

Paul fragt Laura: „Wer war eigentlich Lorenzo Gafà?“ Laura antwortet: „Hmm, keine Ahnung...“ Diese Szene findet ihre Fortsetzung sicher nicht in einer Bibliothek, sondern vor einem Rechner mit Internetanschluss.

Sehr hohe und international ständig steigende Nutzungszahlen des Internets belegen ein großes aber durchaus unterschiedliches Interesse an diesem Medium. Eine Gesellschaft ohne dieses Netzwerk ist in vielen Bereichen kaum noch vorstellbar. Durch die massive Etablierung der Webtechnologien und die fortschreitende Digitalisierung in allen nur denkbaren Bereichen sind die Anforderungen an die Informationsinfrastruktur deutlich gestiegen.

2.1 Grundlagen Internet

Dieses Kapitel gibt einen Überblick über die Entstehung, den Aufbau, die Nutzungsvoraussetzungen, die Dienste (Möglichkeiten) sowie die Orientierung im Internet.

2.1.1 Historie und Basis

1962 wurden die ersten bereits existierenden militärischen Netzwerke, die einen sehr kleinen Umfang aufwiesen und vollständig souverän angelegt waren, auf Vorschlag von Paul Baran zu einem großen Netzwerk ohne Zentraleinheit zusammengeschlossen. Die erstmals als digitale Netzarchitektur umgesetzte Idee entsprach der Struktur eines traditionellen Fischernetzes. Das zugrunde liegende Regelwerk für die Kommunikation der Rechner untereinander war das Netzwerkprotokoll NCP (network control protocol). Dieses frühe Netzwerk trug den Namen ARPAnet.



Abb. 2.1 Paketorientierung

1969 wurden vier separate universitäre Netzwerke in den USA auf die gleiche Art miteinander verbunden und 1972 umfasste dieses Netzwerk vierzig Rechner.

Die noch heute weltweit genutzte Netzwerkprotokollfamilie TCP/IP (Transmission Control Protokoll/Internet Protokoll) wurde 1982 eingeführt und das so neu entstandene Netzwerk bekam den Namen Internet, eine Abkürzung für den englischen Begriff Inter-connected Networks, also zusammengefügte Netze.

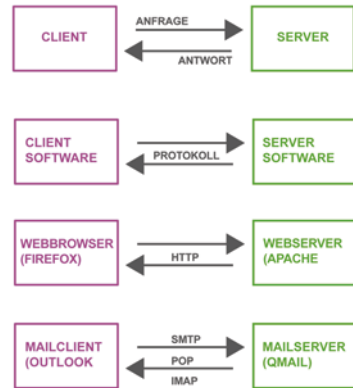
Die Datenübermittlung erfolgte damals wie heute nicht in einer einzelnen Gesamtheit sondern paketorientiert: Die Gesamtdatenmenge wird in einzelne Einheiten, die sogenannten Pakete, zerlegt und mit Headerdaten versehen, die Zugehörigkeit und Ziel des Paketes beinhalten. Wie in Abb. 2.1: Paketorientierung dargestellt, werden diese Pakete unabhängig voneinander auf den unterschiedlichsten Wegen durch das dezentrale Netzwerk geschickt und erst am Ziel beim Empfänger wieder zu einer Dateneinheit zusammengesetzt.

Diese grundlegende global einheitliche Struktur ermöglicht das Netzwerk „Internet“ und ist somit ein eher konzeptueller Begriff. Was das Internet letztlich ausmacht, ist die festgelegte, einheitliche Form (Protokoll) der Datenkommunikation zwischen einem beliebigen Client und einem Server, die physisch miteinander verbunden sind und TCP/IP als Kommunikationsgrundlage nutzen. Auf diesem Netzwerk können nun weitere Techniken aufsetzen, um Serviceleistungen im Internet anzubieten. Es empfiehlt sich hierbei eine standardisierte Vorgehensweise, um eine möglichst große Reichweite zu erzielen. Um diese Standards bemüht, hat Tim Berners-Lee¹ einerseits das WorldWideWeb ins Leben gerufen und andererseits im Oktober 1994 das World Wide Web Consortium, kurz W3C², gegründet. Dieses, in Arbeitsgruppen organisierte, internationale Konsortium entwickelt in einem sehr transparenten Prozess unter Einbeziehung relevanter Usergruppen aktuelle standardisierte Technologien für das WWW. Ein zunächst als Arbeitsentwurf (Working Draft) vorgestellter Standard wird öffentlich diskutiert und letztlich als W3C-Empfehlung

¹ Biographie unter <http://www.w3.org/People/Berners-Lee/Overview.html>.

² Siehe <http://www.w3.org> und <http://www.w3c.de>.

Abb. 2.2 Client-Server-System



(W3C Recommendation) veröffentlicht.³ Prominente Beispiele sind HTML/XHTML, XML und Familie sowie XML Schema und XSLT, CSS, PNG, RDF, DOM, OWL uvm.

2.1.2 Client – Server – System

Zwei Agenten, die miteinander in Kontakt treten möchten, benötigen eine einheitliche Strategie und eine definierte Sprachregelung. Übertragen auf heutige Rechensysteme bedeutet dies, zwei Softwaresysteme – der Client, der die Anfrage stellt und der Server, der diese beantwortet – nutzen die bereits vorhandene physische Verbindung zwischen den Rechnern und TCP/IP, also das Internet, zur allgemeinen Datenübermittlung und verwenden zielorientierte Protokolle (Regelwerke), um eine bestimmte Funktion zu realisieren. Auf Abb. 2.2: Client-Server-System ist dies veranschaulicht.

Das Ziel „Anzeige einer Webseite“ erfordert also die Anfrage der Browsersoftware als Client mittels des Regelwerkes/Protokolls HTTP⁴ an einen Server, der diese Anfrage versteht, also ein Webserver wie z. B. Apache HTTP Server. Dieser übermittelt den gewünschten, dort gespeicherten HTML-Code⁵ an den Browser, der wiederum in der Lage ist, diesen zu interpretieren und anzuzeigen.

Das Ziel „Versenden einer E-Mail“ erfordert entsprechende Client-Software, die das Protokoll SMTP⁶ benutzen kann, um eine Mail an einen Mailserver zu übertragen, der diese dann zur Abholung durch den Empfänger verfügbar macht.

³ Siehe <http://www.w3.org/TR/>.

⁴ HyperTextTransferProtokoll: HTTP/1.0 1996: Standard RFC 1945 und HTTP/1.1 1999: Standard RFC 2616/Tutorial für Webentwickler: <http://net.tutsplus.com/tutorials/tools-and-tips/http-the-protocol-every-web-developer-must-know-part-1/>.

⁵ HyperTextMarkup Language, Kap. 2.2.

⁶ Simple Mail Transfer Protocol, RFC 5321.

Nur wenn die Protokolle – die Regelwerke für ein bestimmtes Ziel – exakt eingehalten werden, kann eine Kommunikation zwischen zwei Rechensystemen gelingen.

2.1.3 Dienste – Funktionalitäten – Möglichkeiten

Die verschiedenen Protokolle, die im Internet Anwendung finden, bedingen die verschiedenen Dienste und damit die Möglichkeiten des Internets. Eine Etablierung neuer Protokolle ermöglicht die Erweiterung der Funktionalität des heutigen Internets auf zukünftige Herausforderungen und neue Ideen. In Tab. 2.1 finden sich einige Beispiele.

2.1.4 Aufbau

Das Internet ist letztlich ein Zusammenschluss von vielen Einzelnetzen, die eine gemeinsame Kommunikationsstrategie verwenden. Jeder einzelne Netzbetreiber ist für sein Netz verantwortlich und trägt damit seinen Teil zum Internet bei. Ressourcen werden über Webserver der Provider zur Verfügung gestellt und können von beliebigen Rechnern angefragt werden. Siehe dazu Abb. 2.3: Internet – Aufbau und Teilhabe.

Voraussetzung ist, dass ein Rechner eine Verbindung zum Internet über einen Provider und eine Client-Software hat, die solche Anfragen formulieren und die Antworten interpretieren kann.

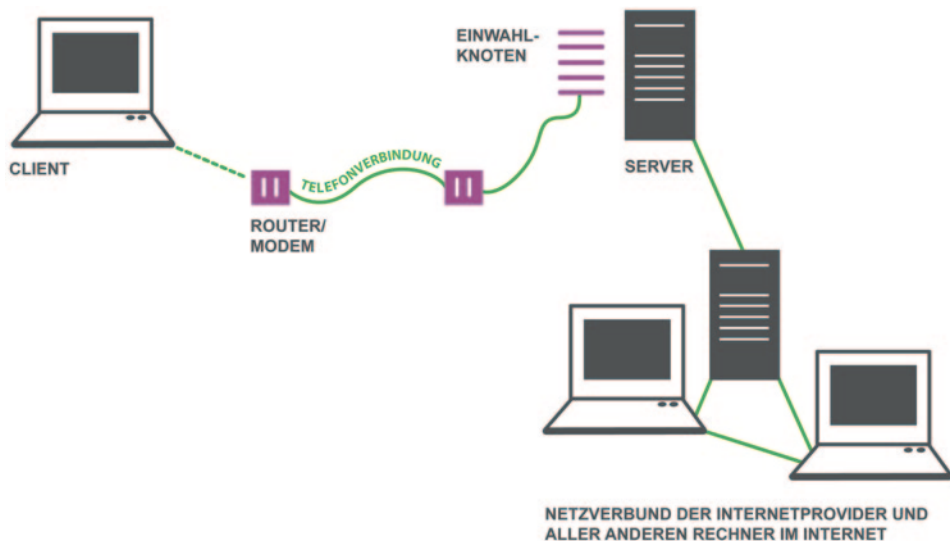


Abb. 2.3 Internet – Aufbau und Teilhabe

Tab. 2.1 Protokolle

Protokoll	Dienst	Client-Software	Server-Software
SMTP Simple Mail Transfer Protocol RFC 5321 POP Post Office Protocol RFC 1939 IMAP Internet Message Access Protocol RFC 3501	E-Mail	Thunderbird Windows Live -Mail Mail iOs AirMail AquaMail	Mercury Postfix hMailServer Exim Qmail
HTTP Hypertext Transfer Protocol RFC 2616	WWW	Firefox Google Chrome Internet-Explorer Safari Opera	Apache HTTP Server MS Internet Information Server IBM HTTP Server
HTTP Hypertext Transfer Protocol RFC 2616	Internetradio	Streaming-Clients	Streaming-Server
IRC RFC 2810	Internet Relay Chat	mIRC XChat	IRCD=IRC-Daemon
XMPP Extensible Messaging and Presence Protocol RFC 6120 – 22	Jabber, Instant Messaging	Psi	Open Fire
Proprietäre Protokolle (kennt nur die NSA...)	Instant Messaging	Skype, ICQ, AIM, MSN Messenger, IMessage	ICQ, AIM, MSN Messenger, IMessage
SIP Session Initiation Protocol RFC 3261	VoIP Internet-Telefonie	Face Time IChat	Voiceserver
LDAP Lightweight Directory Access Protocol RFC 4510/RFC 4511	Verzeichnisdienst	LBE meist integriert in komplexe Kommunikationssysteme	OpenLDAP
IPSec Internet Protocol Security PPTP Point-to-Point Tunneling Protocol GRE Generic Routing Encapsulation Protocol	VPN	Cisco VPN – Client	OpenVPN

2.1.5 Internet Partizipation

Kommerzielle oder öffentliche Internet-Provider unterhalten weltweit eigene Server (Hosts), aus denen sich das Internet zusammensetzt. Mittels eines solchen Providers als Dienstleister und eines Rechners (Desktop/Laptop/Pad/Smartphone und viele mehr), der mit einem Modem, einem Router oder einer Netzwerkkarte ausgestattet sein muss, kann auf die Server zugegriffen werden. Die Verbindung wird entweder über das (mobile) Telefonnetz oder über ein lokales Netzwerk (LAN/WLAN) bzw. einer Kombination aus beidem hergestellt.

Die Kommunikation, hier am Beispiel eines Webseitenaufrufs, erfolgt in folgenden Schritten:

- Anfragen werden von dem eigenen Rechner an den des Providers geschickt,
- dieser vermittelt die Anfrage an den Zielrechner (Webserver, auf dem die Webseite gespeichert ist),
- der wiederum eine Antwort (Quellcode der Webseite) an den Provider schickt,
- der diese an den anfragenden Rechner übermittelt.

Die Bereitstellung von Webseiten erfolgt ebenfalls über den Provider, der als Dienstleistung Speicherplatz auf einem Webserver zur Verfügung stellt.

2.1.6 Orientierung im Internet

Die Orientierung innerhalb des Netzes der unzähligen Server erfolgt über standardisierte IP-Adressen. Jeder Server hat eine eindeutige Adresse, die aus vier Zahlen zwischen 0 und 255 besteht, die durch Punkte getrennt werden. Ein Beispiel: 134.95.80.223.

Seit der IPv4, der vierten Version des Internet Protocol⁷, werden IPv4-Adressen verwendet, die aus 32 Bits bestehen und damit 4.294.967.296 Adressen ($=2^{32}$) abbilden können. Es wird die dotted decimal notation verwendet, die besagt, dass die vier Oktetts als vier durch Punkte voneinander getrennte ganze Zahlen in Dezimaldarstellung im Bereich von 0 bis 255 geschrieben werden müssen.

Aufgrund des kontinuierlichen Wachstums des Internets stellt diese auf etwas über vier Milliarden begrenzte Anzahl an IP-Adressen ein Problem dar und so wurde bereits 1998 IPv6⁸ entwickelt, die sechste Version des Internet Protocol, die 2^{128} Adressen (≈ 340 Sextillionen) zur Verfügung stellt. Dies wird erreicht, indem acht Blöcke mit jeweils 16 Bit durch einen Doppelpunkt voneinander getrennt verwendet und die Zahlen mit vier Hexadezimalstellen angegeben werden: 2001:0db8:0000:08d3:0000:8a2e:0070:7344. Durch Anwendung verschiedener Regeln kann diese sehr lange Schreibweise verkürzt werden.

⁷ Siehe <http://tools.ietf.org/html/rfc791>.

⁸ Siehe <http://tools.ietf.org/html/rfc2460>.

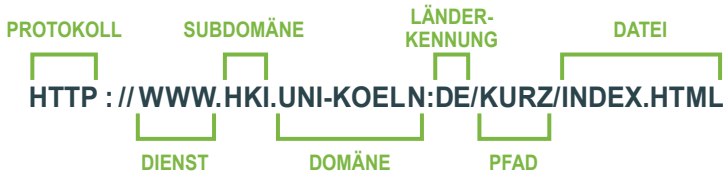


Abb. 2.4 Aufbau einer URL

Da die Repräsentation der IP-Adressen als Zahlenkolonnen zwar für den Rechner gut verarbeitbar, für den Menschen jedoch nur wenig verständlich und einprägsam sind, werden Bezeichnungen/Namen vereinbart, die eine für Menschen verständliche Variante der IP-Adresse darstellen. Ein sogenannter DNS⁹, der Domain Name Server, übersetzt zwischen diesen beiden Repräsentationsformen, also zwischen der URL¹⁰ und der zugehörigen IP-Adresse. Dieser Vorgang ist vergleichbar mit einer Telefonbuchfunktion, bei der einem Telefonanschluss eine Nummer zugeordnet wird.

2.1.6.1 Aufbau von WWW-Adressen/URL

Für Web-Adressen wurde 1994 ein Bezeichnungsstandard festgeschrieben, der unter RFC 1738¹¹ (Request for Comments) von der Network Working Group¹² veröffentlicht ist und seitdem Anwendung findet. Der Uniform Resource Locator, kurz URL, verortet eine bestimmte Web-Ressource auf einem bestimmten Server und dies muss dem in Abb. 2.4 dargestellten standardisierten Aufbau folgen.

Grundsätzlich ist eine URL nicht case-sensitive, Groß- und Kleinschreibung muss also nicht berücksichtigt werden. Jedoch kann ein Webserver je nach Konfiguration die Pfade und Dateinamen case-sensitive verwalten, sodass nach der Länderkennung Groß- und Kleinschreibung unter Umständen relevant ist.

Der Uniform Resource Locator ist ein untergeordnetes Konzept des Uniform Resource Identifier¹³, der ein Konzept abbildet, um real vorhandene oder abstrakte Ressourcen eindeutig zu identifizieren.

2.2 Einfacher Markup

Bei einfachen Markup-Systemen handelt es sich um Auszeichnungen (Tags), die Inhalte eines Dokumentes beschreiben. Ursprünglich entstanden aus den Anweisungen für die Setzer für einen Drucksatz, haben sich die Markup-Systeme zu komplexen Sprachen ent-

⁹ Domain Name Server werden von Internet Providern betrieben.

¹⁰ Uniform Resource Locator, siehe nachfolgendes Kapitel.

¹¹ Siehe <http://tools.ietf.org/html/rfc1738>.

¹² Verantwortlich: T. Berners-Lee (CERN), L. Masinter (Xerox Corporation), M. McCahill (University of Minnesota).

¹³ Siehe <http://tools.ietf.org/html/rfc3986>.

wickelt. Die häufig zunächst als kryptisch empfundenen Tag Namen lassen sich meist auf bekannte Begriffe zurück führen und können so nicht nur besser verstanden sondern auch leichter erinnert werden. Beispiel: Zum Einfügen einer horizontalen Linie wird das zunächst unverständliche `<hr>` verwendet, das eine Abkürzung für horizontal ruler ist.

2.2.1 HTML

HTML, die Hyper Text Markup Language, wird häufig als „Sprache des WWW“ bezeichnet. Es ist jenes Medium, das u. a. benötigt wird, um Texte und Bilder so aufzubereiten, dass sie im Internet mittels eines Webbrowsers betrachtet werden können. HTML ist bis zur Version 4 eine SGML-Anwendung (Standard Generalized Markup Language)¹⁴ und wurde am 13. März 1989 von Tim Berners-Lee am CERN in Genf festgelegt¹⁵.

Der reine uninterpretierte Text, aus dem ein HTML-Dokument besteht, wird „Quelltext“ genannt. Dieser kann auch innerhalb eines Webbrowsers eingesehen werden, wenn auf einer entsprechenden Webseite nach einem Maus-Rechtsklick im Kontextmenü „(Seiten)Quelltext anzeigen“ ausgewählt wird.

Quelltext kann mit jedem beliebigen ASCII-Editor erstellt werden, eine besondere Softwareanwendung ist dazu nicht notwendig. Um die gewünschte Darstellung der Webseite zu erhalten, muss der Quelltext interpretiert werden. Dies geschieht in der Regel durch einen Webbrowser (wie Firefox, Google Chrome, MS Internet Explorer, Safari, Opera, ...).

Ein HTML-Dokument besteht also aus reinem (lesbaren) Text, der sowohl die Inhalte der darzustellenden Webseite, wie auch deren Formatierungsanweisungen enthält. Für das Schreiben solcher Dateien können Softwaresysteme verwendet werden, die direkt auf die entstehenden Bedürfnisse ausgerichtet sind, sogenannte HTML-Editoren. Derartige Anwendungen unterstützen und erleichtern das Erstellen von HTML-Code je nach Produkt durch unterschiedliche Funktionen wie automatisches Einfügen, Syntax-Highlighting¹⁶ etc. bis hin zu sogenannten WYSISYG-Editoren, die mittels des Prinzips „What You See Is What You Get“ HTML-Code aufgrund einer visuellen Vorlage vollständig generieren.

2.2.2 ASCII-Format versus binäres Format

Jede Art der digitalen Darstellung von Text erfordert ein Speichern des Textes selbst zusammen mit mehr oder minder umfangreichen Informationen über seine Darstellung (Formatierung). Die Art der Speicherung dieser Daten bedingt den Unterschied zwischen einem ASCII und einem Binärformat:

¹⁴ Siehe Kap. 2.6 „XHTML“.

¹⁵ Siehe <http://www.w3.org/History/1989/proposal.html>.

¹⁶ Beispiel: `Syntax-Highlighting`.

- Sind diese Informationen so gespeichert, dass sie mit „neutralen“ Anwendungen (z. B. einfacher Texteditor) als Zeichenketten gelesen werden können, handelt es sich um ein ASCII- (oder Unicode-) Format der Datei.
- Sind diese Informationen aber auf eine Art gespeichert, dass sie nur von ganz bestimmten Softwareanwendungen interpretiert werden können, handelt es sich um ein binäres Textformat.

2.2.3 Information versus Metainformation

Fast alle für die Geisteswissenschaften interessanten Rechneranwendungen unterscheiden zwischen zwei Arten von Informationen:

1. der zu verarbeitenden Information z. B. der Zeichenkette „Neuhaus“ und
2. der Information, die wir benötigen, um die Zeichenkette adäquat zu verarbeiten.
 - Ist „Neuhaus“ der Name eines Ortes oder einer Person?
 - Soll „Neuhaus“ im Fettdruck erscheinen?
 - Ist „Neuhaus“ mit einer anderen Information verlinkt?

Es steht also zunächst die reine Information („Neuhaus“) zur Verfügung, die mit einer weiteren Aussage versehen wird, die sich auf diese reine Information bezieht und diese erweitert. Diese Information über eine reine Information wird als Metainformation bezeichnet. Die Unterscheidung zwischen Information und Metainformation wird auch in HTML vorgenommen. Hier wird dem Begriff der Information der Text, der dargestellt werden soll, zugeordnet und die dazugehörige Darstellungsanweisung ist die Metainformation. Bei Auszeichnungssprachen wie HTML ist diese Differenzierung einfach nachvollziehbar, da die Metainformation immer in spitze Klammern geschrieben wird:

```
< Metainformation > Information </ Metainformation >
```

Beispiel in HTML: **b** für bold = Fettdruck

```
<b> Neuhaus </b>
```

Für andere Auszeichnungssprachen

```
<Ort> Neuhaus </Ort>  
<Name> Neuhaus </Person>
```

Die Metainformation steht also innerhalb von Paaren spitzer Klammern, die „Tags“ genannt werden.

2.2.4 Grundsätzlicher Aufbau von HTML-Dokumenten

Für alle HTML-Dokumente gilt grundsätzlich: Jede Datei beginnt und endet mit der Meta-information `html`, da der Inhalt der gesamten Datei den Regeln der Markupsprache HTML folgt:

```
<html> ... </html>
```

Das Dokument ist immer in zwei Teile geteilt. Der erste Teil (Header) wird verwendet, um allgemeine Angaben zum Dokument festzuhalten:

```
<head> ... </head>
```

In dem zweiten Teil, dem sogenannten Body, stehen alle Informationen, die im Anzeigebereich des Browsers dargestellt werden sollen:

```
<body> ... </body>
```

Innerhalb des Headers wird u. a. die Beschriftung des Browserfensters festgelegt. Dies erfolgt mittels des Tags

```
<title> ... </title>.
```

Ein HTML-Dokument sieht also grundsätzlich wie folgt aus:

```
<html>
  <head>
    <title>Titel der Seite</title>
  </head>
  <body>
    Inhalt der Seite
  </body>
</html>
```

Der aktuelle Standard HTML5 erfordert vor dem ersten öffnenden `<html>` Tag folgende Information zum Dokumententyp:

```
<!DOCTYPE html>
```

Paare aus festgelegten „Start-Tags“ und „End-Tags“ (`<abc> ... </abc>`) verleihen dem dazwischenstehenden Text eine bestimmte (logische) Eigenschaft, die bei den ursprünglichen Version von HTML meist Konsequenzen auf die Darstellung des Textes hat. Welche

Digital Humanities

Grundlagen und Technologien für die Praxis

Kurz, S.

2016, XXVI, 220 S. 112 Abb., 23 Abb. in Farbe.,

Softcover

ISBN: 978-3-658-11212-7