Was bedeutet ... ?

Apache-Lizenz

Die Apache-Lizenz ist eine Freie-Software-Lizenz der Apache Software Foundation, die keinen Copyleft-Vermerk besitzt.

Beta-Code Altgriechisch

Griechischer Beta Code ist die 7-Bit-sichere Kodierung mittels des US-ASCII-Zeichensatzes. Jedes diakritische Zeichen wird durch ein eigenes Zeichen dargestellt, welches dem Buchstaben folgt (Ausnahme: bei Großbuchstaben vor dem Buchstaben). Beta Code unterscheidet nicht zwischen Klein-/Großschreibung, Großbuchstaben werden durch Voranstellung von * Asteriskos (griech. ἀστερίσκος) gekennzeichnet. Einige Projekte benutzen nur Großbuchstaben (z.B. TLG), andere nur Kleinbuchstaben (z.B. das Perseus Project).

Siehe auch: Betacode-Transkriptions-Tabelle Altgriechisch.

ἀστερίσκος in Beta-Code Altgriechisch:

```
a)steri/skos
```

Big Data

Analyse großer Datenmengen aus verschiedenen Quellen mit dem Ziel, wirtschaftlichen Nutzen daraus zu erzeugen.

CC

Unter dem Begriff **C**reative **C**ommons (CC) wird eine Sammlung von Lizenzen verstanden, mit denen ein Autor Nutzungsrechte für sein Werk einräumen kann. Durch die Kombination der Rechtemodule

- by (Attribution) Namensnennung
- nc (Non-Commercial) Nicht kommerziell
- nd (No Derivatives) Keine Bearbeitung
- sa (Share Alike) Weitergabe unter gleichen Bedingungen

kann die Freigabe nach den Wünschen des Urhebers abgestuft werden.

Copyleft

Als Copyleft wird eine Klausel in Nutzungslizenzen bezeichnet, die festlegt, dass alle Änderungen an einem Werk nur dann statthaft sind, wenn sie im Wesentlichen unter den gleichen Lizenzbedingungen verbreitet werden.

CSV

Das textbasierte Dateiformat CSV (Comma-separated values) ist eine Form von DSV (Delimiter-separated values). Die Daten sind in Tabellenform, also zweidimensional, gespeichert. Jede Zeile ist ein Datensatz. Felder werden mittels **Komma** oder **Semikolon** separiert.

Parallelstellen von TATIANUS (TLG) im CSV-Format:

beispiel.csv

```
Original Sentence; Reference; Original Author; Original Publication; Original DC; Author;
Publication; DC; Similarity; Dating; Author Name; Author Epiteths; Author ID; AuthorID-WorkID
"Τυρρηνοὶ σάλπιγγα, χαλκεύειν Κύκλωπες, καὶ ἐπιστολὰς συντάσσειν ἡ Περσῶν ποτε ἡγησαμένη γυνή, καθά φησιν
Ἕλλάνικος:";"εὗρεν) ἡ Περσῶν ποτε ἡγησαμένη γυνή, καθά φησιν Ἑλλάνικος:";"TATIANUS  Apol.  [1766]";"Oratio ad
Graecos, ed. E.J. Goodspeed, Die altesten  Apologeten. Gottingen: Vandenhoeck & Ruprecht, 1915:
268-305.  (Cod: 10,694: Apol., Orat.)  ";"1T/2/2 to 1T/2/4 (Schema:Chapter/section/line
)";"HELLANICUS  Hist.  [0539]";"Fragmenta, FGrH #4, #323a, #601a, #608a, #645a, #687a:
1A:107-152, *6-*8 addenda; 3B:41-50, 732-733; 3C:1-2, 190,  412-414.  fr. 124b (PSI 1173): vol.
1A, p. *6 addenda.  fr. 189 (P. Oxy. 10.1241): vol. 1A, p. 150.  fr. 201 bis (P. Giss. 307v):
vol. 1A, p. *7 addenda.  (Pap: 18,331: Hist., Myth.)  ";"1a,4,F/179a/3 to 1a,4,F/179a/4
(Schema:Volume-Jacoby#-F//fragment/line )";67;-450.5;"HELLANICUS ";"Hist. ";"0539";"0539-002"
"Τυρρηνοὶ σάλπιγγα, χαλκεύειν Κύκλωπες, καὶ ἐπιστολὰς συντάσσειν ἡ Περσῶν ποτε ἡγησαμένη γυνή, καθά φησιν
Ἕλλάνικος:";"εὗρεν) ἡ Περσῶν ποτε ἡγησαμένη γυνή, καθά φησιν Ἑλλάνικος:";"TATIANUS  Apol.  [1766]";"Oratio ad
Graecos, ed. E.J. Goodspeed, Die altesten  Apologeten. Gottingen: Vandenhoeck & Ruprecht, 1915:
268-305.  (Cod: 10,694: Apol., Orat.)  ";"1T/2/2 to 1T/2/4 (Schema:Chapter/section/line
)";"HELLANICUS  Hist.  [0539]";"Fragmenta, FGrH #4, #323a, #601a, #608a, #645a, #687a:
1A:107-152, *6-*8 addenda; 3B:41-50, 732-733; 3C:1-2, 190,  412-414.  fr. 124b (PSI 1173): vol.
1A, p. *6 addenda.  fr. 189 (P. Oxy. 10.1241): vol. 1A, p. 150.  fr. 201 bis (P. Giss. 307v):
vol. 1A, p. *7 addenda.  (Pap: 18,331: Hist., Myth.)  ";"3c,687a,F/8a/3 to 3c,687a,F/8a/3
(Schema:Volume-Jacoby#-F//fragment/line )";67;-450.5;"HELLANICUS ";"Hist. ";"0539";"0539-002"
```

CTS

Das Notationssystem CTS (**C**anoncial **T**ext **S**ervices) als Teil der CITE Architektur bietet einen netzbasierten Service zur Identifikation klassischer Texte basierend auf URN. CTS URNs sind in fünf Teile untergliedert, die von Doppelpunkten voneinander getrennt sind: urn:ctn:ctnNameSpace:WorkIdentifier:PassageIdentifier.

DOI

Digital **O**bject **I**dentifier (DOI) werden seit 1998 durch die International DOI Foundation (IDF) koordiniert. Mit DOI können sowohl physische, digitale als auch abstrakte Objekte dauerhaft eindeutig identifiziert und lokalisiert werden. Dem Schema, welches immer mit 10 beginnt, wird zur Identifikation eine doi vorangestellt: **doi**:10.ORGANISATION/ID.

Ein Beispiel:

```
Ch. Schubert (Hg.): Working Papers Contested Order (NO. 10): Das Portal eAQUA – Neue Methoden in der
geisteswissenschaftlichen Forschung V
DOI: http://dx.doi.org/10.11588/ea.2013.2
```

Editierdistanz

siehe Levenshtein-Distanz

Entropie

Entropie in der Informationstheorie gibt an, wieviel Bits im Durchschnitt benötigt werden, um einen Wert einer Zuvallsvariablen als ein Ereignis (als Teil einer Nachricht) zu codieren. Je mehr Bits benötigt werden, desto höher ist die Entropie und umso schwieriger die Vorhersagen eines Ereignisses.

GPL

Die GNU **G**eneral **P**ublic **L**icense (auch GPL oder GNU GPL) ist eine Lizenz, die es erlaubt, eine Software kostenlos zu nutzen, zu verbreiten, zu studieren oder auch zu verändern. Alle von der Software abgeleitete Programme müssen ebenfalls zu den Bedingungen der GPL lizenziert werden (Copyleft).

HTML

Hyper**t**ext **M**arkup **L**anguage ist eine textbasierte Auszeichnungssprache zur strukturierten Darstellung von Inhalten in elektronischen Dokumenten.

JPEG

Verschiedene Methoden der Bildkompression, die vom Gremium **J**oint **P**hotographic **E**xperts **G**roup 1992 in Form einer Norm vorgestellt wurden, werden unter dem Begriff JPEG zusammengefasst.

JSON

Java**S**cript **O**bject **N**otation ist ein kompaktes Datenformat, welches zur Übertragung von Daten zwischen Client und Server konzipiert wurde.

Auszug von TLG-Metadaten in JSON:

beispiel.json

```
{
"corpora_author_id":2064,
"author":"ACACIUS",
"works":
  [
  {"corpora_work_id":"002","work":"Fragmenta in epistulam ad Romanos (in catenis)"}
  ]
},
{
"corpora_author_id":1832,
"author":"ACESANDER",
"works":
  [
  {"corpora_work_id":"001","work":"Fragmenta "},
  {"corpora_work_id":"002","work":"Fragmentum (P. Oxy. 32.2637)"}
  ]
}
```

Kookkurrenz

Das gemeinsame Auftreten zweier lexikalischer Einheiten, z.B. Wörter, innerhalb eines übergeordneten Segmentes, z.B. Satz, wird in der Allgemeinen Linguistik als Kookkurrenz bezeichnet.

Lemmatisierung

Reduktion auf die Grundform eines Wortes, also diejenige Form, unter der der Begriff in einem Nachschlagewerk zu finden ist.

Levenshtein-Distanz

Anzahl von Einfüge-, Lösch- und Ersetz-Operationen, um eine Zeichenkette in eine andere zu verwandeln.

Siehe auch: Editierdistanz bei der Parallelstellensuche.

Markup

Eine Markup language (ML) oder Auszeichnungssprache beschreibt den Inhalt eines Dokumentes oder das Verfahren, welches zur Verarbeitung der Daten notwendig ist. HTML, XML oder LaTeX sind Auszeichnungssprachen.

Metadaten

Metadaten oder auch Metainformationen sind allgemein Daten, die Informationen über Merkmale beinhalten, die nicht Bestandteil der Daten selbst sind. Bei einer Korpusanalyse werden z.B. alle bibliographischen Informationen als Metadaten behandelt.

MIT-Lizenz

Die MIT-Lizenz (auch X-Lizenz oder X11-Lizenz) ist eine aus dem Massachusetts Institute of Technology stammende Lizenz für die Software-Benutzung, die erlaubt, die Software zu verwenden, kopieren, ändern, fusionieren, verlegen, verbreiten, unterlizenzieren und/oder zu verkaufen, sofern ein Urheberrechtsvermerk und der Erlaubnisvermerk den Kopien beigelegt sind.

N3

Notation 3 ist eine formale Sprache, die beispielsweis als Syntax für RDF-Daten genutzt werden kann:

```
<#Tim Berners-Lee> <#entwickelte> <#N3> .
```

N-Gramm

Zerlegung eines Textes in einzelne Fragmente der Anzahl N. Die Fragmente können Buchstaben, Phoneme oder auch Wörter sein. In der Computerlinguistik finden sich oft Bi- oder Trigramme aus Zeichen (Buchstaben und/oder Satzzeichen).

NER

Named **E**ntity **R**ecognition - Eigennamenerkennung. Begriffe eines Textes werden bestimmten Klassen zugeordnet, z.B. Orte oder Personen.

Normalisierung

Allgemein wird darunter die Vereinheitlichung von Text verstanden.

Parser

Ein Parser ist ein Programm, welches eine Eingabe zerlegt und in ein für die Weiterverarbeitung brauchbares Format umwandelt.

Persistent Identifier

Ein künstlich zugewiesenes Merkmal zur eindeutigen, dauerhaften Identifizierung eines Subjektes / Objektes wird als persistent Identifier (persistent ID oder PID) bezeichnet.

PNG

Portable **N**etwork **G**raphics ist ein Grafikformat, welches verlustfrei komprimieren kann. Es wurde als freier Ersatz für Graphics Interchange Format (GIF) entwickelt und unterstützt die Transparenz per Alphakanal.

PoS

Part-**o**f-**S**peech Tagging ordnet die Wörter eines Textes Wortarten zu.

PURL

Ein **P**ersistent **U**niform **R**esource **L**ocator verweist in Form einer URL nicht direkt auf eine Ressource, sondern auf einen Resolver, der die aktuelle Internet-URL liefert. DOI oder URN existieren alternativ dazu.

Resolver

Als Resolver wird in der Informatik allgemein eine Software zur Namensauflösung bezeichnet. Ein Linkresolver löst Metadaten z.B. in Form einer URN in lokale Bestandsdaten auf und liefert den dazu passenden Hyperlink.

RDA

Resource **D**escription and **A**ccess bezeichnet einen neuen Standard für die Erschließung von Ressourcen in Bibliotheken, Archiven und Museen als Nachfolger der Anglo-American Cataloguing Rules (AACR2).

RDF

Das **R**esource **D**escription **F**ramework wurde vom World Wide Web Consortium (W3C) zur Beschreibung von Metadaten entwickelt. Es gilt mittlerweile als wesentlicher Bestandteil des sogenannten semantischen Webs. Aussagen im RDF-Modell werden als Tripel von Subjekt, Prädikat und Objekt gebildet, zumeist in Form von XML oder N3.

Signifikanz

In der Statistik wird unter Signifikanz eine Kennzahl verstanden, welche die Wahrscheinlichkeit eines systematischen Zusammenhangs zwischen Variablen bezeichnet.

Similar-Text

Ein Algorithmus, der die Ähnlichkeit zweier Texte auf Zeichenbasis und mit Hilfe der **Editierdistanz** berechnet.

SQL

Datenbanksprache in relationalen Datenbanken. SQL (Allgemeiner Sprachgebrauch: **S**tructured **Q**uery **L**anguage) unterscheidet drei Befehlskategorien:

- Data Manipulation Language (DML) - Befehle zur Datenmanipulation
- Data Definition Language (DDL) - Befehle zur Definition des Datenbankschemas
- Data Control Language (DCL) - Befehle für die Rechteverwaltung und Transaktionskontrolle.

Stoppwords

Eine Liste von Wörtern, die bei der Verarbeitung eines Textes nicht berücksichtigt werden sollen.

SVG

Scalable **V**ector **G**raphics basiert auf XML und beschreibt zweidimensionale Vektorgrafiken.

siehe XML

Table has no rows

Bei der Datentabelle Zitation kann es vorkommen, dass die eingestellten Filterkriterien eine Anzeige von Datensätzen verhindern, obgleich Daten verfügbar sind. In diesem Fall zeigt die Visualisierung mit dem Hinweis „Table has no rows“ an, dass keine Datensätze den Filterkriterien entsprechen.

TEI

Das von der **T**ext **E**ncoding **I**nitiative entwickelte, gleichnamige Dokumentenformat basiert in der aktuellen Version P5 auf XML und hat sich zum De-facto-Standard zur Kodierung gedruckter Werke in den Geisteswissenschaften entwickelt.

siehe XML

TIFF

Tagged **I**mage **F**ile **F**ormat ist ein Bilddateiformat, welches insbesondere für hochaufgelöste Bilder in druckfähiger, verlustfreier Qualität benutzt wird.

Tokenesierung

In der Computerlinguistik wird damit die Zerlegung in Segmente auf Wortebene bezeichnet.

TSV

Das textbasierte Dateiformat TSV (**T**ab-**S**eparated **V**alues) ist eine Form von DSV (Delimiter-separated values). Die Daten sind in Tabellenform, also zweidimensional, gespeichert. Jede Zeile ist ein Datensatz. Felder werden mittels **Tab-Stop** separiert.

siehe CSV

URI

Laut RFC 1630 von T. Berners-Lee aus dem Jahr 1994 ist URI ein Akronym für Universal Resource Identifiers, inzwischen wird es als **U**niform **R**esource **I**dentifier verstanden. Ein URI dient zur Identifizierung einer abstrakten oder physischen Ressource und kann aus fünf Teilen bestehen, von denen aber nur scheme und path zwingend vorhanden sein müssen: scheme://[authority]/path?[query]#[fragment] .

URL

Uniform **R**esource **L**ocator identifizieren eine Ressource anhand der zu verwendenden Zugriffsmethode. Der eAQUA-Internetauftritt wird z.B. über **http://www.eaqua.net** erreichbar gemacht, eine E-Mail-Adresse mit dem

Schema **mailto:max.mustermann@example.org** erkannt.

URN

Publikationen können im Netz dauerhaft und zuverlässig zitiert werden, indem eindeutige, standortunabhängige Identifikatoren URNs (**U**niform **R**esource **N**ame) anstelle von URLs verwendet werden. URNs sind URIs mit dem Schema urn:namensraum:namensraum-spezifischerTeil, also z.B. urn:nbn:de:101-2012121200 für das Werk “Policy für die Vergabe von URNs im Namensraum urn:nbn:de (Version 1.0, Stand: 29. November 2012)” der Deutschen Nationalbibliothek.

UTF

Unicode **T**ransformation **F**ormat. Zeichen werden zum Zwecke der elektronischen Verarbeitung auf eine Folge von Bytes abgebildet. Übliche Kodierungsverfahren sind

- UTF-8 - Zwischen 1 und 4 Byte. Die Codepoints 0 bis 127, die dem ASCII-Zeichensatz entsprechen, werden mit Hilfe von sieben Bits kodiert. Das achte leitet ein längeres Unicode-Zeichen ein, welches die nachfolgenden 1-3 Bytes belegt. UTF-8 speichert lateinische Zeichen am effizientesten.
- UTF-16 - Ein oder zwei 16-Bit-Einheiten (2 oder 4 Bytes) werden zur Kodierung eines Zeichens verwendet.
- UTF-32 - Kodiert immer 32 Bit (4 Byte). Durch die feste Länge am einfachsten zu handhaben, benötigt dafür mehr Speicher.

Wortstammreduktion

Auch Stemming, Stammformreduktion oder Normalformenreduktion genannt. Verschiedene morphologische Varianten eines Wortes werden auf ihren gemeinsamen Wortstamm zurückgeführt.

XLS

Binäres Dateiformat von Microsoft Excel, welches bis 2007 ausschließlich gebräuchlich war.

XML

E**x**tensible **M**arkup **L**anguage ist eine Auszeichnungssprache zur Darstellung strukturierte Daten in Textform. Sie wird vor allem als Austauschformat zwischen verschiedenen Computersystemen genutzt.

Beginn eines TEI-XML Dokuments aus der Perseus Digital Library:

```
<?xml version="1.0"?>
<!DOCTYPE TEI.2
  PUBLIC "-//TEI P4//DTD Main DTD Driver File//EN" "http://www.tei-c.org/Guidelines/DTD/tei2.dtd" [
<!ENTITY % TEI.XML "INCLUDE">
<!ENTITY % PersProse PUBLIC "-//Perseus P4//DTD Perseus Prose//EN"
"http://www.perseus.tufts.edu/DTD/1.0/PersProse.dtd" >
%PersProse;
]>
<TEI.2>
   <teiHeader type="text" status="new">
```

```
      <fileDesc>
         <titleStmt>
            <title>De liberis educandis</title>
            <title type="sub">Machine readable text</title>
            <author n="Plut.">Plutarch</author>
            <editor role="editor" n="Teubner">Gregorius N.
Bernardakis</editor>&responsibility;&fund.NEH;</titleStmt>
```

W3C

Das World Wide Web Consortium standardisiert die Techniken im World Wide Web. Es wurde 1994 am MIT gegründet.

Wahrscheinlichkeitsverteilung

Die Wahrscheinlichkeitsverteilung ist das theoretische Pendant zur empirisch ermittelbaren Häufigkeitsverteilung. Sie beschreibt, mit welchen Wahrscheinlichkeiten eine Zufallsvariable ihre möglichen Werte annimmt.

Zipf'sches Gesetz

Das Gesetz besagt, wenn man die Typen eines Textes ihrer Häufigkeit *f* nach ordnet und ihnen dabei jeweils einen Rang *r* zuweist, dann ergibt das Produkt von *f* und *r* jeweils einen konstanten Wert *k*.