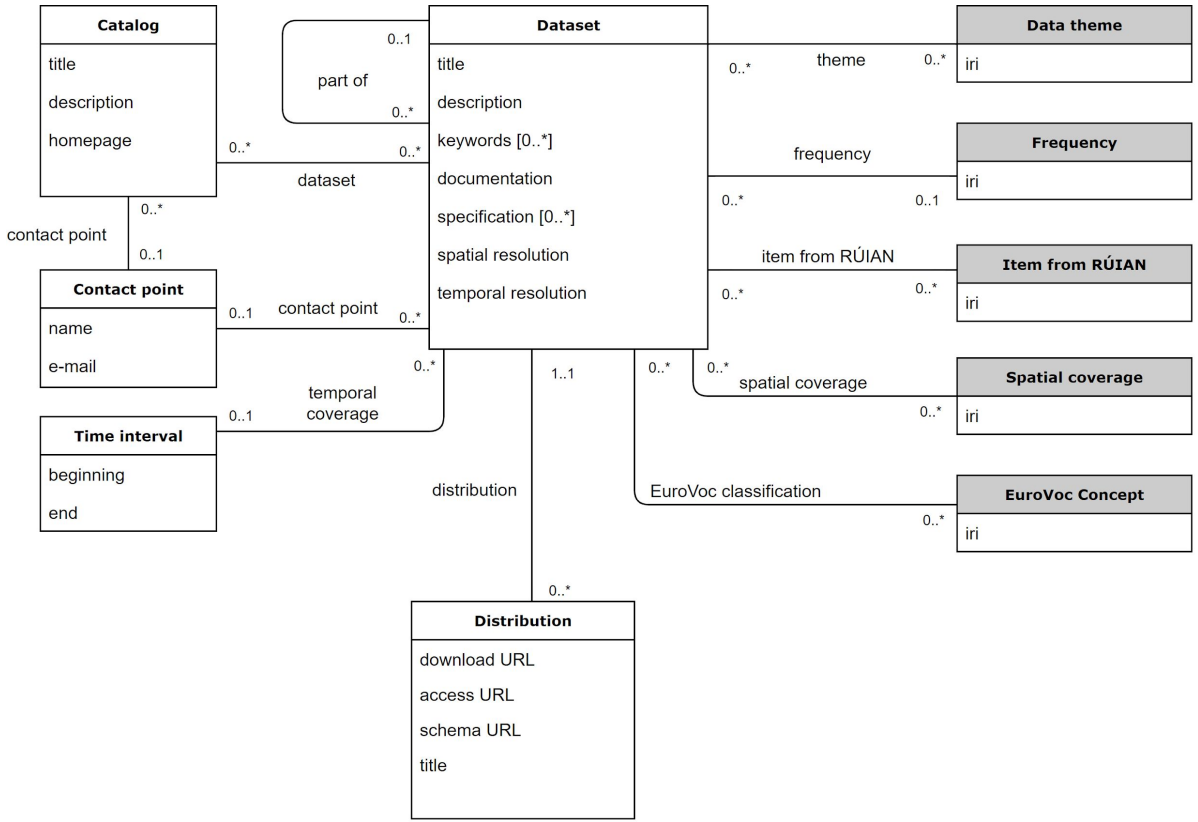
- theme
- frequency
- item from RÚIAN
- spatial coverage
- EuroVoc classification
- Time interval



Conceptual domain model - UML Class diagrams

Association: `distribution`

- Association end 1:
  `Dataset`, 1..1
- Association end 2:
  `Distribution`, 0..*



Conceptual domain model - UML Class diagrams

Association: `media type`

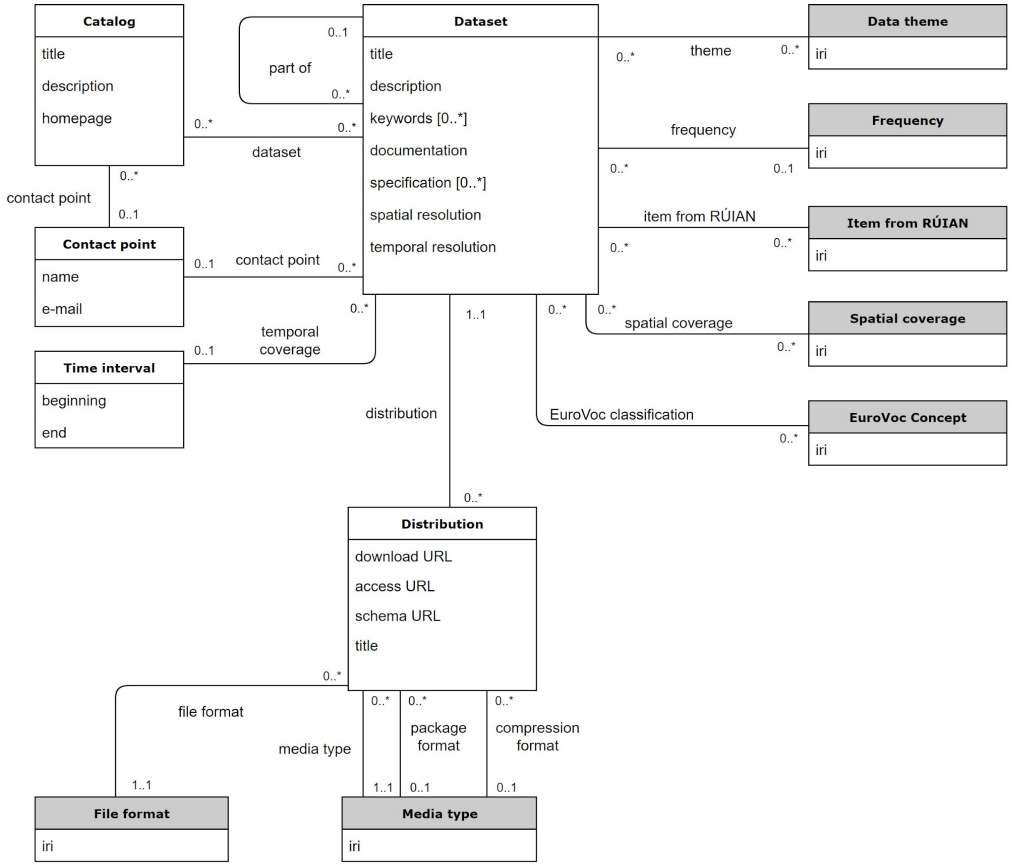
- Association end 1: `Distribution`, 0..*
- Association end 2: `Media type`, 1..1

Association: `package format`

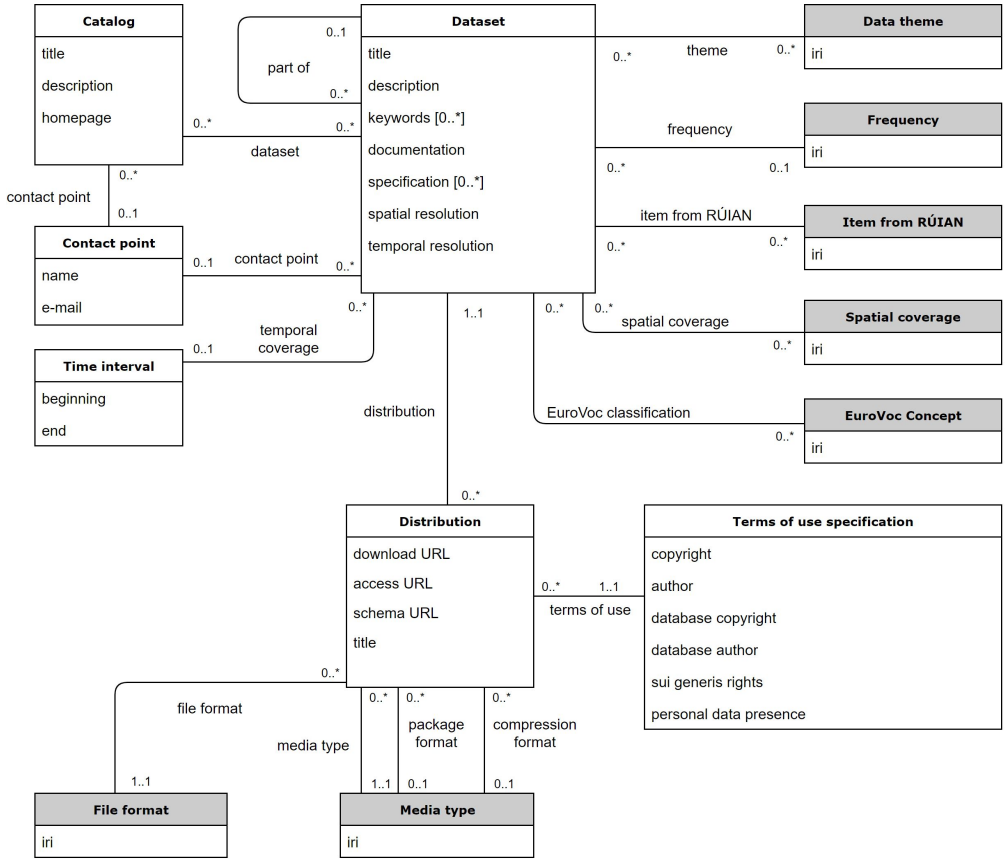
- Association end 1: `Distribution`, 0..*
- Association end 2: `Media type`, 0..1

Association: `compression format`

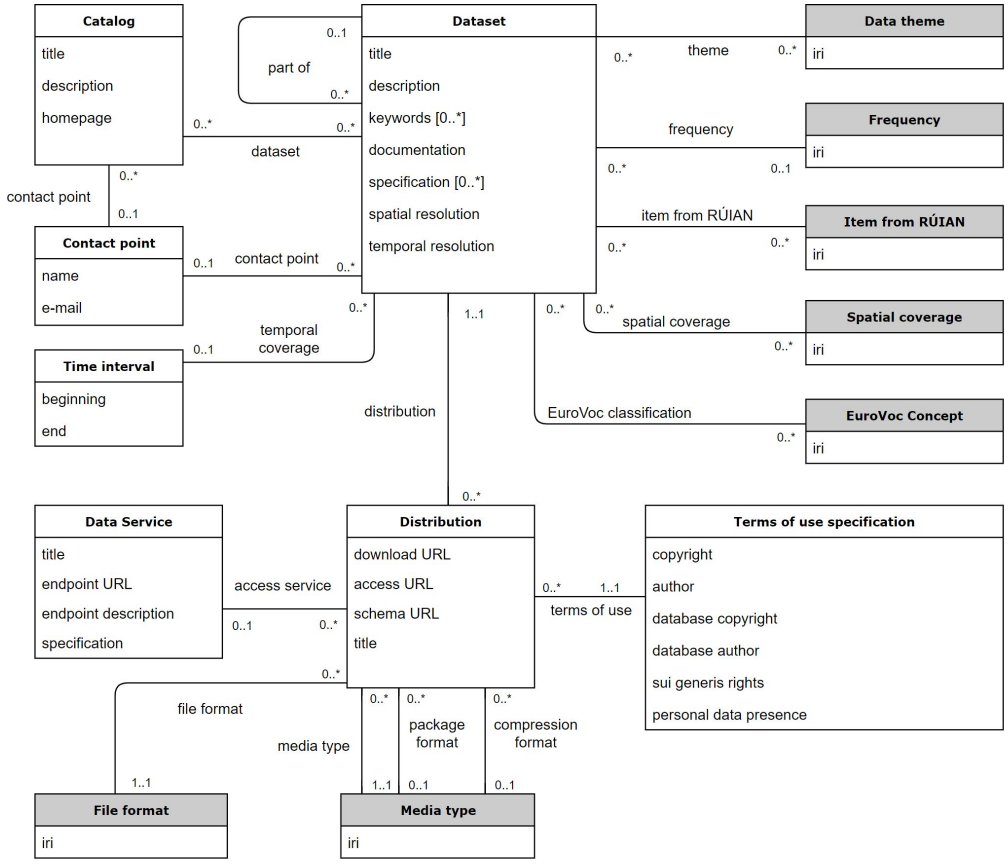
- Association end 1: `Distribution`, 0..*
- Association end 2: `Media type`, 0..1



Conceptual domain model - UML Class diagrams



Conceptual domain model - UML Class diagrams



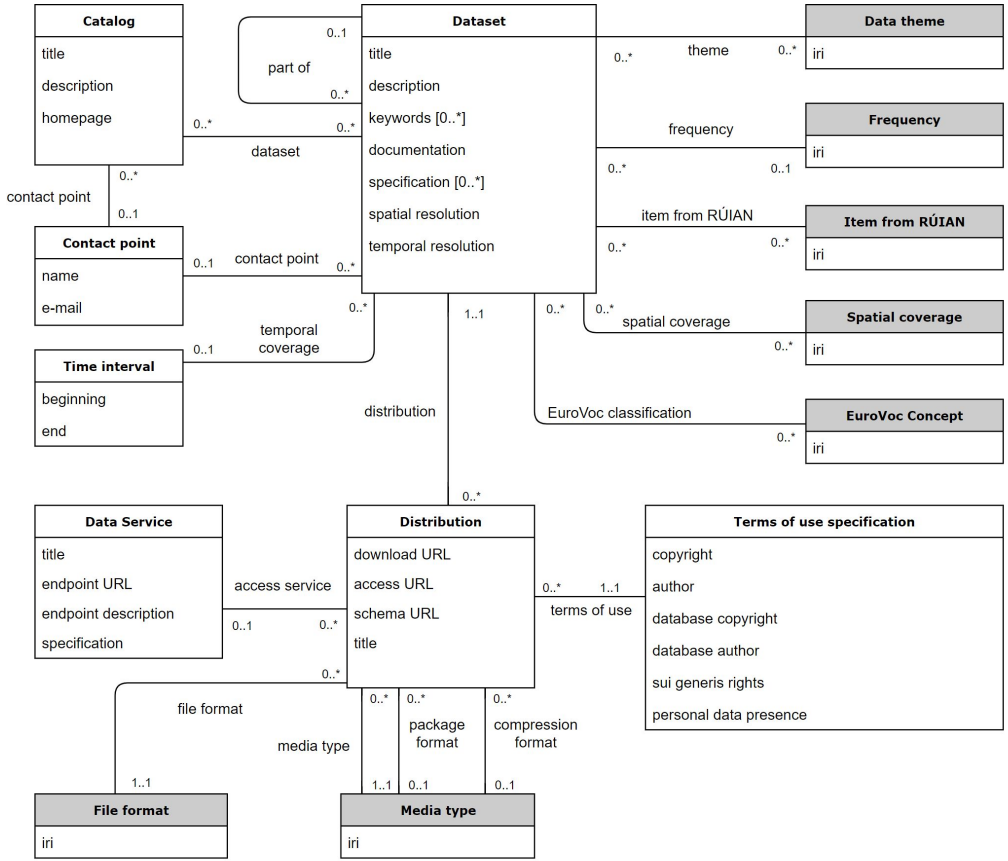
Conceptual domain model - UML Class diagrams

DCAT-AP-CZ

- Czech specification of how to represent data catalogs
  - 2021
- Based on DCAT-AP 2.0.1
  - European application profile
  - European Commission, 2020
- Based on DCAT 2
  - W3C Recommendation, 2020

Representations in:

RDF, JSON, CSV



Data models

vs.

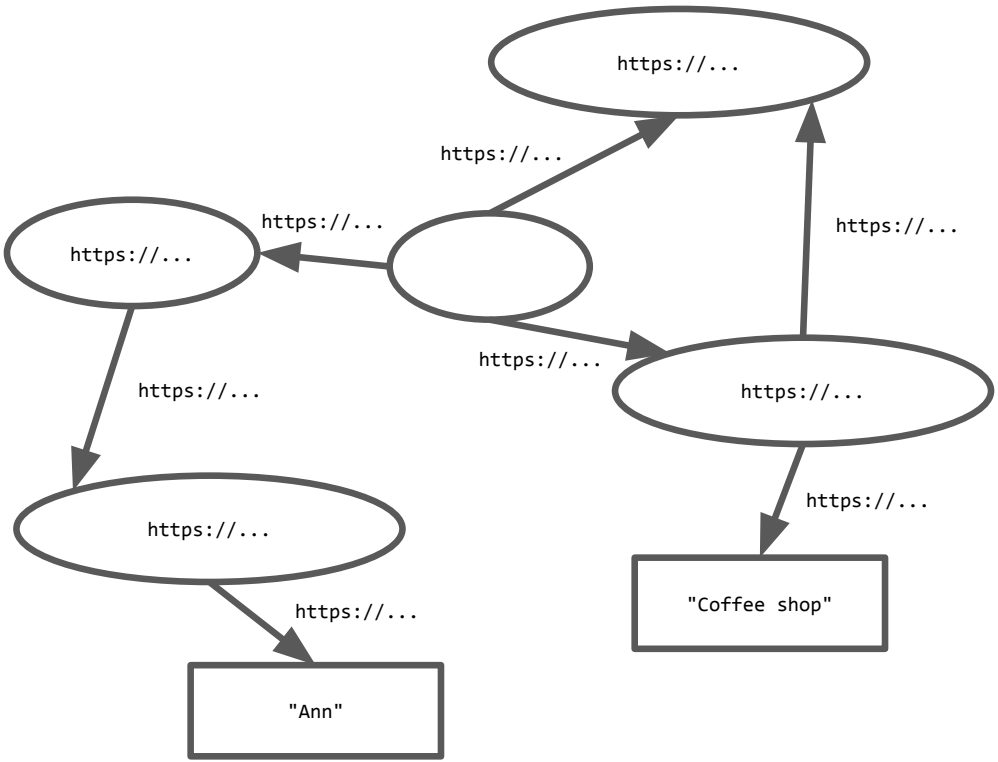
Data formats

vs.

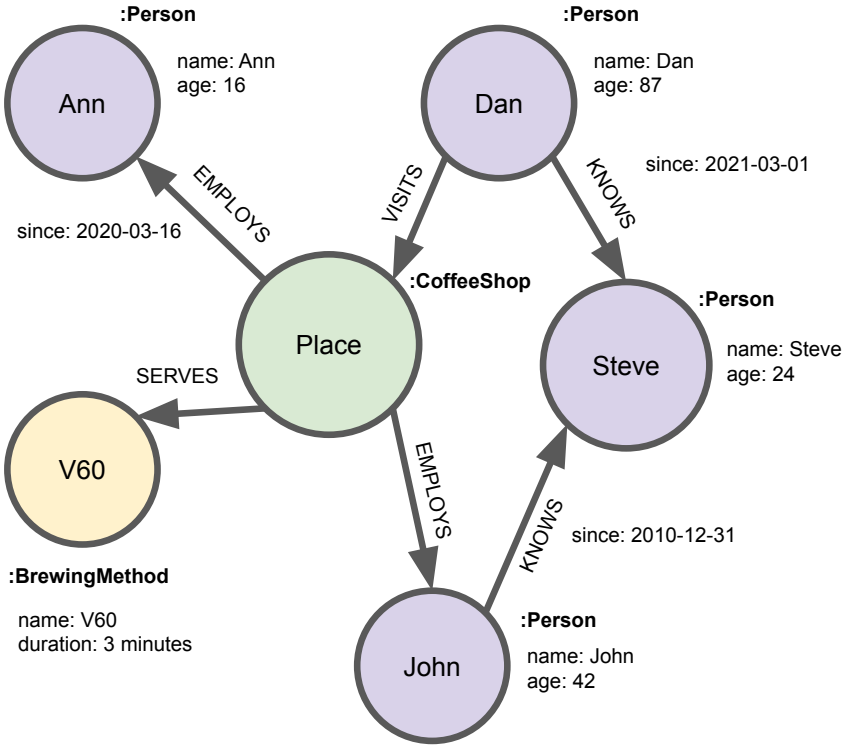
Data schemas

Data models - logical view of data - graphs

Resource Description Framework (RDF) model

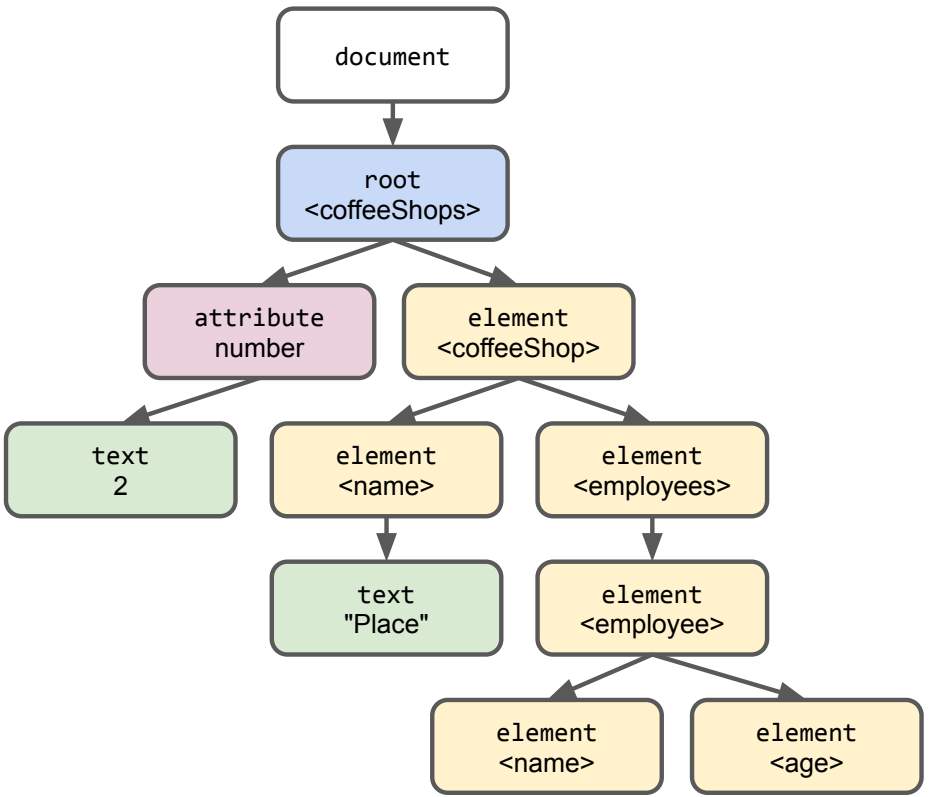


Labeled Property Graph (LPG) model

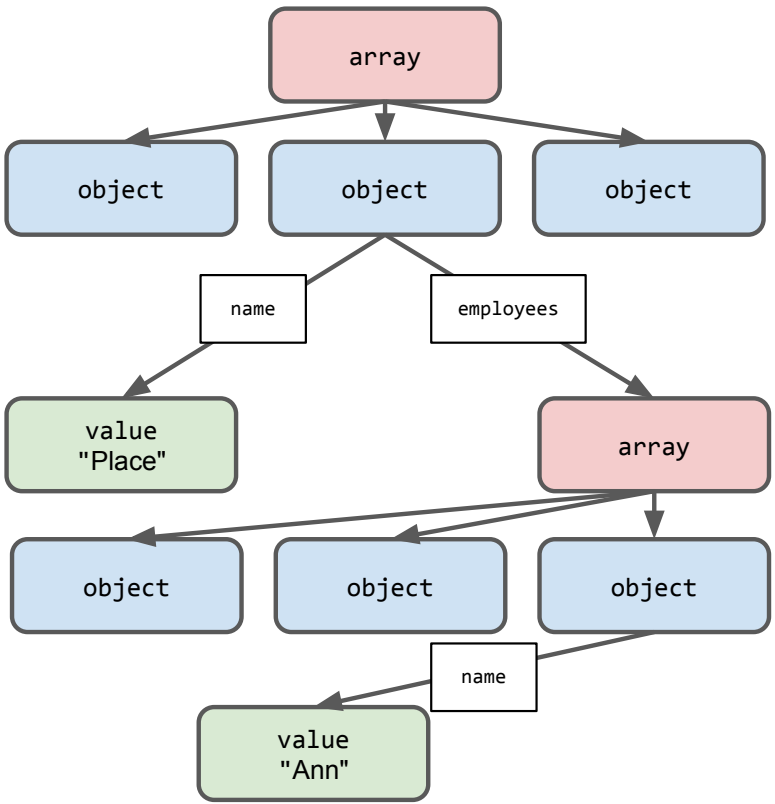


Data models - logical view of data - hierarchies/trees

Document Object Model (DOM)

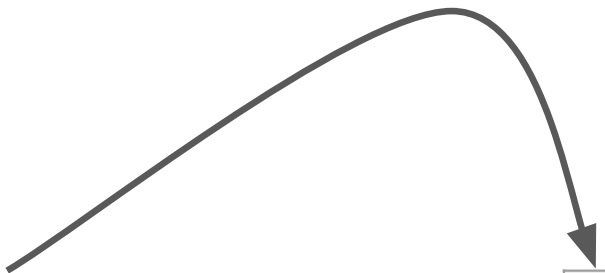


JSON (both format and model)



Data models - logical view of data - tables

Relational model



name	employee
Place	Ann
Place	John

name	age	knows
Ann	16	
John	42	Steve
Steve	24	
Dan	87	Steve

Data formats - physical view of data

How data using a certain data model is serialized into files / sent over network

Graph

- RDF graph model
  - Text-based: N-Triples, N-Quads, Turtle, TriG, RDF/XML, JSON-LD, RDFa
  - Binary: HDT

- Property graph
  - CSV, JSON, GraphML, Cypher Script

Hierarchical

- DOM
  - XML, HTML
- JSON
  - JSON, XML

Relational

- CSV, SQL dump

Data schemas

Annotations and constraints applicable to instances of data formats, allowing the data to be better described and validated

CSV

- Schema language
  - **CSV on the Web**

RDF

- Schema language
  - SHACL
  - ShEx

JSON

- Schema language
  - **JSON Schema**

XML

- Schema languages
  - DTD
  - **XML Schema**
  - Relax NG
  - Schematron

Specific data formats using meta-formats

CSV, JSON, XML, … sometimes called meta-formats

They serve as “host” formats for use-case specific formats

Data schemas used to define these specific formats

JSON

- GeoJSON

CSV

- General Transit Feed Specification (GTFS)
- Example of Prague public transport

XML

- SVG, Atom, RSS 2.0, Office Open XML (.docx, .xlsx), …

RDF

- DCAT, Schema.org

Generic data format properties
open vs. closed
machine-readability
binary vs. text-based

Open vs. closed formats

Open

Specification available on the Web, freely accessible to anyone, with no limitation on its usage.

- Meta-formats e.g.: XML, JSON, CSV, RDF
- Specific formats
  - SVG, GeoJSON, ...

Closed

- specification not accessible
- need for payment for access to specification
- need for registration
- need for certification of library/application claiming compatibility

Examples

- railML.org
  - XML based
  - need for certification

Machine-readable format

“All files are machine-readable, because all files are, in the end, read by machines”

The second part is true, but not what machine readability is about

Machine-readable format

❓ ❓ ❓

“Open formats like CSV, XML, JSON or RDF Turtle and Excel .xlsx files are machine readable”

Machine-readable format - CSV, JSON

```
,,,,,,,,,,,
Back to TOC,,,,,,,,,,,
r2 : R2. Do you have permanent residence in
Brno?,,,,,,,,,,,
 ,%,count,,,,,,,,,
Yes,89.1%,1385,,,,,,,,,
No,10.9%,169,,,,,,,,,
TOTAL,100.0%,1554,,,,,,,,,
"Total sample, Weight: Weight, base n =
1554",,,,,,,,,,,
,,,,,,,,,,,
Back to TOC,,,,,,,,,,,
r3 : R3. For how long have you lived in Brno?
,,,,,,,,,,,
 ,%,count,,,,,,,,,
```

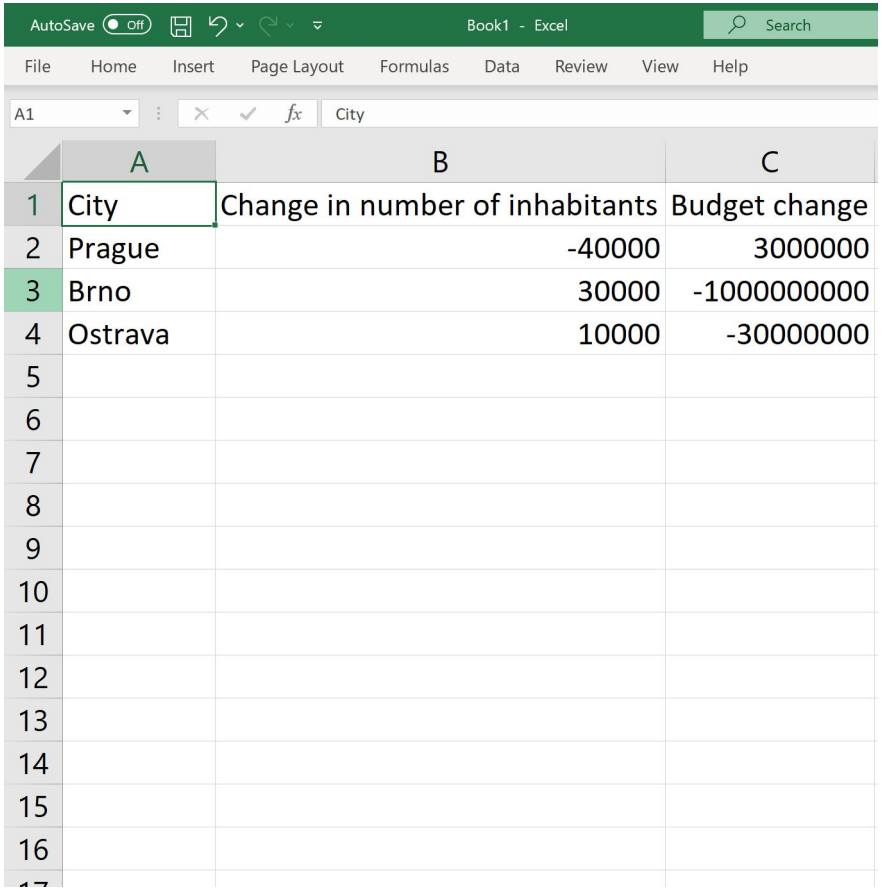
```
"ico","nazev","udaje","vymazDatum","zapisDatum"

"3571092","Nadace RK CARE","[{hlavicka=Spisová
značka;zapisDatum=2014-11-20;hodnotaText=N
521/KSBR;udajTyp={kod=SPIS_ZN;nazev=spisová
značka};spisZn={soud={kod=KSBR;nazev=Krajský soud v
Brně};oddil=N;vlozka=521}},
{hlavicka=Název;zapisDatum=2014-11-20;hodnotaText=Nadace RK
CARE;udajTyp={kod=NAZEV;nazev=název}},
{hlavicka=Sídlo;zapisDatum=2014-11-20;udajTyp={kod=SIDLO;naz
ev=sídlo};adresa={statNazev=Česká
republika;obec=Lipůvka;castObce=Lipůvka;cisloPo=385;psc=6792
2;okres=Blansko}}, {hlavicka=Identifikační
číslo;zapisDatum=2014-11-20;hodnotaText=3571092;udajTyp={kod=
ICO;nazev=identifikační číslo}}, {hlavicka=Právní
forma;zapisDatum=2014-11-20;hodnotaText=nad;udajTyp={kod=PRA
VNI_FORMA;nazev=právní
forma};pravniForma={kod=nad;nazev=Nadace;zkratka=nad}},
{hlavicka=Účel
nadace;zapisDatum=2014-11-20;udajTyp={kod=UCEL_SUBJEKTU_SEKC
```

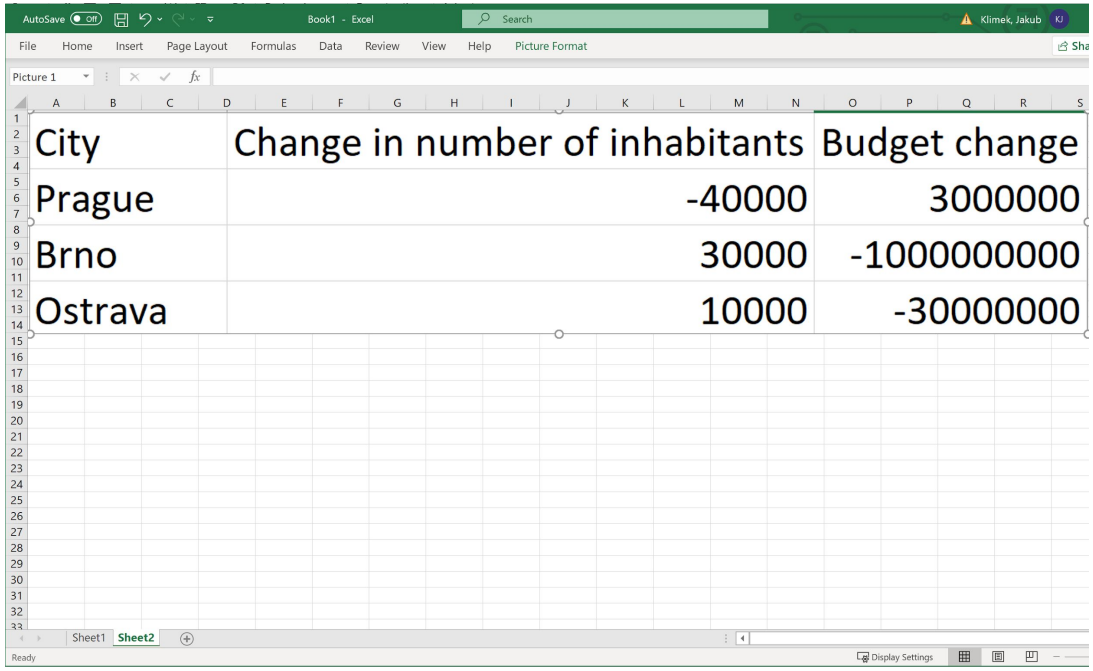
Machine-readable format - XML


```xml
<?xml version="1.0" encoding="UTF-8"?>
<PvsRejstrikData rejstrik="" operace="1" xmlns="http://portal.gov.cz/portal/xsd/PvsRejstrikData">
<TYPE>datová sada</TYPE>
<NAZEV>Smlouvy SŽDC 2017</NAZEV>
<POPIS>Uzavřené smlouvy organizace Správa železniční a dopravní cesty (resort dopravy) v roce 2017</POPIS>
<HOMEPAGE></HOMEPAGE>
<PERIODICITY></PERIODICITY>
<SPATIAL_TYPE></SPATIAL_TYPE>
<SPATIAL_TYPE_TXT></SPATIAL_TYPE_TXT>
<SPATIAL_CODE></SPATIAL_CODE>
<SPATIAL_CODE_TXT>Česká republika</SPATIAL_CODE_TXT>
<THEME></THEME>
<THEME_TXT>-</THEME_TXT>
<KEYWORDS>smlouva</KEYWORDS>
<STAV>zpracováno 2017-03-29 15:54:05</STAV>
<PROBLEMY></PROBLEMY>
<x-priloha MimeTyp="application/xml"
Jmeno="data.xml">PD94bWwgdmVyc2lvbj0iMS4wIiBlbmNvZGluZz0iVVRGLTgiIHN0YW5kYWxvbmU9Im5vIj8+CjxkYXRhc2V0IHhtbG5zPSJodHRwOi8vcG9ydGFsLmdvdi5jei9wb3J0YWwveHNkL1B2c1JlanN0cmlrRGF0YSIgSUQ9IiI
gb3BlcmFjZT0iMSI+CiAgPHRpdGxlPlNtbG91dnkgU8W9REMgMjAxNzwvdGl0bGU+CiAgPGRlc2NyaXB0aW9uPlV6YXbFmWVuw6kgc21sb3V2eSBvcmdhbml6YWNlIFNwcsOhdmEgxb5lbGV6bmnEjW7DrSBhIGRvcHJhdm7DrSBjZXN0eSAocmV
zb3J0IGRvcHJhdnkpIHYgcm9jZSAyMDE3PC9kZXNjcmlwdGlvbj4KICA8YWNjcnVhbFBlcmlvZGljaXR5PlIvUDFNPC9hY2NydWFsUGVyaW9kaWNpdHk+CiAgPHNwYXRpYWw+CiAgICA8dHlwZT5TVDwvdHlwZT4KICAgIDxub3RhdGlvbj4xPC9
ub3RhdGlvbj4KICA8L3NwYXRpYWw+CiAgPHRlbXBvcmFsPgogICAgPHN0YXJ0RGF0ZT4yMDE3LTAxLTAxPC9zdGFydERhdGU+CiAgICA8ZW5kRGF0ZT4yMDE3LTEyLTMxPC9lbmREYXRlPgogIDwvdGVtcG9yYWw+CiAgPGtleXdvcmQ+c21sb3V
2YTwva2V5d29yZD4KICA8ZGlzdHJpYnV0aW9uPgogICAgPGFjY2Vzc1VSTD5odHRwOi8vd3d3Lm1kY3IuY3ovTURDUi9tZWRpYS9vdGV2cmVuYWRhdGEvc21sb3V2eS8yMDE3L3NtbG91dnlfc3pkY18yMDE3LmNzdjwvYWNjZXNzVVJMPgogICA
gPGRvd25sb2FkVVJMPmh0dHA6Ly93d3cubWRjci5jei9NRENSL211ZGlhL290ZXZyZW5hZGF0YS9zbWxvdXZ5LzIwMTcvc21sb3V2eV9zemRjXzIwMTcuY3N2PC9kb3dubG9hZFVSTD4KICAgIDxmb3JtYXQ+dGV4dC9jc3Y8L2Zvcm1hdD4KICA
gIDxsaWNlbnNlPmh0dHBzOi8vcG9ydGFsLmdvdi5jei9wb3J0YWwvb3N0YXRuaS92b2xueS1wcmlzdHVwLWstZHMuaHRtbDwvbGljZW5zZT4KICA8L2Rpc3RyaWJ1dGlvbj4KPC9kYXRhc2V0Pgo=</x-priloha>
</PvsRejstrikData>
```

Machine-readable format - XLSX



	A	B	C
1	City	Change in number of inhabitants	Budget change
2	Prague	-40000	3000000
3	Brno	30000	-1000000000
4	Ostrava	10000	-30000000
5			
6			
7			
8			
9			
10			
11			
12			
13			
14			
15			
16			



Machine-readable format

Machine readability is not a property of a format

- Depends on the form of a particular data instance
- Says whether the data is **easily** processed by **appropriate** applications
- For example
  - tabular data: structured as table: rows, columns and cells well formatted, easily processed
  - textual data: individual characters easily accessible, i.e. without OCR
  - ...

Binary vs. text based formats

“Binary format means that the file is stored as 1s and 0s”

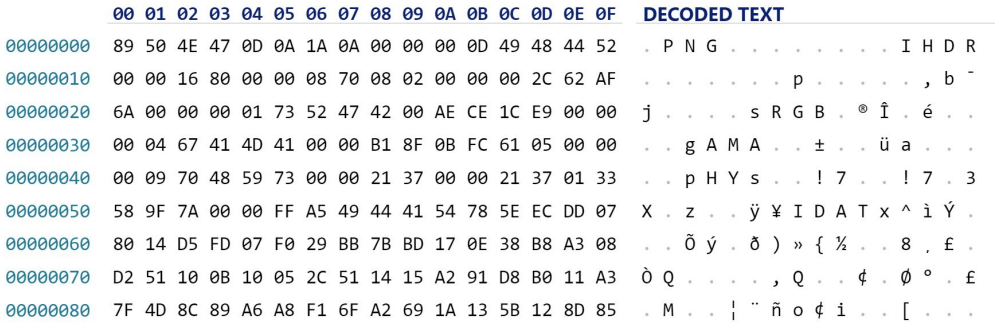
-- a student at a recent state exam

This is, of course, true also for text-based file formats

Binary vs. text based formats

Binary files

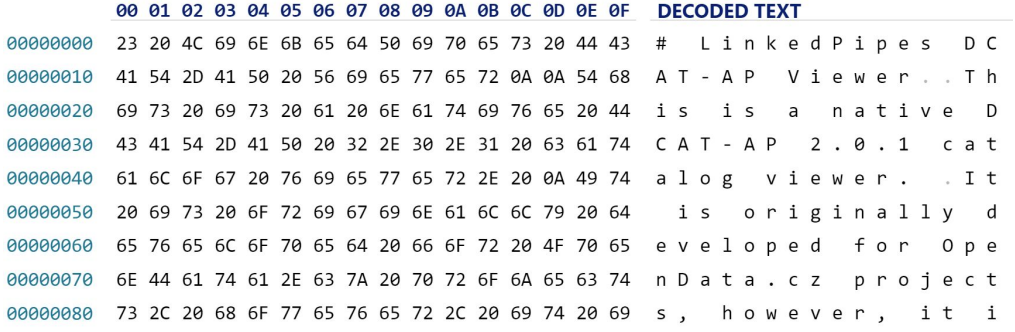
- Their structure may be defined on bit by bit level
- a.k.a. “non-text” file
- Not readable by text editors
- Viewable by hex editors



	00 01 02 03 04 05 06 07 08 09 0A 0B 0C 0D 0E 0F	DECODED TEXT
00000000	89 50 4E 47 0D 0A 1A 0A 00 00 00 0D 49 48 44 52	. P N G . . . . . I H D R
00000010	00 00 16 80 00 00 08 70 08 02 00 00 00 2C 62 AF	. . . . . p . . . . . , b ¯
00000020	6A 00 00 00 01 73 52 47 42 00 AE CE 1C E9 00 00	j . . . . s R G B . ® Î . é . .
00000030	00 04 67 41 4D 41 00 00 B1 8F 0B FC 61 05 00 00	. . g A M A . . ± . . ü a . . .
00000040	00 09 70 48 59 73 00 00 21 37 00 00 21 37 01 33	. . p H Y s . . ! 7 . . ! 7 . 3
00000050	58 9F 7A 00 00 FF A5 49 44 41 54 78 5E EC DD 07	X . z . . ÿ ¥ I D A T x ^ ì Ý .
00000060	80 14 D5 FD 07 F0 29 BB 7B BD 17 0E 38 B8 A3 08	. . Õ ý . ð) » { ½ . . 8 ¸ £ .
00000070	D2 51 10 0B 10 05 2C 51 14 15 A2 91 D8 B0 11 A3	Ò Q . . . . , Q . . ¢ . Ø ° . £
00000080	7F 4D 8C 89 A6 A8 F1 6F A2 69 1A 13 5B 12 8D 85	. M . . ¦ ¨ ñ o ¢ i . . [. . .

Text-based files

- Contains text
- Typically structured as characters on lines
- Viewable by text editors
- Also viewable by hex editors
- Text is encoded into 1s and 0s using *character encoding*



	00 01 02 03 04 05 06 07 08 09 0A 0B 0C 0D 0E 0F	DECODED TEXT
00000000	23 20 4C 69 6E 6B 65 64 50 69 70 65 73 20 44 43	# LinkedPipes DC
00000010	41 54 2D 41 50 20 56 69 65 77 65 72 0A 0A 54 68	AT-AP Viewer..Th
00000020	69 73 20 69 73 20 61 20 6E 61 74 69 76 65 20 44	is is a native D
00000030	43 41 54 2D 41 50 20 32 2E 30 2E 31 20 63 61 74	CAT-AP 2.0.1 cat
00000040	61 6C 6F 67 20 76 69 65 77 65 72 2E 20 0A 49 74	alog viewer. .It
00000050	20 69 73 20 6F 72 69 67 69 6E 61 6C 6C 79 20 64	is originally d
00000060	65 76 65 6C 6F 70 65 64 20 66 6F 72 20 4F 70 65	eveloped for Ope
00000070	6E 44 61 74 61 2E 63 7A 20 70 72 6F 6A 65 63 74	nData.cz project
00000080	73 2C 20 68 6F 77 65 76 65 72 2C 20 69 74 20 69	s, however, it i

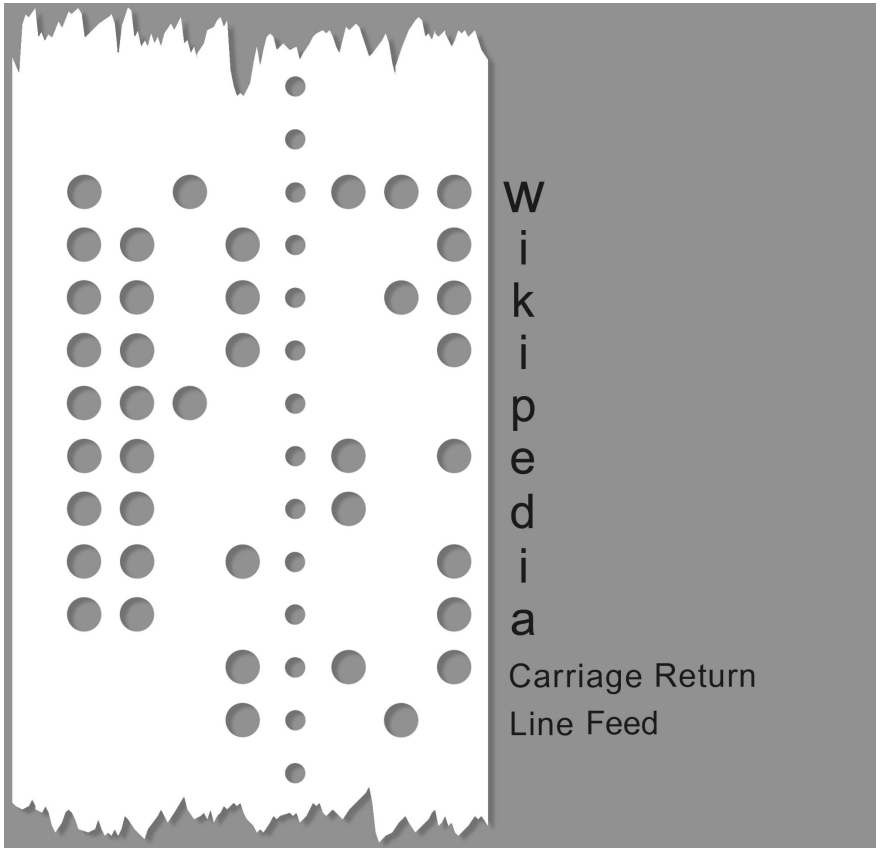
Text-based formats - character encoding - US-ASCII

Character encoding - representation of characters as binary sequences (numbers)

US-ASCII using 7 bits to represent 1 character

USASCII code chart

Bits b_7 b_6 b_5 →	0 0 0	0 0 1	0 1 0	0 1 1	1 0 0	1 0 1	1 1 0	1 1 1
b_4 b_3 b_2 b_1 ↓ Row ↓ / Column →	0	1	2	3	4	5	6	7
0 0 0 0	0	NUL	DLE	SP	0	@	P	`	p
0 0 0 1	1	SOH	DC1	!	1	A	Q	a	q
0 0 1 0	2	STX	DC2	"	2	B	R	b	r
0 0 1 1	3	ETX	DC3	#	3	C	S	c	s
0 1 0 0	4	EOT	DC4	$	4	D	T	d	t
0 1 0 1	5	ENQ	NAK	%	5	E	U	e	u
0 1 1 0	6	ACK	SYN	&	6	F	V	f	v
0 1 1 1	7	BEL	ETB	'	7	G	W	g	w
1 0 0 0	8	BS	CAN	(	8	H	X	h	x
1 0 0 1	9	HT	EM	)	9	I	Y	i	y
1 0 1 0	10	LF	SUB	*	:	J	Z	j	z
1 0 1 1	11	VT	ESC	+	;	K	[	k	{
1 1 0 0	12	FF	FS	,	<	L	\	l	|
1 1 0 1	13	CR	GS	-	=	M	]	m	}
1 1 1 0	14	SO	RS	.	>	N	^	n	~
1 1 1 1	15	SI	US	/	?	O	_	o	DEL



Text-based formats - newline representations

CR - carriage return - \r

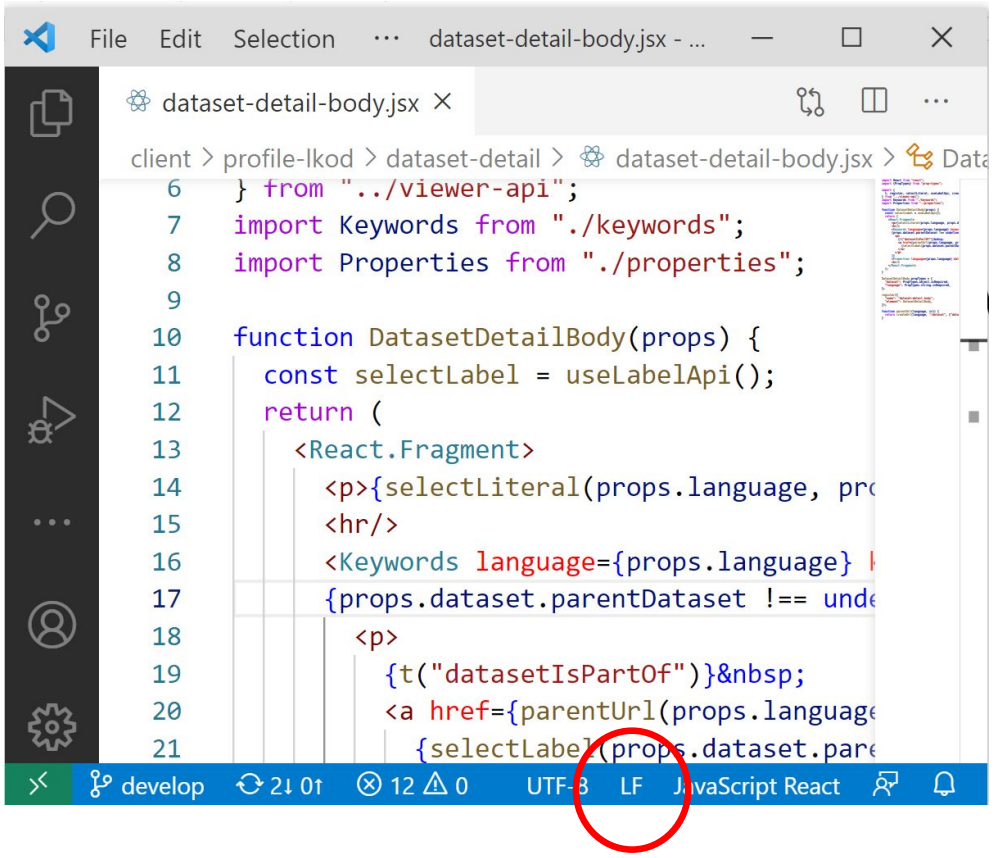
LF - line feed - \n - Unix/Linux, MacOS

CR LF - both of them - \r\n - Windows

[See all variants (Wikipedia)]

USASCII code chart

Bits b_4	b_3	b_2	b_1	Row ↓ / Column → 0	0	1	2	3	4	5	6	7
0	0	0	0	0	NUL	DLE	SP	0	@	P	`	p
0	0	0	1	1	SOH	DC1	!	1	A	Q	a	q
0	0	1	0	2	STX	DC2	"	2	B	R	b	r
0	0	1	1	3	ETX	DC3	#	3	C	S	c	s
0	1	0	0	4	EOT	DC4	$	4	D	T	d	t
0	1	0	1	5	ENQ	NAK	%	5	E	U	e	u
0	1	1	0	6	ACK	SYN	&	6	F	V	f	v
0	1	1	1	7	BEL	ETB	'	7	G	W	g	w
1	0	0	0	8	BS	CAN	(	8	H	X	h	x
1	0	0	1	9	HT	EM	)	9	I	Y	i	y
1	0	1	0	10	LF	SUB	*	:	J	Z	j	z
1	0	1	1	11	VT	ESC	+	;	K	[	k	{
1	1	0	0	12	FF	FS	,	<	L	\	l	|
1	1	0	1	13	CR	GS	-	=	M	]	m	}
1	1	1	0	14	SO	RS	.	>	N	^	n	~
1	1	1	1	15	SI	US	/	?	O	_	o	DEL



Text-based formats - newline representations

`CR` - carriage return - `\r`

`LF` - line feed - `\n` - Unix/Linux, MacOS

`CR LF` - both of them - `\r\n` - Windows

[See all variants (Wikipedia)]

USASCII code chart

Bits $b_7 b_6 b_5$ →	0 0 0	0 0 1	0 1 0	0 1 1	1 0 0	1 0 1	1 1 0	1 1 1
b_4 b_3 b_2 b_1 ↓ Row / Column →	0	1	2	3	4	5	6	7
0 0 0 0	0	NUL	DLE	SP	0	@	P	`	p
0 0 0 1	1	SOH	DC1	!	1	A	Q	a	q
0 0 1 0	2	STX	DC2	"	2	B	R	b	r
0 0 1 1	3	ETX	DC3	#	3	C	S	c	s
0 1 0 0	4	EOT	DC4	$	4	D	T	d	t
0 1 0 1	5	ENQ	NAK	%	5	E	U	e	u
0 1 1 0	6	ACK	SYN	&	6	F	V	f	v
0 1 1 1	7	BEL	ETB	'	7	G	W	g	w
1 0 0 0	8	BS	CAN	(	8	H	X	h	x
1 0 0 1	9	HT	EM	)	9	I	Y	i	y
1 0 1 0	10	LF	SUB	*	:	J	Z	j	z
1 0 1 1	11	VT	ESC	+	;	K	[	k	{
1 1 0 0	12	FF	FS	,	<	L	\	l	|
1 1 0 1	13	CR	GS	-	=	M	]	m	}
1 1 1 0	14	SO	RS	.	>	N	^	n	~
1 1 1 1	15	SI	US	/	?	O	_	o	DEL

http://penguinpetes.com/

Text-based formats - character encoding - UTF-8

1 to 4 bytes representing one character

first byte compatible with US-ASCII

most frequently used characters use 2 bytes

emojis use 4 bytes

Number of bytes	First code point	Last code point	Byte 1	Byte 2	Byte 3	Byte 4
1	U+0000	U+007F	0xxxxxxx			
2	U+0080	U+07FF	110xxxxx	10xxxxxx		
						)xxxxxx

```
:metrikaDostupnostiCORSPodmínekUžitíZvláštníPrávoPořizovateleDatabáze
    a dqv:Metric ;
    skos:definition "It checks if CORS is available for the value of pu:databáze-chráněná-zvláštními-
právy."@en , "Kontroluje, zda je pro hodnotu pu:datab��ze-chráněná-zvláštními-právy dostupná technika
CORS."@cs;
    dqv:inDimension ldqd:availability ;
    dqv:expectedDataType xsd:boolean
    .
```

á

Character encoding - BOM - Byte order mark

Magic number at the beginning of a text file

Indicates

- unicode encoding
- encoding type
  - UTF-8 - `EF BB BF`
  - UTF-16 BE - `FE FF`
  - UTF-16 LE - `FF FE`
  - UTF-32 `00 00 FE FF`
  - more...
- byte order (endianness) for multi-byte encodings

Most data formats use UTF-8 **without BOM**

- since other UTF variants are rarely used (on the Web)

	00 01 02 03 04 05 06 07 08 09 0A 0B 0C 0D 0E 0F	DECODED TEXT
00000000	EF BB BF 74 65 73 74 +	ï » ¿ t e s t +

Character encoding - other encodings

- 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
  - 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

In Czechia, from legacy systems mainly

- iso-8859-2 (Latin 2)
- windows-1250

- 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

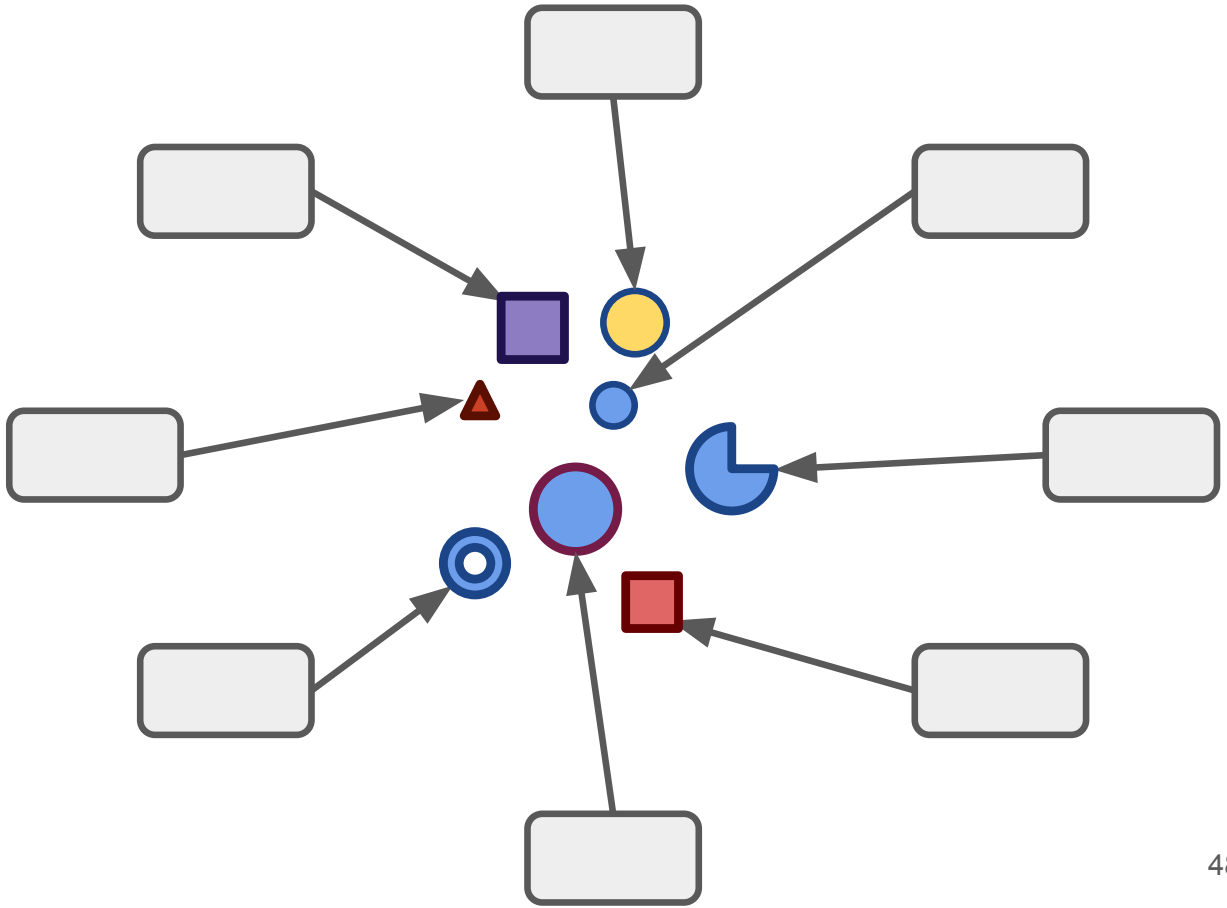
  Taiwan Big5 (a more famous variant is Microsoft Code page 950)
  - Hong Kong HKSCS
- Korean
  - KS X 1001 is a Korean double-byte character encoding standard
  - EUC-KR
  - ISO-2022-KR
- Unicode (and subsets thereof, such as the 16-bit 'Basic Multilingual Plane')
  - UTF-8
  - UTF-16
  - UTF-32
- ANSEL or ISO/IEC 6937

Standardization

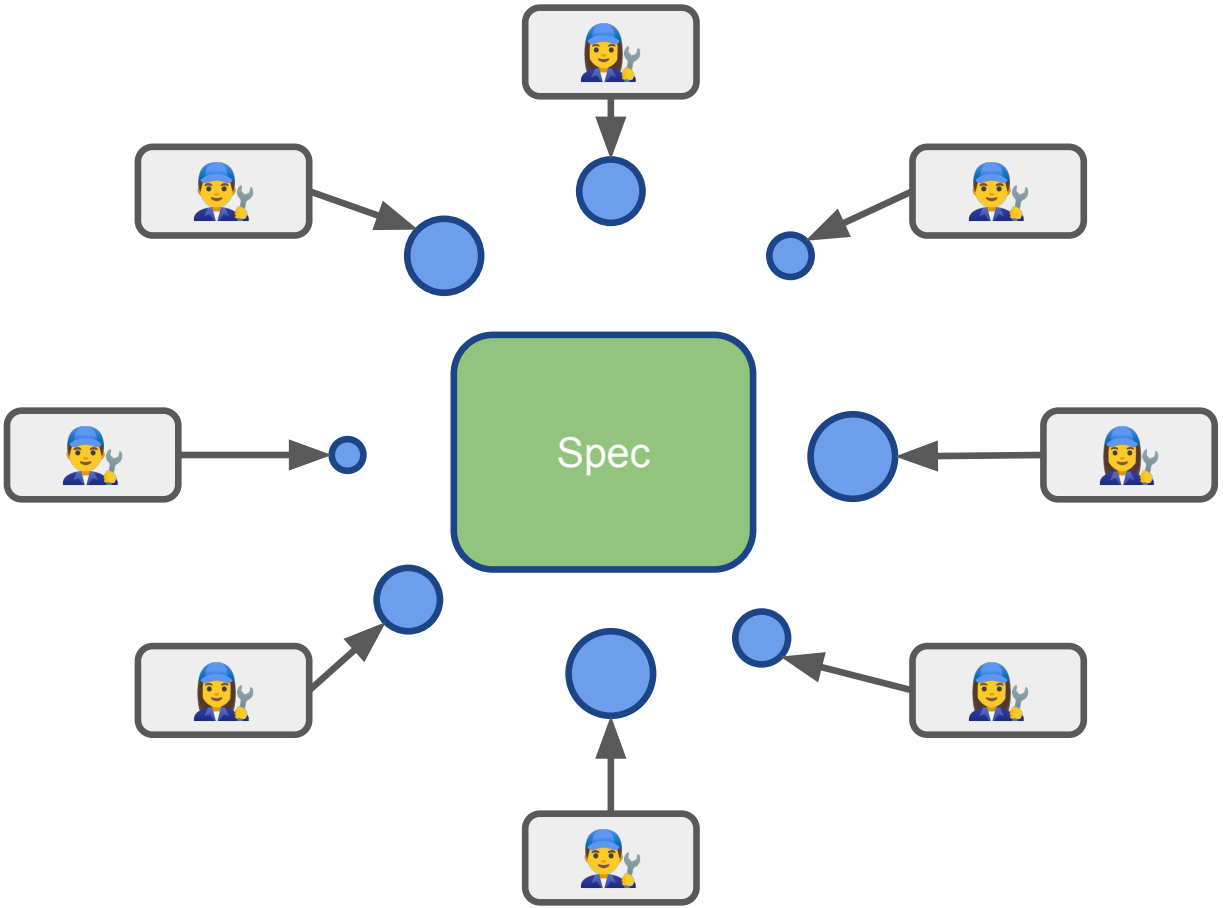
Standards - for data formats and other things

Why do we need standards?

- Interoperability, naturally
- But also business
  - So it is clear who is doing something wrong...
  - … and who will pay to fix it

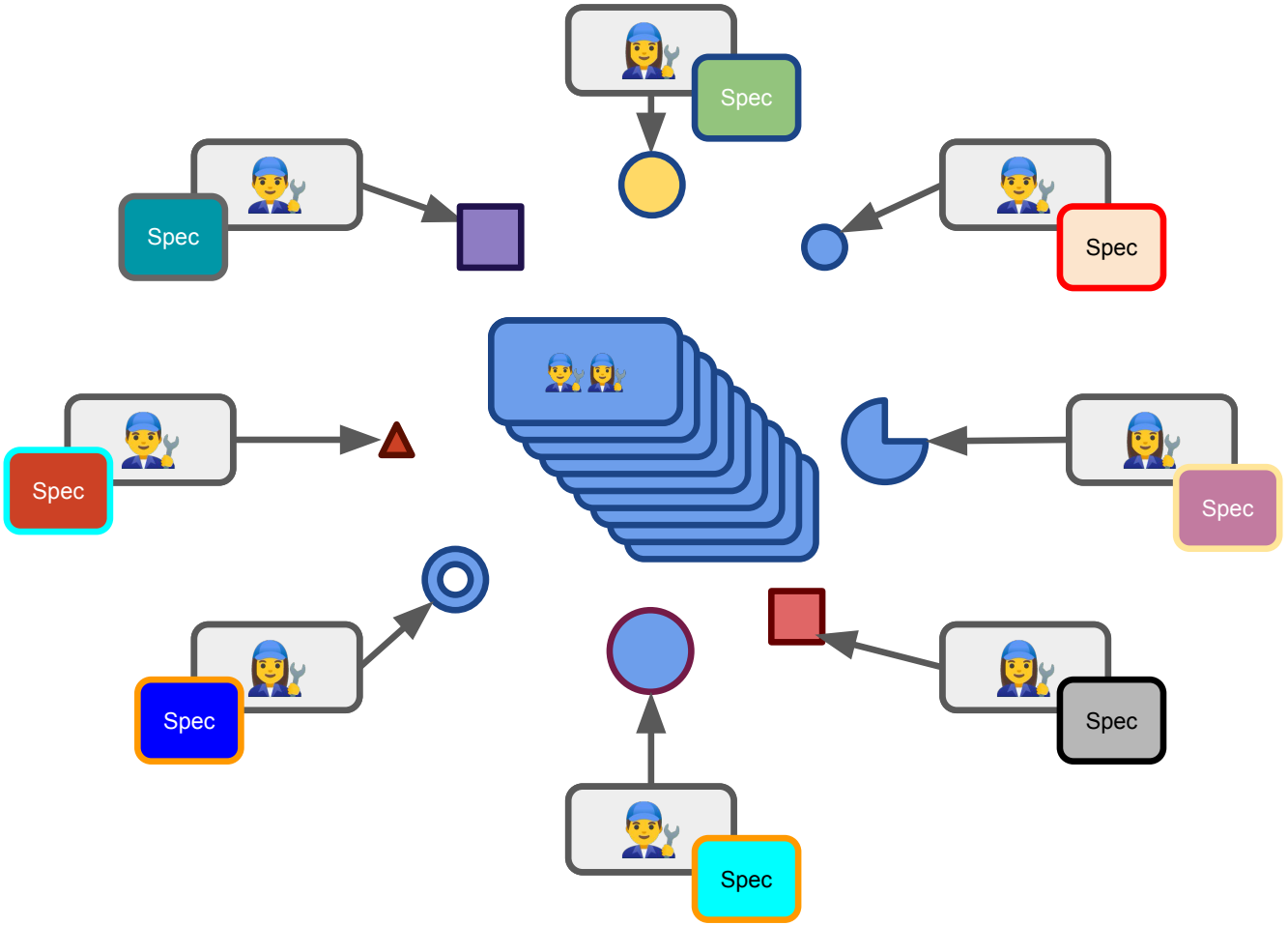


Interoperability is costly. For each dataset:



- 1 specification created
- each provider needs to learn the specification
- each provider needs to adjust their data publication process
- each consumer learns 1 specification to process all data

Low interoperability is even costlier! For each dataset:



- each provider creates specification
- each provider needs to learn the specification
- each provider needs to adjust their data publication process
- each consumer learns all specifications to process all data

Internet Engineering Task Force - IETF

Open standards organization

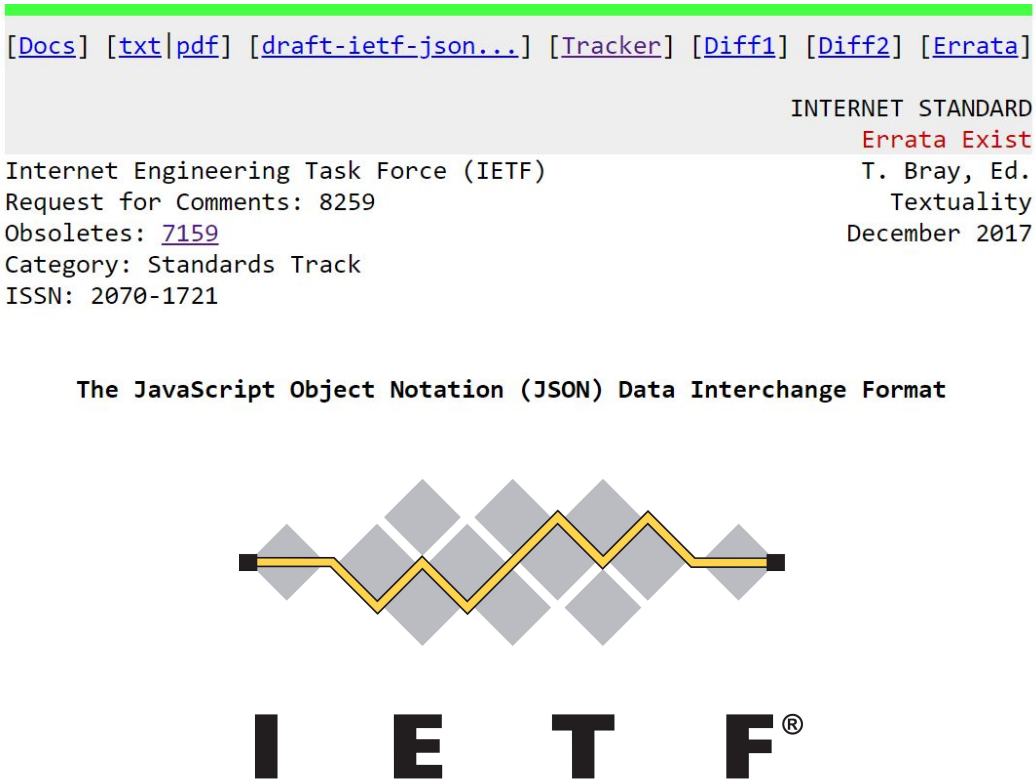
- founded 1986
- initially supported by the US federal government
- now under ISOC
- participants are volunteers

IETF Working Groups

- topic, chairperson, charter, focus, deadline
- open to all

Internet Engineering Steering Group (IESG)

- final technical review of Internet standards



[Docs] [txt|pdf] [draft-ietf-json...] [Tracker] [Diff1] [Diff2] [Errata]

INTERNET STANDARD
Errata Exist

Internet Engineering Task Force (IETF) — T. Bray, Ed.
Request for Comments: 8259 — Textuality
Obsoletes: 7159 — December 2017
Category: Standards Track
ISSN: 2070-1721

The JavaScript Object Notation (JSON) Data Interchange Format

IETF®

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

Internet Society - ISOC

Americat non-profit

- founded 1992
- to provide leadership in Internet-related standards, education, access, and policy
- deals mainly with political issues
- standards are created by the Internet Engineering Task Force (IETF) to which ISOC is related
- RFC - Request for Comments
  - some become Internet Standards

“to promote the open development, evolution, and use of the Internet for the benefit of all people throughout the world”

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

World Wide Web Consortium - W3C

International standards organization for the WWW

- founded 1994
- by Tim Berners-Lee - inventor of the Web
- issues Recommendations
  - e.g. HTML, CSS, RDF, XML, RSS…

Specification maturation process

1. Working draft (WD)
2. Candidate recommendation (CR)
3. Proposed recommendation (PR)
4. W3C recommendation (REC)

- membership is paid and must be approved
  - universities, non-profits, businesses, governments, individuals

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

Internet Corporation for Assigned Names and Numbers - ICANN

Standards organization

- founded 1998
- manages IANA - Internet Assigned Numbers Authority
- IPv4 and IPv6 address space management
- autonomous system number allocation
- root zone management in DNS
- …
- media types

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

MIME-Type, Media-type

Multipurpose Internet Mail Extensions (MIME) type

The list: Media Types

Managed by
Internet Assigned Numbers Authority (IANA)

Examples:

- text/html
- text/xml
- application/xml
- application/soap+xml
  - + suffix - specifies serialization - e.g. +xml, +json, +zip
- application/vnd.openxmlformats-officedocument.wordprocessingml.document
  - vnd. - publicly available products, e.g. Microsoft Office
- text/x-turtle
  - x- & x. - should not be used - experimental, deprecated, local, etc.

Ecma International

Standards organization

- founded 1961
- membership-based
  - IT companies, IT trade associations, universities, foundations and public institutions
- rebranded in 1994 from European Computer Manufacturers Association (ECMA)
- HQ: Geneva, Switzerland

Examples:

- ECMA-262 – ECMAScript Language Specification (based on JavaScript)
- ECMA-334 – C# Language Specification
- ECMA-376 – Office Open XML
- ECMA-404 – JSON

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

RFC 2119 - Key words for use in RFCs to Indicate Requirement Levels

`MUST, REQUIRED, SHALL`

- an absolute requirement of the specification

`MUST NOT, SHALL NOT`

- an absolute prohibition of the specification

`SHOULD, RECOMMENDED`

- there may exist valid reasons in particular circumstances to ignore a particular item
- full implications must be understood and carefully weighed before choosing a different course

`SHOULD NOT, NOT RECOMMENDED`

- there may exist valid reasons in particular circumstances when the particular behavior is acceptable or even useful
- full implications should be understood and the case carefully weighed before implementing any behavior described with this label

RFC 5234 - Augmented Backus-Naur Form (ABNF)

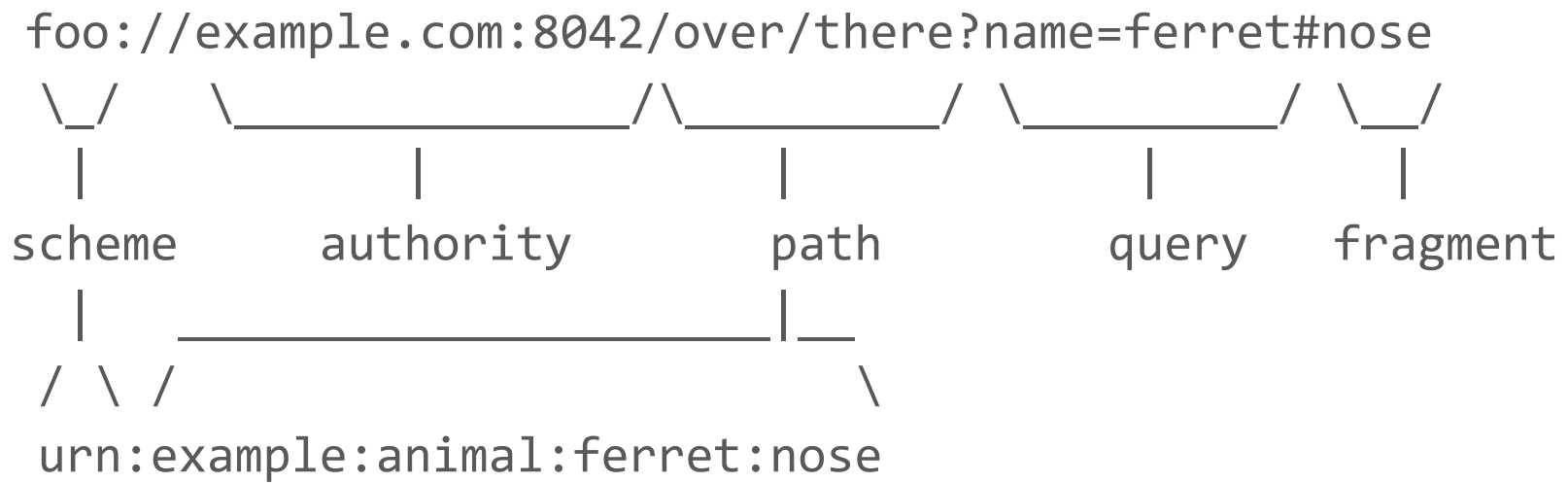
Example

```
fragment      = *( pchar / "/" / "?" )
pchar         = unreserved / pct-encoded / sub-delims / ":" / "@"
pct-encoded   = "%" HEXDIG HEXDIG
unreserved    = ALPHA / DIGIT / "-" / "." / "_" / "~"
sub-delims    = "!" / "$" / "&" / "'" / "(" / ")" / "*" / "+" / "," / ";" / "="
```

Identifiers

URI, URL, IRI, URN

URI - Uniform Resource Identifier - RFC 3986
URN - Uniform Resource Name - RFC 8141, IANA URN namespace registry
URL - Uniform Resource Locator - RFC 3986
IRI - Internationalized Resource Identifier - RFC 3987



RFC 3986 - Uniform Resource Identifier - examples

- `ftp://ftp.is.co.za/rfc/rfc1808.txt`
- `http://www.ietf.org/rfc/rfc2396.txt`
- `ldap://[2001:db8::7]/c=GB?objectClass?one`
- `mailto:John.Doe@example.com`
- `news:comp.infosystems.www.servers.unix`
- `tel:+1-816-555-1212`
- `telnet://192.0.2.16:80/`
- `urn:oasis:names:specification:docbook:dtd:xml:4.1.2`

RFC 3987 - IRI - Internationalized Resource Identifier

Examples

- https://opendata.gov.cz/špatná-praxe:start
- https://linked.opendata.cz/zdroj/💩
- https://en.wiktionary.org/wiki/Ῥόδος

Percent-encoding

- For some usages only URIs are acceptable
  - e.g. HTTP
- IRIs are encoded in URIs
- each byte represented as `'%'` and two hexadecimal digits
- e.g. 💩 => **`%F0%9F%92%A9`**
  - emojis are 4 bytes in UTF-8

The same examples of IRIs percent-encoded into URIs

- https://opendata.gov.cz/%C5%A1patn%C3%A1-praxe:start
- https://linked.opendata.cz/zdroj/%F0%9F%92%A9
- https://en.wiktionary.org/wiki/%E1%BF%AC%CF%8C%CE%B4%CE%BF%CF%82

Tip: Copy Unicode URLs - Chrome Web Store

RFC 3492 - Punycode

IRIs not to be confused with IDN - internationalized domain name:

- e.g. https://www.háčkyčárky.cz = https://www.xn--hkyrky-ptac70bc.cz/
- even less readable than percent-encoding
- punycoded name is used with DNS

Data types

Common data types in text-based structured data formats

The same data types used in all common formats - RDF syntaxes, XML, JSON, CSV
Based on XML Schema data type system

- boolean
  - `true`
  - `false`
- number
  - integer
    - `42`
  - decimal
    - `42.42`
  - float/double
    - `4.2e2`

- date - ISO-8601-compliant
  - `YYYY-MM-DD`
  - `2021-03-01`
- time
  - `HH:MM:SS.sss`
  - `10:40:00`
- dateTime
  - `YYYY-MM-DDTHH:MM:SS.sss`
  - `2021-03-01T10:40:00`
- time zones
  - `2021-03-01T10:40:00+02:00`
  - `2021-03-01-02:00`
  - `2021-03-01Z`