03) Information und Daten

Im allgemeinen Sprachgebrauch wird „Information“ mit „Bedeutung“ oder „Wissen“ gleichgesetzt. Im Vergleich dazu sind „Daten“ Angaben zu „Sachverhalten“ und „Vorgängen“. Daten sind also Werte und Inhalte, die eine Information darstellen können.

Informationen und Daten haben keinen Ort. Sie können jederzeit von einem ortsgebundenen materiellen Träger zu einem anderen wandern. Das bedeutet auch, dass sich Informationen nicht zweifelsfrei lokalisieren lassen.

Gesetze und Regeln, die Informationen und Daten an einen Ort binden, scheinen nützlich und sinnvoll zu sein. Sie sind aber kaum umsetzbar und damit sinnlos. Denn jede einzelne Informationseinheit kann jederzeit an jeden Ort der Welt übertragen und gespeichert werden.

Nur dann, wenn man Informationen und Daten durchgehend verschlüsselt, ist es fast egal, wo die Informationen und Daten gespeichert sind.

Information

- als Beseitigung von Unwissenheit
- als eine Nachricht, welche der Absender dem Empfänger über einen Kanal vermittelt. Die Nachricht wird dann interpretiert/verstanden.
- als darstellbar als Folge von 0 und 1
- als über Raum und Zeit sich ein physikalisches veränderliches Signal

Daten

- stellen Informationen dar!
- sind Träger von Informationen
- können in Schrift, Ton und Bild auftreten
- können analog oder digital gespeichert werden
- können steuern, nutzen oder adressieren

EDV (Elektronische Datenverarbeitung)

Die elektronische bzw. digitale Datenverarbeitung kennt im Prinzip nur zwei Zustände. Diese beiden Zustände werden als logisch „High“ und „Low“ bezeichnet und häufig als „1“ und „0“ dargestellt. Etwas, was nur diese zwei Zustände kennt, bezeichnet man als binäres System oder Binärlogik. Alle Daten, die elektronisch verarbeitet werden sollen, müssen in dieses Binärsystem übersetzt werden.

Das bedeutet, Schrift in Form von Buchstaben, Zahlen und Zeichen, und Bilder mit der Darstellung von Personen, Gegenständen oder Landschaften und jegliche andere Daten und Informationen werden als elektronisch lesbare Codierung in Form einer Folge aus „0“ und „1“ verarbeitet und gespeichert.

Codierung von Zeichen

Eine **Zeichenkodierung (englisch Character encoding, kurz Encoding)** erlaubt die eindeutige Zuordnung von Schriftzeichen (i. A. Buchstaben oder Ziffern) und Symbolen innerhalb eines Zeichensatzes. In der elektronischen Datenverarbeitung werden Zeichen über einen Zahlenwert kodiert, um sie zu übertragen oder zu speichern. Der deutsche Umlaut Ü wird zum Beispiel im ISO-8859-1-Zeichensatz mit dem Dezimalwert 220 kodiert. Im EBCDIC-Zeichensatz kodiert derselbe Wert 220 die geschweifte Klammer }. Zur richtigen Darstellung eines Zeichens muss also die Zeichenkodierung bekannt sein; der Zahlenwert allein reicht nicht aus.

Zahlenwerte aus Zeichenkodierungen lassen sich auf verschiedene Art speichern oder übertragen, z. B. als Morsezeichen, verschieden hohe Töne (Faxgerät), verschieden hohe Spannungen.

Geschichte des Character Encoding

Mit der Entwicklung des Computers begann die Umsetzung der im Grunde schon seit dem Baudot-Code verwendeten binären Zeichenkodierung in Bit-Folgen, bzw. intern meist in verschiedene elektrische Spannungswerte als Unterscheidungskriterium, ganz analog zu der bisher zur Unterscheidung der Signalwerte genutzten Tonhöhe oder Signaldauer.

Um diesen Bit-Folgen darstellbare Zeichen zuzuordnen, mussten Übersetzungstabellen, sogenannte Zeichensätze, engl. Charsets, festgelegt werden. 1963 wurde eine erste **7-Bit-Version des ASCII-Codes durch die ASA (American Standards Association)** definiert, um eine **Vereinheitlichung der Zeichenkodierung** zu erreichen. Obwohl IBM an der Definition mitgearbeitet hatte, führte man 1964 einen eigenen **8-Bit-Zeichencode EBCDIC** ein. Beide finden bis heute in der Computertechnik Verwendung.

Da für viele Sprachen jeweils unterschiedliche diakritische Zeichen benötigt werden, mit denen Buchstaben des lateinischen Schriftsystems modifiziert werden, gibt es für viele Sprachgruppen jeweils eigene Zeichensätze. Die **ISO** hat mit der **Normenreihe ISO 8859 Zeichenkodierungen für alle europäischen Sprachen** (einschließlich Türkisch) und Arabisch, Hebräisch sowie Thai standardisiert.

Das **Unicode Consortium** schließlich veröffentlichte 1991 eine erste Fassung des gleichnamigen Standards, der es sich zum Ziel gesetzt hat, alle Zeichen aller Sprachen in Codeform zu definieren. **Unicode** ist gleichzeitig die **internationale Norm ISO 10646**.

Bevor ein Text elektronisch verarbeitet wird, muss der verwendete Zeichensatz und die Zeichenkodierung festgelegt werden. Dazu dienen beispielsweise folgende Angaben:

Definition des Zeichensatzes in einer HTML-Seite

```
<meta http-equiv="Content-Type" content="text/html; charset=UTF-8">
```

Definition des Zeichensatzes in den Kopfzeilen (Headern) einer E-Mail oder eines HTTP-Pakets

```
Content-Type: text/plain; charset="ISO-8859-1"
```

ASCII - American Standard Code for Information Interchange

Der **American Standard Code for Information Interchange (ASCII, deutsch „Amerikanischer Standard-Code für den Informationsaustausch“)** ist eine **7-Bit-Zeichenkodierung**; sie entspricht der US-Variante von ISO 646 und dient als Grundlage für spätere, auf mehr Bits basierende Kodierungen für Zeichensätze.

Der ASCII-Code wurde zuerst am 17. Juni 1963 von der **American Standards Association (ASA) als Standard** ASA X3.4-1963 gebilligt und 1967/1968 wesentlich sowie zuletzt im Jahr 1986 von ihren Nachfolgeinstitutionen aktualisiert. Die Zeichenkodierung definiert **128 Zeichen**, bestehend aus **33 nicht druckbaren** sowie **95 druckbaren Zeichen**.

ASCII-Code Tabelle

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

www.VirtualUniversity.ch

Fakten

- sehr verbreiteter Standard, u.a. in PCs (Hardware-Ebene)
- von ISO genormt

- ursprünglich 7-Bit-Code, also 128 Zeichen
- davon einige nach nationalem Bedarf abgewandelt z.B. deutsche Umlaute statt [ ] { } \ |
- unterschiedliche 8-Bit-Erweiterungen mit zusätzlichen Zeichen im Bereich 128 – 255
- Erweiterungen oft problematisch bei älteren Rechnern, im Internet1) etc. → deshalb E-Mail, HTTP etc. auf 7 Bit beschränkt (→ 2. Sem.)

Probiere weitere Zeichen der ASCII-Tabelle in Microsoft Word aus. Drücke die ALT-Taste und tippe einen dezimalen ASCII Code ein! (z.B. ALT + 65)

UTF8 - UCS Transformation Format

UTF-8 (Abk. für 8-Bit UCS Transformation Format, wobei UCS wiederum Universal Character Set abkürzt) ist die am weitesten verbreitete Kodierung für Unicode-Zeichen (Unicode und UCS sind praktisch identisch). Die Kodierung wurde im September 1992 von Ken Thompson und Rob Pike bei Arbeiten am Plan-9-Betriebssystem festgelegt.

UTF-8 ist in den **ersten 128 Zeichen (Indizes 0–127) deckungsgleich mit ASCII** und eignet sich mit in der Regel nur **einem Byte Speicherbedarf** für Zeichen vieler westlicher Sprachen besonders für die Kodierung englischsprachiger Texte, die sich im Regelfall ohne Modifikation daher sogar mit nicht-UTF-8-fähigen Texteditoren ohne Beeinträchtigung bearbeiten lassen, was einen der Gründe für den Status als **De-facto-Standard-Zeichenkodierung des Internets** und damit verbundener Dokumenttypen darstellt. Im Oktober 2017 verwendeten **89,9 % aller Websites UTF-8**

In anderen Sprachen ist der Speicherbedarf in Byte pro Zeichen größer, wenn diese vom ASCII-Zeichensatz abweichen: Bereits die **deutschen Umlaute erfordern zwei Byte**, ebenso griechische oder kyrillische Zeichen. Zeichen fernöstlicher Sprachen und von Sprachen aus dem afrikanischen Raum belegen dagegen bis zu 4 Byte je Zeichen. Da die Verarbeitung von UTF-8 als Multibyte-Zeichenfolge wegen der notwendigen Analyse jedes Bytes im Vergleich zu Zeichenkodierungen mit fester Byteanzahl je Zeichen mehr Rechenaufwand und für bestimmte Sprachen auch mehr Speicherplatz erfordert, werden abhängig vom Einsatzszenario auch andere UTF-Kodierungen zur Abbildung von Unicode-Zeichensätzen verwendet: Microsoft Windows als meistgenutztes Desktop-Betriebssystem verwendet intern als Kompromiss zwischen UTF-8 und UTF-32 etwa UTF-16 Little Endian

Fakten Unicode (UTF-8, UTF-16, UTF-32)

- neuere Kodierung, ebenfalls ISO-standardisiert 1993 und heutzutage Standard
- Ziel: Berücksichtigung möglichst vieler Sprachen und ihrer Eigenheiten
- Buchstaben-, Silben-, und Ideogrammsprachen
- Schreibrichtungen (links-rechts, rechts-links, oben-unten)
- außerdem diverse Sonderzeichen, mathematisch-technische Symbole, Diakritika, geometrische Formen, Pfeile, Piktogramme u.v.m.
- dafür 16-Bit-Darstellung (erlaubt 65.536 Zeichen)
- Erweiterung auf 32 Bit für künftigen Bedarf
- Standard auf unixoiden Betriebssystemen