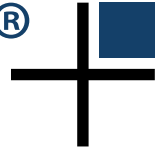


Ordnung muss sein

Peter Eisentraut

2ndQuadrant[®] 
PostgreSQL

peter.eisentraut@2ndquadrant.com
[@petereisentraut](https://twitter.com/petereisentraut)

schon, schön, schoen, Schöne

Maser, Masse, Maße

deluge, de luge, de-luge

Göbel, Göthe, Götz, Goldmann, Goethe

cote, coté, côte, côté

Unicode, ISO, DIN usw.

- Unicode
- UTR #10 (UCA)
- ISO 14651
- DIN 5007

Beispiel UCA

'abc' <=> 'def' ?

Gewichtungen

```
0061 ; [.1E89.0020.0002] # LATIN SMALL LETTER A
0062 ; [.1EA3.0020.0002] # LATIN SMALL LETTER B
0063 ; [.1EBD.0020.0002] # LATIN SMALL LETTER C
0064 ; [.1ED2.0020.0002] # LATIN SMALL LETTER D
0065 ; [.1EED.0020.0002] # LATIN SMALL LETTER E
0066 ; [.1F28.0020.0002] # LATIN SMALL LETTER F
```

Sortierschlüssel

```
'abc' => 1E89 1EA3 1EBD 0000 0020 0020 0020 0000 0002 0002 0002
'def' => 1ED2 1EED 1F28 0000 0020 0020 0020 0000 0002 0002 0002
```

Beispiel UCA

'abc' <=> 'abcd' ?

Gewichtungen

```
0061 ; [.1E89.0020.0002] # LATIN SMALL LETTER A
0062 ; [.1EA3.0020.0002] # LATIN SMALL LETTER B
0063 ; [.1EBD.0020.0002] # LATIN SMALL LETTER C
0064 ; [.1ED2.0020.0002] # LATIN SMALL LETTER D
0065 ; [.1EED.0020.0002] # LATIN SMALL LETTER E
0066 ; [.1F28.0020.0002] # LATIN SMALL LETTER F
```

Sortierschlüssel

```
'abc'  => 1E89 1EA3 1EBD 0000 0020 0020 0020 0000 0002 0002 0002
'abcd' => 1E89 1EA3 1EBD 1ED2 0000 0020 0020 0020 0020 0000 0002 0002 0002 0002
```

Beispiel

' äBc ' <=> ' déF ' ?

Gewichtungen

```
00E4 ; [.1E89.0020.0002][.0000.002B.0002] # LATIN SMALL LETTER A WITH DIAERESIS
0042 ; [.1EA3.0020.0008] # LATIN CAPITAL LETTER B
0063 ; [.1EBD.0020.0002] # LATIN SMALL LETTER C
0064 ; [.1ED2.0020.0002] # LATIN SMALL LETTER D
00E9 ; [.1EED.0020.0002][.0000.0024.0002] # LATIN SMALL LETTER E WITH ACUTE
0046 ; [.1F28.0020.0008] # LATIN CAPITAL LETTER F
```

Sortierschlüssel

```
' äBc ' => 1E89 1EA3 1EBD 0000 0020 002B 0020 0020 0000 0002 0008 0002
'déF' => 1ED2 1EED 1F28 0000 0020 0020 0024 0020 0000 0002 0002 0008
```

Beispiel Umlaut

'schon' <=> 'schön' ?

Gewichtungen

```
0063 ; [.1EBD.0020.0002] # LATIN SMALL LETTER C
0068 ; [.1F5B.0020.0002] # LATIN SMALL LETTER H
006E ; [.1FFD.0020.0002] # LATIN SMALL LETTER N
006F ; [.2021.0020.0002] # LATIN SMALL LETTER O
00F6 ; [.2021.0020.0002][.0000.002B.0002] # LATIN SMALL LETTER O WITH DIAERESIS
0073 ; [.20B6.0020.0002] # LATIN SMALL LETTER S
```

Sortierschlüssel

```
'schon' => 20B6 1EBD 1F5B 2021 1FFD 0000 0020 0020 0020 0020 0020 0000 0002 0002
           0002 0002 0002
'schön' => 20B6 1EBD 1F5B 2021 1FFD 0000 0020 0020 0020 0020 0020 002B 0020 0000 0002
           0002 0002 0002 0002 0002
```

Beispiel » ß«

'Maser', 'Masse', 'Maße'

Gewichtungen

```
004D ; [.1FEE.0020.0008] # LATIN CAPITAL LETTER M
0061 ; [.1E89.0020.0002] # LATIN SMALL LETTER A
0065 ; [.1EED.0020.0002] # LATIN SMALL LETTER E
0072 ; [.2078.0020.0002] # LATIN SMALL LETTER R
0073 ; [.20B6.0020.0002] # LATIN SMALL LETTER S
00DF ; [.20B6.0020.0004][.0000.0116.0004][.20B6.0020.0004] # LATIN SMALL LETTER SHARP S
```

Sortierschlüssel

```
'Maser' => 1FEE 1E89 20B6 1EED 2078 0000 0020 0020 0020 0020 0020 0000 0008 0002 0002 0002 0002
'Masse' => 1FEE 1E89 20B6 20B6 1EED 0000 0020 0020 0020 0020 0020 0000 0008 0002 0002 0002 0002
'Maße'  => 1FEE 1E89 20B6 20B6 1EED 0000 0020 0020 0020 0116 0020 0020 0000 0008 0002 0004 0004 0004 0002
```

Deutsche Sortierungen

Göbel, Göthe, Götz, Goldmann, Goethe

Auflösung

DIN 5007 (1)

Göbel, Goethe, Goldmann, Göthe, Götz

DIN 5007 (2)

Göbel, Goethe, Göthe, Götz, Goldmann

Österreichisch

Goethe, Goldmann, Göbel, Göthe, Götz

Die spinnen, die Gallier

COLLATE "und-x-icu"

cote, coté, côte, côté

COLLATE "fr-CA-x-icu"

cote, côte, coté, côté

Sonderzeichen

'death', 'deluge', 'de luge'

Gewichtungen

```
0020 ; [*0209.0020.0002] # SPACE
0061 ; [.1E89.0020.0002] # LATIN SMALL LETTER A
0064 ; [.1ED2.0020.0002] # LATIN SMALL LETTER D
0065 ; [.1EED.0020.0002] # LATIN SMALL LETTER E
0067 ; [.1F37.0020.0002] # LATIN SMALL LETTER G
0068 ; [.1F5B.0020.0002] # LATIN SMALL LETTER H
006C ; [.1FBB.0020.0002] # LATIN SMALL LETTER L
0074 ; [.20DA.0020.0002] # LATIN SMALL LETTER T
0075 ; [.20FA.0020.0002] # LATIN SMALL LETTER U
```

Sortierschlüssel

Option "non-ignorable"

```
'de luge' => 1ED2 1EED 0209 1FBB 20FA 1F37 1EED 0000 0020 0020 0020 0020 0020 0020 0020 0000
              0002 0002 0002 0002 0002 0002 0002
'death'   => 1ED2 1EED 1E89 20DA 1F5B 0000 0020 0020 0020 0020 0020 0000 0002 0002 0002 0002
              0002
'deluge'  => 1ED2 1EED 1FBB 20FA 1F37 1EED 0000 0020 0020 0020 0020 0020 0020 0000 0002 0002
              0002 0002 0002 0002
```

Sonderzeichen

'death', 'deluge', 'de luge'

Gewichtungen

Option "shifted"

```
0020 ; [.0000.0000.0000.0209] # SPACE
0061 ; [.1E89.0020.0002.FFFF] # LATIN SMALL LETTER A
0064 ; [.1ED2.0020.0002.FFFF] # LATIN SMALL LETTER D
0065 ; [.1EED.0020.0002.FFFF] # LATIN SMALL LETTER E
0067 ; [.1F37.0020.0002.FFFF] # LATIN SMALL LETTER G
0068 ; [.1F5B.0020.0002.FFFF] # LATIN SMALL LETTER H
006C ; [.1FBB.0020.0002.FFFF] # LATIN SMALL LETTER L
0074 ; [.20DA.0020.0002.FFFF] # LATIN SMALL LETTER T
0075 ; [.20FA.0020.0002.FFFF] # LATIN SMALL LETTER U
```

Sortierschlüssel

```
'death'    => 1ED2 1EED 1E89 20DA 1F5B 0000 0020 0020 0020 0020 0020 0000 0002 0002 0002 0002
              0002 0000 FFFF FFFF FFFF FFFF FFFF
'de luge'  => 1ED2 1EED 1FBB 20FA 1F37 1EED 0000 0020 0020 0020 0020 0020 0020 0000 0002 0002
              0002 0002 0002 0002 0000 FFFF FFFF 0209 FFFF FFFF FFFF FFFF
'deluge'   => 1ED2 1EED 1FBB 20FA 1F37 1EED 0000 0020 0020 0020 0020 0020 0020 0000 0002 0002
              0002 0002 0002 0002 0000 FFFF FFFF FFFF FFFF FFFF FFFF
```

Sonderzeichen

'death', 'deluge', 'de luge'

Gewichtungen

Option "shift-trimmed"

```
0020 ; [.0000.0000.0000.0209] # SPACE
0061 ; [.1E89.0020.0002.FFFF] # LATIN SMALL LETTER A
0064 ; [.1ED2.0020.0002.FFFF] # LATIN SMALL LETTER D
0065 ; [.1EED.0020.0002.FFFF] # LATIN SMALL LETTER E
0067 ; [.1F37.0020.0002.FFFF] # LATIN SMALL LETTER G
0068 ; [.1F5B.0020.0002.FFFF] # LATIN SMALL LETTER H
006C ; [.1FBB.0020.0002.FFFF] # LATIN SMALL LETTER L
0074 ; [.20DA.0020.0002.FFFF] # LATIN SMALL LETTER T
0075 ; [.20FA.0020.0002.FFFF] # LATIN SMALL LETTER U
```

Sortierschlüssel

```
'death'    => 1ED2 1EED 1E89 20DA 1F5B 0000 0020 0020 0020 0020 0020 0000 0002 0002 0002 0002
              0002 0000
'deluge'   => 1ED2 1EED 1FBB 20FA 1F37 1EED 0000 0020 0020 0020 0020 0020 0020 0000 0002 0002
              0002 0002 0002 0002 0000
'de luge'  => 1ED2 1EED 1FBB 20FA 1F37 1EED 0000 0020 0020 0020 0020 0020 0020 0000 0002 0002
              0002 0002 0002 0002 0000 FFFF FFFF 0209
```

Variable Weighting Optionen

- Non-ignorable
- Blanked
- Shifted
- Shift-Trimmed

Variable Weighting Optionen

- alte glibc: **shift-trimmed**
 - (CentOS 7, Debian 9)
- neue glibc: **shifted**
 - (Fedora 29, RHEL 8, Debian testing)
- ICU: **non-ignorable**
 - Option: **shifted**

Collation Provider

- libc
- ICU

libc

```
CREATE COLLATION c1 (provider = libc, locale = 'de_DE.utf8');
```

```
CREATE TABLE t1 (  
    id int PRIMARY KEY,  
    x text COLLATE c1  
);
```

ICU

```
CREATE COLLATION c1 (provider = icu, locale = 'de');  
CREATE COLLATION c1 (provider = icu, locale = 'de-AT');  
  
CREATE TABLE t1 (  
    id int PRIMARY KEY,  
    x text COLLATE c1  
);
```

Datenbankweite Sortierfolge

immer libc!

```
CREATE DATABASE d1 LC_COLLATE = 'de_DE.utf8';
```

Collation-Optionen mit ICU

```
CREATE COLLATION c1 (provider = icu, locale = 'de');
```

Hacker, Häcker, Hackmann, Haecker, Hafermann

```
CREATE COLLATION c2 (provider = icu, locale = 'de-u-co-phonebk');
```

```
CREATE COLLATION c2 (provider = icu, locale = 'de@collation=phonebook');
```

Hacker, Hackmann, Häcker, Haecker, Hafermann

Weitere ICU-Varianten

```
CREATE COLLATION c1 (provider = icu, locale = 'de');
```

bar, Bar, foo, Foo

```
CREATE COLLATION c2 (provider = icu, locale = 'de-u-kf-upper');
```

```
CREATE COLLATION c2 (provider = icu, locale = 'de@colCaseFirst=upper');
```

Bar, bar, Foo, foo

Weitere ICU-Varianten

```
CREATE COLLATION c1 (provider = icu, locale = 'de');
```

A-123, A-12n, A-21, B-100

```
CREATE COLLATION c2 (provider = icu, locale = 'de-u-kn-true');
```

```
CREATE COLLATION c2 (provider = icu, locale = 'de@colNumeric=yes');
```

A-12n, A-21, A-123, B-100

Weitere ICU-Varianten

```
CREATE COLLATION c1 (provider = icu, locale = 'und');  
CREATE COLLATION c1 (provider = icu, locale = '');
```



```
CREATE COLLATION c2 (provider = icu, locale = 'und-u-co-emoji');  
CREATE COLLATION c2 (provider = icu, locale = '@collation=emoji');
```



(UTR #51)

Case-insensitive

```
SELECT x FROM t1 WHERE lower(x) = lower('foo');
```

x
foo
Foo
FOO

citext

```
CREATE TABLE t1 (x citext);
```

```
SELECT x FROM t1 WHERE x = 'foo';
```

x
foo
Foo
FOO

Sonderfälle

```
SELECT x FROM t2;
```

x

ὀδυσσεύς

```
SELECT x FROM t2
```

```
WHERE lower(x) = lower( 'ὈΔΥΣΣΕΎΣ' );
```

```
(0 rows)
```

Nicht-deterministische Collations

(ab PostgreSQL 12)

```
CREATE COLLATION case_insensitive (provider = icu,  
    locale = 'und-u-ks-level2', deterministic = false);
```

'abc' = 'ABC'

```
CREATE COLLATION ignore_accents (provider = icu,  
    locale = 'und-u-ks-level1-kc-true', deterministic = false);
```

'äbc' = 'abc'

Aktualisierung der Locale-Daten

```
SELECT * FROM foo ORDER BY 1;
```

WARNING: collation "xx-x-icu" has version mismatch

DETAIL: The collation in the database was created using
version 1.2.3.4, but the operating system provides version 2.3.4.5.

HINT: Rebuild all objects affected by this collation and run
ALTER COLLATION pg_catalog."xx-x-icu" REFRESH VERSION,
or build PostgreSQL with the right library version.

glibc 2.27

ANDREAS
AS
A/S
CA
côlon
coop
co-op
COOP
CO-OP
Copenhagen
MÂCON
McArthur
Mc Arthur
Mc Mahon

glibc 2.28

ANDREAS
A/S
AS
CA
côlon
co-op
coop
CO-OP
COOP
Copenhagen
MÂCON
Mc Arthur
McArthur
Mc Mahon

Aktualisierung der Locale-Daten in glibc

- Collations in glibc haben keine Version
- glibc 2.28 hat generalüberholte Locale-Daten
 - alt: CentOS 7, Debian 9, Ubuntu 18.04 LTS
 - neu: RHEL 8, Debian testing, Ubuntu 18.10, Fedora 29
- 
- https://wiki.postgresql.org/wiki/Locale_data_changes

Zusammenfassung

- UCA
- libc hat Probleme.
- ICU ist die Zukunft.
- REINDEX nach OS-Update