

# Algorithmus zur komprimierten Übertragung von Textdaten an mobile Endgeräte

Sven Reinck

17. Januar 2007

# Motivation

# Wörterbuch

# Wörterbuch

- Häufig vorkommende Buchstabengruppen werden in einem Wörterbuch gespeichert

# Wörterbuch

- Häufig vorkommende Buchstabengruppen werden in einem Wörterbuch gespeichert
- Im Text steht dann nur noch eine Referenz

# Wörterbuch

- Häufig vorkommende Buchstabengruppen werden in einem Wörterbuch gespeichert
- Im Text steht dann nur noch eine Referenz
- ASCII-Tabelle enthält viele unnötige Zeichen

# Wörterbuch

- Häufig vorkommende Buchstabengruppen werden in einem Wörterbuch gespeichert
- Im Text steht dann nur noch eine Referenz
- ASCII-Tabelle enthält viele unnötige Zeichen
- Ersparnis =  $(\text{Vorkommen} - 1) \cdot (\text{Länge} - 1)$

# Wörterbuch

- Häufig vorkommende Buchstabengruppen werden in einem Wörterbuch gespeichert
- Im Text steht dann nur noch eine Referenz
- ASCII-Tabelle enthält viele unnötige Zeichen
- Ersparnis =  $(\text{Vorkommen} - 1) \cdot (\text{Länge} - 1)$
- Zweier- und Dreier-Gruppen sind am ergiebigsten

# Wörterbuch

- Häufig vorkommende Buchstabengruppen werden in einem Wörterbuch gespeichert
- Im Text steht dann nur noch eine Referenz
- ASCII-Tabelle enthält viele unnötige Zeichen
- Ersparnis =  $(\text{Vorkommen} - 1) \cdot (\text{Länge} - 1)$
- Zweier- und Dreier-Gruppen sind am ergiebigsten
- Einschränkung auf Zweier-Gruppen mit Rekursion

# Statistiken aus Beispieltext

fischers fritz fischt frische fische

fischers fritz fischt frische fische

Paar	Vorkommen	Ersparnis
'sc'	4	3
'is'	4	3
'ch'	4	3
' f'	4	3
'he'	3	2
'fi'	3	2
'ri'	2	1
'fr'	2	1

Code	Ersetzung	Rekursiv
0	'sc'	'sc'

fi0hers fritz fi0ht fri0he fi0he

Paar	Vorkommen	Ersparnis
'i0'	4	3
'0h'	4	3
' f'	4	3
'he'	3	2
'fi'	3	2
'ri'	2	1
'fr'	2	1

Code	Ersetzung	Rekursiv
0	'sc'	'sc'
1	'i0'	'isc'

f1hers fritz f1ht fr1he f1he

Paar	Vorkommen	Ersparnis
'1h'	4	3
' f'	4	3
'he'	3	2
'f1'	3	2
'fr'	2	1

Code	Ersetzung	Rekursiv
0	'sc'	'sc'
1	'i0'	'isc'
2	'1h'	'isch'

f2ers fritz f2t fr2e f2e

Paar	Vorkommen	Ersparnis
' f'	4	3
'f2'	3	2
'2e'	3	2
'fr'	2	1

Code	Ersetzung	Rekursiv
0	'sc'	'sc'
1	'i0'	'isc'
2	'1h'	'isch'
3	' f'	' f'

f2ers3ritz32t3r2e32e

Paar	Vorkommen	Ersparnis
'2e'	3	2
'3r'	2	1
'32'	2	1

Code	Ersetzung	Rekursiv
0	'sc'	'sc'
1	'i0'	'isc'
2	'1h'	'isch'
3	' f'	' f'
4	'2e'	'ische'

f4rs3ritz32t3r434

Paar	Vorkommen	Ersparnis
'3r'	2	1

Code	Ersetzung	Rekursiv
0	'sc'	'sc'
1	'i0'	'isc'
2	'1h'	'isch'
3	' f'	' f'
4	'2e'	'ische'
5	'3r'	' fr'

f4rs5itz32t5434

## 1. Abschnitt

Methode	Overhead	Daten (%)	Gesamt (%)
unkomprimiert	0	2359 (100%)	2359 (100%)

## ganzer Text

Methode	Overhead	Daten (%)	Gesamt (%)
unkomprimiert	0	16896 (100%)	16896 (100%)

## 1. Abschnitt

Methode	Overhead	Daten (%)	Gesamt (%)
unkomprimiert	0	2359 (100%)	2359 (100%)
Wörterbuch	449	848 (36%)	1297 (55%)

## ganzer Text

Methode	Overhead	Daten (%)	Gesamt (%)
unkomprimiert	0	16896 (100%)	16896 (100%)
Wörterbuch	437	7899 (47%)	8336 (49%)

# Huffman-Codierung

# Huffman-Codierung

- unterschiedliche Häufigkeiten

# Huffman-Codierung

- unterschiedliche Häufigkeiten
- häufige Buchstaben → kurze Codierung

# Huffman-Codierung

- unterschiedliche Häufigkeiten
- häufige Buchstaben → kurze Codierung
- seltene Buchstaben → lange Codierung

# Statistiken aus Beispieltext

Mississippi

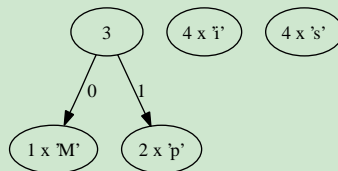
1 x 'M'

2 x 'p'

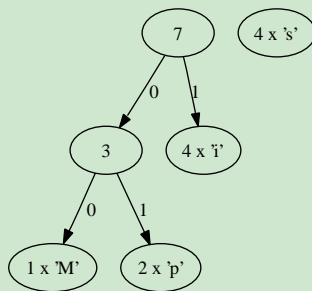
4 x 'i'

4 x 's'

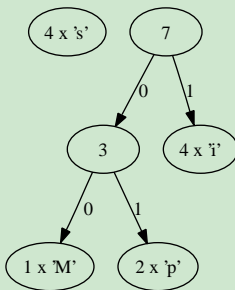
Mississippi



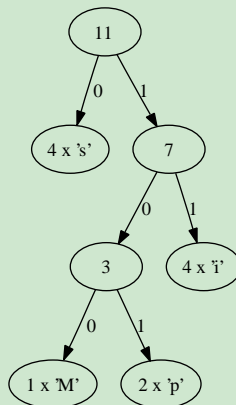
Mississippi



## Mississippi



## Mississippi



## Tabelle

Zeichen	Bits
's'	0
'M'	100
'p'	101
'i'	11

## Beispiel

M i s s i s s i p p i

## Tabelle

Zeichen	Bits
's'	0
'M'	100
'p'	101
'i'	11

## Beispiel

M i s s i s s i p p i  
100

## Tabelle

Zeichen	Bits
's'	0
'M'	100
'p'	101
'i'	11

## Beispiel

M i s s i s s i p p i

100 11

## Tabelle

Zeichen	Bits
's'	0
'M'	100
'p'	101
'i'	11

## Beispiel

M	i	s	s	i	s	s	i	p	p	i
100	11	0								

## Tabelle

Zeichen	Bits
's'	0
'M'	100
'p'	101
'i'	11

## Beispiel

M	i	s	s	i	s	s	i	p	p	i
100	11	0	0							

## Tabelle

Zeichen	Bits
's'	0
'M'	100
'p'	101
'i'	11

## Beispiel

M	i	s	s	i	s	s	i	p	p	i
100	11	0	0	11						

## Tabelle

Zeichen	Bits
's'	0
'M'	100
'p'	101
'i'	11

## Beispiel

M	i	s	s	i	s	s	i	p	p	i
100	11	0	0	11	0					

## Tabelle

Zeichen	Bits
's'	0
'M'	100
'p'	101
'i'	11

## Beispiel

M	i	s	s	i	s	s	i	p	p	i
100	11	0	0	11	0	0				

## Tabelle

Zeichen	Bits
's'	0
'M'	100
'p'	101
'i'	11

## Beispiel

M	i	s	s	i	s	s	i	p	p	i
100	11	0	0	11	0	0	11			

## Tabelle

Zeichen	Bits
's'	0
'M'	100
'p'	101
'i'	11

## Beispiel

M	i	s	s	i	s	s	i	p	p	i
100	11	0	0	11	0	0	11	101		

## Tabelle

Zeichen	Bits
's'	0
'M'	100
'p'	101
'i'	11

## Beispiel

M	i	s	s	i	s	s	i	p	p	i
100	11	0	0	11	0	0	11	101	101	

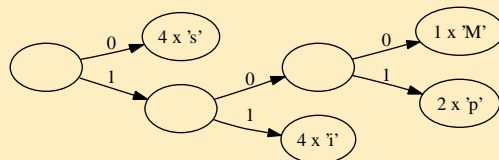
## Tabelle

Zeichen	Bits
's'	0
'M'	100
'p'	101
'i'	11

## Beispiel

M	i	s	s	i	s	s	i	p	p	i
100	11	0	0	11	0	0	11	101	101	11

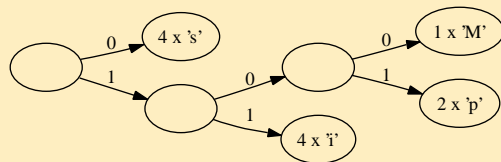
## Huffman-Baum



## Beispiel

100 11 0 0 11 0 0 11 101 101 11

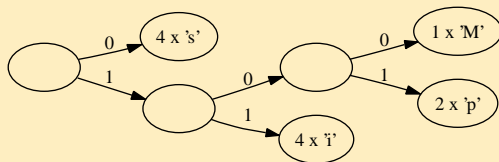
## Huffman-Baum



## Beispiel

100 11 0 0 11 0 0 11 101 101 11  
M

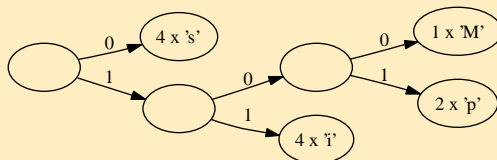
## Huffman-Baum



## Beispiel

100	11	0	0	11	0	0	11	101	101	11
M	i									

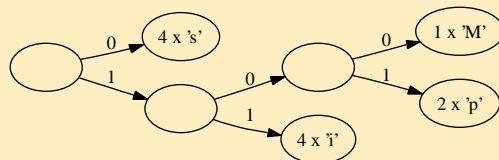
## Huffman-Baum



## Beispiel

100	11	0	0	11	0	0	11	101	101	11
M	i	s								

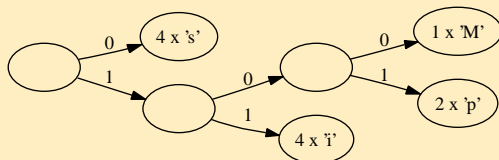
## Huffman-Baum



## Beispiel

100	11	0	0	11	0	0	11	101	101	11
M	i	s	s							

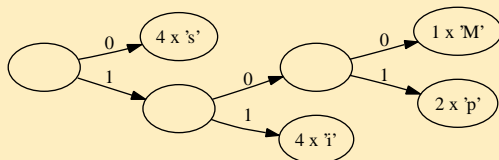
## Huffman-Baum



## Beispiel

100	11	0	0	11	0	0	11	101	101	11
M	i	s	s	i						

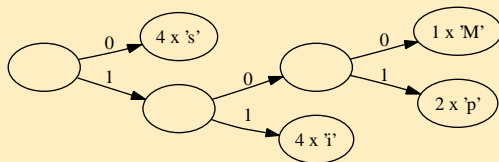
## Huffman-Baum



## Beispiel

100	11	0	0	11	0	0	11	101	101	11
M	i	s	s	i	s					

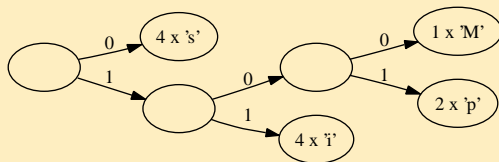
## Huffman-Baum



## Beispiel

100	11	0	0	11	0	0	11	101	101	11
M	i	s	s	i	s	s				

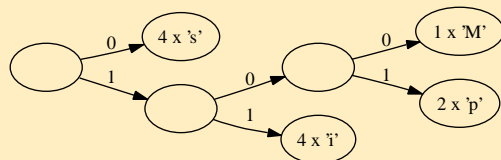
## Huffman-Baum



## Beispiel

100	11	0	0	11	0	0	11	101	101	11
M	i	s	s	i	s	s	i			

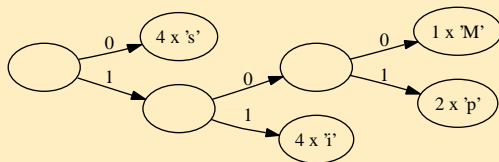
## Huffman-Baum



## Beispiel

100	11	0	0	11	0	0	11	101	101	11
M	i	s	s	i	s	s	i	p		

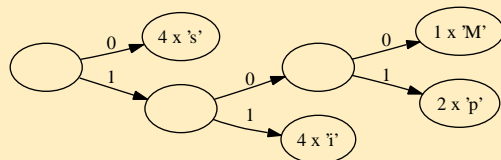
## Huffman-Baum



## Beispiel

100	11	0	0	11	0	0	11	101	101	11
M	i	s	s	i	s	s	i	p	p	

## Huffman-Baum



## Beispiel

100	11	0	0	11	0	0	11	101	101	11
M	i	s	s	i	s	s	i	p	p	i

## Tabelle

Zeichen	Länge	Bits
's'	1	0
'M'	3	100
'p'	3	101
'i'	2	11

## Tabelle

Zeichen	Länge	Bits
's'	1	
'M'	3	
'p'	3	
'i'	2	

## Tabelle

Zeichen	Länge	Bits
'M'	3	
'p'	3	
'i'	2	
's'	1	

## Tabelle

Zeichen	Länge	Bits
'M'	3	000
'p'	3	
'i'	2	
's'	1	

## Tabelle

Zeichen	Länge	Bits
'M'	3	000
'p'	3	001
'i'	2	
's'	1	

## Tabelle

Zeichen	Länge	Bits
'M'	3	000
'p'	3	001
'i'	2	01
's'	1	

## Tabelle

Zeichen	Länge	Bits
'M'	3	000
'p'	3	001
'i'	2	01
's'	1	1

## Tabelle

Zeichen	Länge	Bits
'M'	3	000
'p'	3	001
'i'	2	01
's'	1	1

## Beispiel

M i s s i s s i p p i

## Tabelle

Zeichen	Länge	Bits
'M'	3	000
'p'	3	001
'i'	2	01
's'	1	1

## Beispiel

M i s s i s s i p p i  
000

## Tabelle

Zeichen	Länge	Bits
'M'	3	000
'p'	3	001
'i'	2	01
's'	1	1

## Beispiel

M i s s i s s i p p i  
000 01

## Tabelle

Zeichen	Länge	Bits
'M'	3	000
'p'	3	001
'i'	2	01
's'	1	1

## Beispiel

M	i	s	s	i	s	s	i	p	p	i
000	01	1								

## Tabelle

Zeichen	Länge	Bits
'M'	3	000
'p'	3	001
'i'	2	01
's'	1	1

## Beispiel

M	i	s	s	i	s	s	i	p	p	i
000	01	1	1							

## Tabelle

Zeichen	Länge	Bits
'M'	3	000
'p'	3	001
'i'	2	01
's'	1	1

## Beispiel

M	i	s	s	i	s	s	i	p	p	i
000	01	1	1	01						

## Tabelle

Zeichen	Länge	Bits
'M'	3	000
'p'	3	001
'i'	2	01
's'	1	1

## Beispiel

M	i	s	s	i	s	s	i	p	p	i
000	01	1	1	01	1					

## Tabelle

Zeichen	Länge	Bits
'M'	3	000
'p'	3	001
'i'	2	01
's'	1	1

## Beispiel

M	i	s	s	i	s	s	i	p	p	i
000	01	1	1	01	1	1				

## Tabelle

Zeichen	Länge	Bits
'M'	3	000
'p'	3	001
'i'	2	01
's'	1	1

## Beispiel

M	i	s	s	i	s	s	i	p	p	i
000	01	1	1	01	1	1	01			

## Tabelle

Zeichen	Länge	Bits
'M'	3	000
'p'	3	001
'i'	2	01
's'	1	1

## Beispiel

M	i	s	s	i	s	s	i	p	p	i
000	01	1	1	01	1	1	01	001		

## Tabelle

Zeichen	Länge	Bits
'M'	3	000
'p'	3	001
'i'	2	01
's'	1	1

## Beispiel

M	i	s	s	i	s	s	i	p	p	i
000	01	1	1	01	1	1	01	001	001	

## Tabelle

Zeichen	Länge	Bits
'M'	3	000
'p'	3	001
'i'	2	01
's'	1	1

## Beispiel

M	i	s	s	i	s	s	i	p	p	i
000	01	1	1	01	1	1	01	001	001	01

## Tabelle

Zeichen	Länge	Bits
'M'	3	000
'p'	3	001
'i'	2	01
's'	1	1

## Header

## Tabelle

Zeichen	Länge	Bits
'M'	3	000
'p'	3	001
'i'	2	01
's'	1	1

## Header

3

## Tabelle

Zeichen	Länge	Bits
'M'	3	000
'p'	3	001
'i'	2	01
's'	1	1

## Header

3 2 1 1

## Tabelle

Zeichen	Länge	Bits
'M'	3	000
'p'	3	001
'i'	2	01
's'	1	1

## Header

3 2 1 1 'M' 'p' 'i' 's'

## Decodierer

--	--	--

## Codierer

Zeichen	Länge	Bits
'M'	3	000
'p'	3	001
'i'	2	01
's'	1	1

## Header

3 2 1 1 'M' 'p' 'i' 's'

## Decodierer

Länge		
3		
2		
1		

## Codierer

Zeichen	Länge	Bits
'M'	3	000
'p'	3	001
'i'	2	01
's'	1	1

## Header

3

2 1 1 'M' 'p' 'i' 's'

## Decodierer

Länge	Bits
3	000
2	01
1	1

## Codierer

Zeichen	Länge	Bits
'M'	3	000
'p'	3	001
'i'	2	01
's'	1	1

## Header

3 2 1 1

'M' 'p' 'i' 's'

## Decodierer

Länge	Bits	Zeichen
3	000	'M' 'p'
2	01	'i'
1	1	's'

## Codierer

Zeichen	Länge	Bits
'M'	3	000
'p'	3	001
'i'	2	01
's'	1	1

## Header

3 2 1 1 'M' 'p' 'i' 's'

## Decodierer

Länge	Bits	Zeichen
3	000	'M' 'p'
2	01	'i'
1	1	's'

## Beispiel

000 01 1 1 01 1 1 01 001 001 01

## Decodierer

Länge	Bits	Zeichen
3	000	'M' 'p'
2	01	'i'
1	1	's'

## Beispiel

000 01 1 1 01 1 1 01 001 001 01  
 M

## Decodierer

Länge	Bits	Zeichen
3	000	'M' 'p'
2	01	'i'
1	1	's'

## Beispiel

000 01 1 1 01 1 1 01 001 001 01  
 M i

## Decodierer

Länge	Bits	Zeichen
3	000	'M' 'p'
2	01	'i'
1	1	's'

## Beispiel

000	01	1	1	01	1	1	01	001	001	01
M	i	s								

## Decodierer

Länge	Bits	Zeichen
3	000	'M' 'p'
2	01	'i'
1	1	's'

## Beispiel

000	01	1	1	01	1	1	01	001	001	01
M	i	s	s							

## Decodierer

Länge	Bits	Zeichen
3	000	'M' 'p'
2	01	'i'
1	1	's'

## Beispiel

000	01	1	1	01	1	1	01	001	001	01
M	i	s	s	i						

## Decodierer

Länge	Bits	Zeichen
3	000	'M' 'p'
2	01	'i'
1	1	's'

## Beispiel

000	01	1	1	01	1	1	01	001	001	01
M	i	s	s	i	s					

## Decodierer

Länge	Bits	Zeichen
3	000	'M' 'p'
2	01	'i'
1	1	's'

## Beispiel

000	01	1	1	01	1	1	01	001	001	01
M	i	s	s	i	s	s				

## Decodierer

Länge	Bits	Zeichen
3	000	'M' 'p'
2	01	'i'
1	1	's'

## Beispiel

000	01	1	1	01	1	1	01	001	001	01
M	i	s	s	i	s	s	i			

## Decodierer

Länge	Bits	Zeichen
3	000	'M' 'p'
2	01	'i'
1	1	's'

## Beispiel

000	01	1	1	01	1	1	01	001	001	01
M	i	s	s	i	s	s	i	p		

## Decodierer

Länge	Bits	Zeichen
3	000	'M' 'p'
2	01	'i'
1	1	's'

## Beispiel

000	01	1	1	01	1	1	01	001	001	01
M	i	s	s	i	s	s	i	p	p	

## Decodierer

Länge	Bits	Zeichen
3	000	'M' 'p'
2	01	'i'
1	1	's'

## Beispiel

000	01	1	1	01	1	1	01	001	001	01
M	i	s	s	i	s	s	i	p	p	i

## 1. Abschnitt

Methode	Overhead	Daten (%)	Gesamt (%)
unkomprimiert	0	2359 (100%)	2359 (100%)
Wörterbuch	449	848 (36%)	1297 (55%)

## ganzer Text

Methode	Overhead	Daten (%)	Gesamt (%)
unkomprimiert	0	16896 (100%)	16896 (100%)
Wörterbuch	437	7899 (47%)	8336 (49%)

## 1. Abschnitt

Methode	Overhead	Daten (%)	Gesamt (%)
unkomprimiert	0	2359 (100%)	2359 (100%)
Wörterbuch	449	848 (36%)	1297 (55%)
Wb. + Huffman	672	773 (33%)	1445 (61%)

## ganzer Text

Methode	Overhead	Daten (%)	Gesamt (%)
unkomprimiert	0	16896 (100%)	16896 (100%)
Wörterbuch	437	7899 (47%)	8336 (49%)
Wb. + Huffman	705	7282 (43%)	7987 (47%)

# vorhersagendes Codieren

- bisher: kein Kontext

# vorhersagendes Codieren

- bisher: kein Kontext
- In einigen Kontexten können nur bestimmte Zeichen folgen

# vorhersagendes Codieren

- bisher: kein Kontext
- In einigen Kontexten können nur bestimmte Zeichen folgen
- Kombination mit Wörterbuch

# vorhersagendes Codieren

- bisher: kein Kontext
- In einigen Kontexten können nur bestimmte Zeichen folgen
- Kombination mit Wörterbuch
- nur das letzte Zeichen eines Eintrags wird verwendet

# Statistiken aus Beispieltext

## Tabellen

Anfang	M	i	p	s
'M'	'i'	'p'   0	'i'   0	'i'   0
		's'   1	'p'   1	's'   1

## Beispiel

M	i	s	s	i	s	s	i	p	p	i
000	01	1	1	01	1	1	01	001	001	01

## Tabellen

Anfang	M	i	p	s
'M'	'i'	'p'   0	'i'   0	'i'   0
		's'   1	'p'   1	's'   1

## Beispiel

M	i	s	s	i	s	s	i	p	p	i
000	01	1	1	01	1	1	01	001	001	01
x										

## Tabellen

Anfang	M	i	p	s
'M'	'i'	'p'   0	'i'   0	'i'   0
		's'   1	'p'   1	's'   1

## Beispiel

M	i	s	s	i	s	s	i	p	p	i
000	01	1	1	01	1	1	01	001	001	01
x	x									

## Tabellen

Anfang	M	i	p	s
'M'	'i'	'p'   0	'i'   0	'i'   0
		's'   1	'p'   1	's'   1

## Beispiel

M	i	s	s	i	s	s	i	p	p	i
000	01	1	1	01	1	1	01	001	001	01
x	x	1								

## Tabellen

Anfang	M	i	p	s
'M'	'i'	'p'   0	'i'   0	'i'   0
		's'   1	'p'   1	's'   1

## Beispiel

M	i	s	s	i	s	s	i	p	p	i
000	01	1	1	01	1	1	01	001	001	01
x	x	1	1							

## Tabellen

Anfang	M	i	p	s
'M'	'i'	'p'   0	'i'   0	'i'   0
		's'   1	'p'   1	's'   1

## Beispiel

M	i	s	s	i	s	s	i	p	p	i
000	01	1	1	01	1	1	01	001	001	01
x	x	1	1	0						

## Tabellen

Anfang	M	i	p	s
'M'	'i'	'p'   0	'i'   0	'i'   0
		's'   1	'p'   1	's'   1

## Beispiel

M	i	s	s	i	s	s	i	p	p	i
000	01	1	1	01	1	1	01	001	001	01
x	x	1	1	0	1					

## Tabellen

Anfang	M	i	p	s
'M'	'i'	'p'   0	'i'   0	'i'   0
		's'   1	'p'   1	's'   1

## Beispiel

M	i	s	s	i	s	s	i	p	p	i
000	01	1	1	01	1	1	01	001	001	01
x	x	1	1	0	1	1				

## Tabellen

Anfang	M	i	p	s
'M'	'i'	'p'   0	'i'   0	'i'   0
		's'   1	'p'   1	's'   1

## Beispiel

M	i	s	s	i	s	s	i	p	p	i
000	01	1	1	01	1	1	01	001	001	01
x	x	1	1	0	1	1	0			

## Tabellen

Anfang	M	i	p	s
'M'	'i'	'p'   0	'i'   0	'i'   0
		's'   1	'p'   1	's'   1

## Beispiel

M	i	s	s	i	s	s	i	p	p	i
000	01	1	1	01	1	1	01	001	001	01
x	x	1	1	0	1	1	0	0		

## Tabellen

Anfang	M	i	p	s
'M'	'i'	'p'   0	'i'   0	'i'   0
		's'   1	'p'   1	's'   1

## Beispiel

M	i	s	s	i	s	s	i	p	p	i
000	01	1	1	01	1	1	01	001	001	01
x	x	1	1	0	1	1	0	0	1	

## Tabellen

Anfang	M	i	p	s
'M'	'i'	'p'   0	'i'   0	'i'   0
		's'   1	'p'   1	's'   1

## Beispiel

M	i	s	s	i	s	s	i	p	p	i
000	01	1	1	01	1	1	01	001	001	01
x	x	1	1	0	1	1	0	0	1	0

## 1. Abschnitt

Methode	Overhead	Daten (%)	Gesamt (%)
unkomprimiert	0	2359 (100%)	2359 (100%)
Wörterbuch	449	848 (36%)	1297 (55%)
Wb. + Huffman	672	773 (33%)	1445 (61%)

## ganzer Text

Methode	Overhead	Daten (%)	Gesamt (%)
unkomprimiert	0	16896 (100%)	16896 (100%)
Wörterbuch	437	7899 (47%)	8336 (49%)
Wb. + Huffman	705	7282 (43%)	7987 (47%)

## 1. Abschnitt

Methode	Overhead	Daten (%)	Gesamt (%)
unkomprimiert	0	2359 (100%)	2359 (100%)
Wörterbuch	449	848 (36%)	1297 (55%)
Wb. + Huffman	672	773 (33%)	1445 (61%)
Wb. + Context	1216	464 (20%)	1680 (71%)

## ganzer Text

Methode	Overhead	Daten (%)	Gesamt (%)
unkomprimiert	0	16896 (100%)	16896 (100%)
Wörterbuch	437	7899 (47%)	8336 (49%)
Wb. + Huffman	705	7282 (43%)	7987 (47%)
Wb. + Context	2446	5253 (31%)	7699 (46%)

# Xenia: Komprimierung von XML

# Xenia: Komprimierung von XML

computerwoche.de (18.12.2006):

”Mit 'Xenia' können XML-Daten auf 1 bis 15 Prozent ihrer ursprünglichen Größe komprimiert werden”