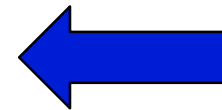

Summer School Digital Tools for Humanists

Pisa – June 13-22 2023

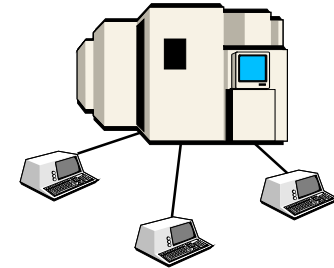
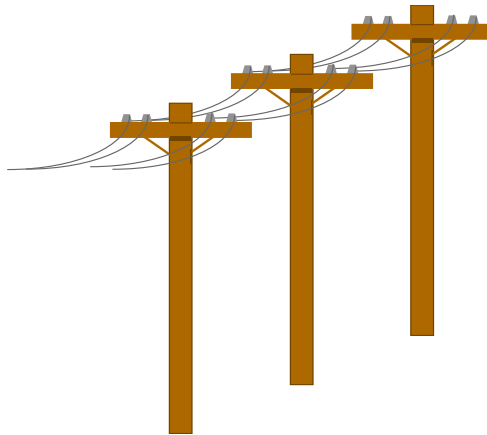
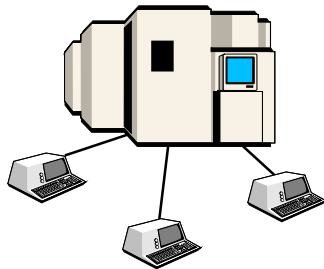
Refresher on Computer Fundamentals and Networking

- History of computers
- Architecture of a computer
- Data representation within a computer
- Computer networks and the Internet
- The Semantic Web



- Computer technology
 - CPU and integrated chips
 - Random Access Memories
 - RAM – from KB to GB
 - External memories
 - Tapes, hard disks, floppy disks
 - Memory sticks
 - CDs
 - DVDs
 - from MB to GB to TB to PB to EB
- Communication technology (networks)
 - (Telephone) line speed
 - Point to point (leased lines)
 - Local Area Networks
 - Inter-networking (TCP/IP)

From mainframe to mainframe
through telephone lines
(point to point connection)

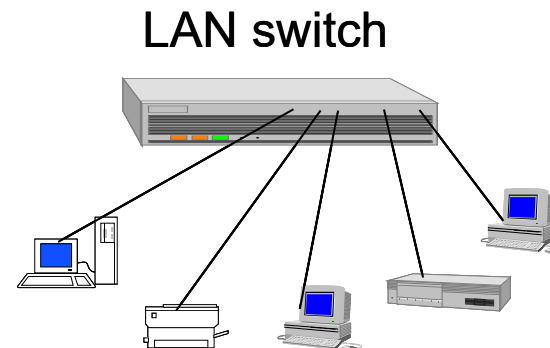
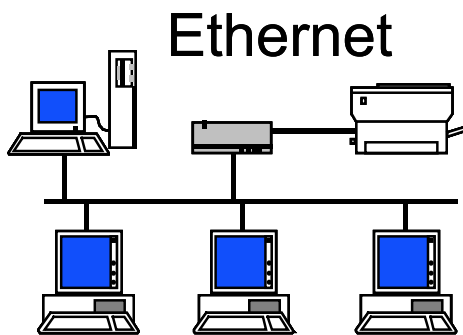
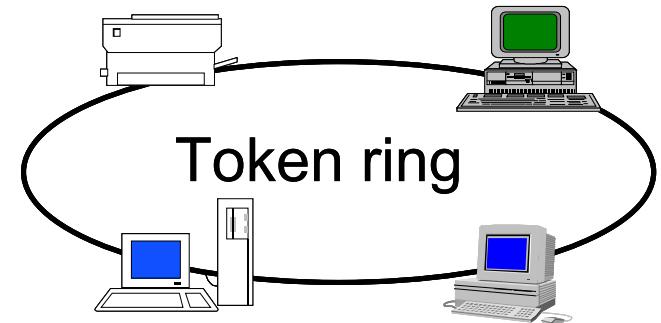


Telephone lines:
slow
expensive
regulated

- In the sixties, first studies on “networking”
 - Networking means communication between node A and node B through one or more intermediate nodes
- In the seventies, fragmentation of the market with the arrival of “minicomputers” provided further motivation for research on networking
- At the same time (in the seventies), the arrival of the LANs (Local Area Networks) provided the final impulse for the development of networking

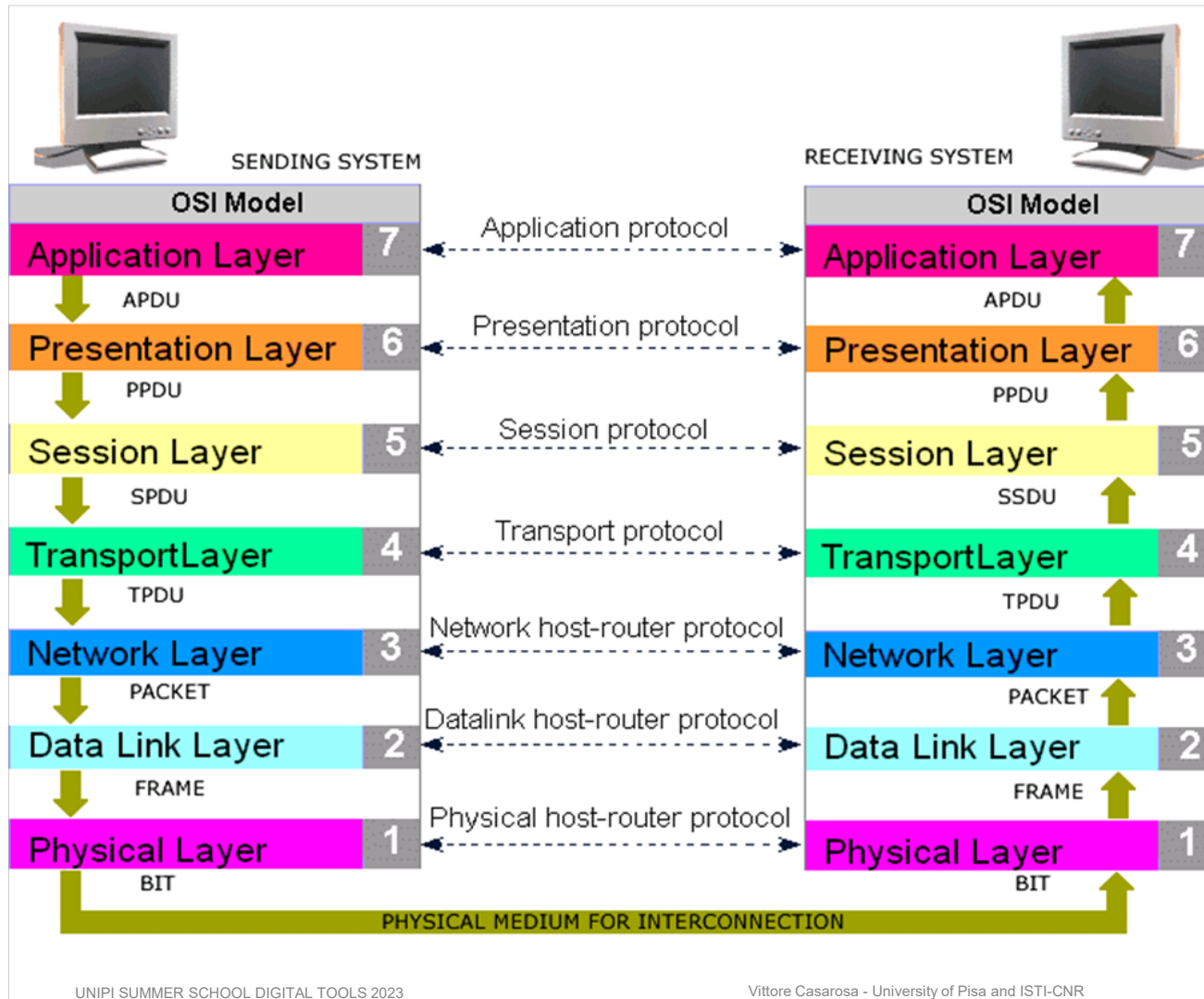
LAN - Local Area Networks

Private networks
Up to several kilometers
Speed up to 100 Mb/sec



- Starting in the late sixties, many research projects on networking, both from universities and industry
 - Arpanet, Cyclades, SNA (IBM), DECnet
- In the late seventies ISO (International Standard Organization), under pressure of a group of computer manufacturer, started the work for the proposal of a “new” communication standard, called OSI: Open System Interconnection
- The OSI model, though no longer in use today, has established a number of networking concepts and is still used as a “reference model”
- The main concept introduced by OSI is the “communication layer”

The OSI model 1980-1990

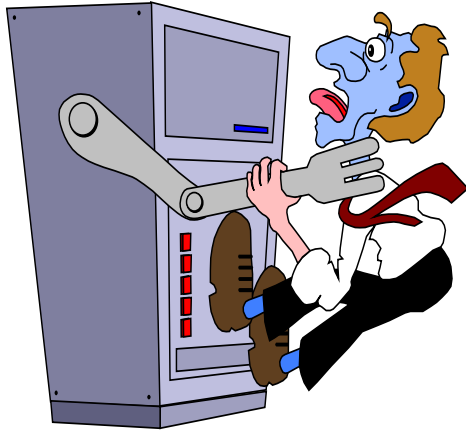


Protocol: formats and rules for exchanging messages between “partners” (e.g. computers)

Packet switching: messages are broken down into “packets”, and each packet gets to destination independently from the others.

- The OSI effort provided a sound and durable foundation for networking, but never became a “market leader”
 - Slow development
 - Initial opposition from IBM
 - “Designed by a Committee”
 - Expensive development
 - Heavy and slow in operation
- In the same period **the Internet** was defining a number of “light weight” protocols
- Most of the market preferred them to OSI

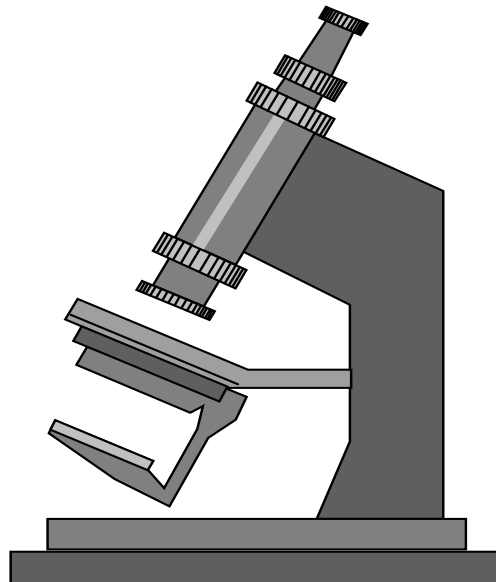
Internet evolution 1960-1990



Research
Network
NSF
Internet



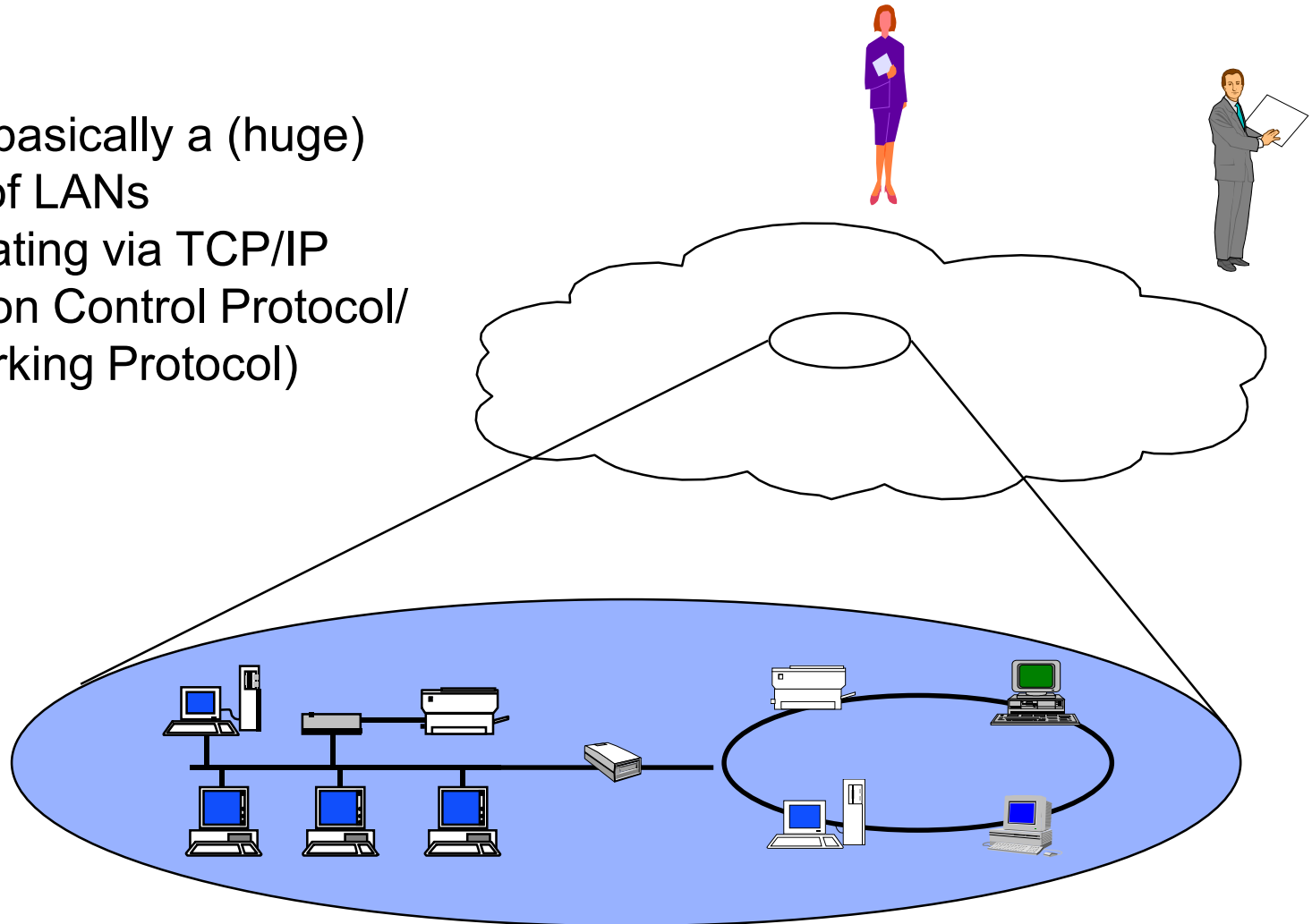
Experimental
Network
DARPA
Arpanet



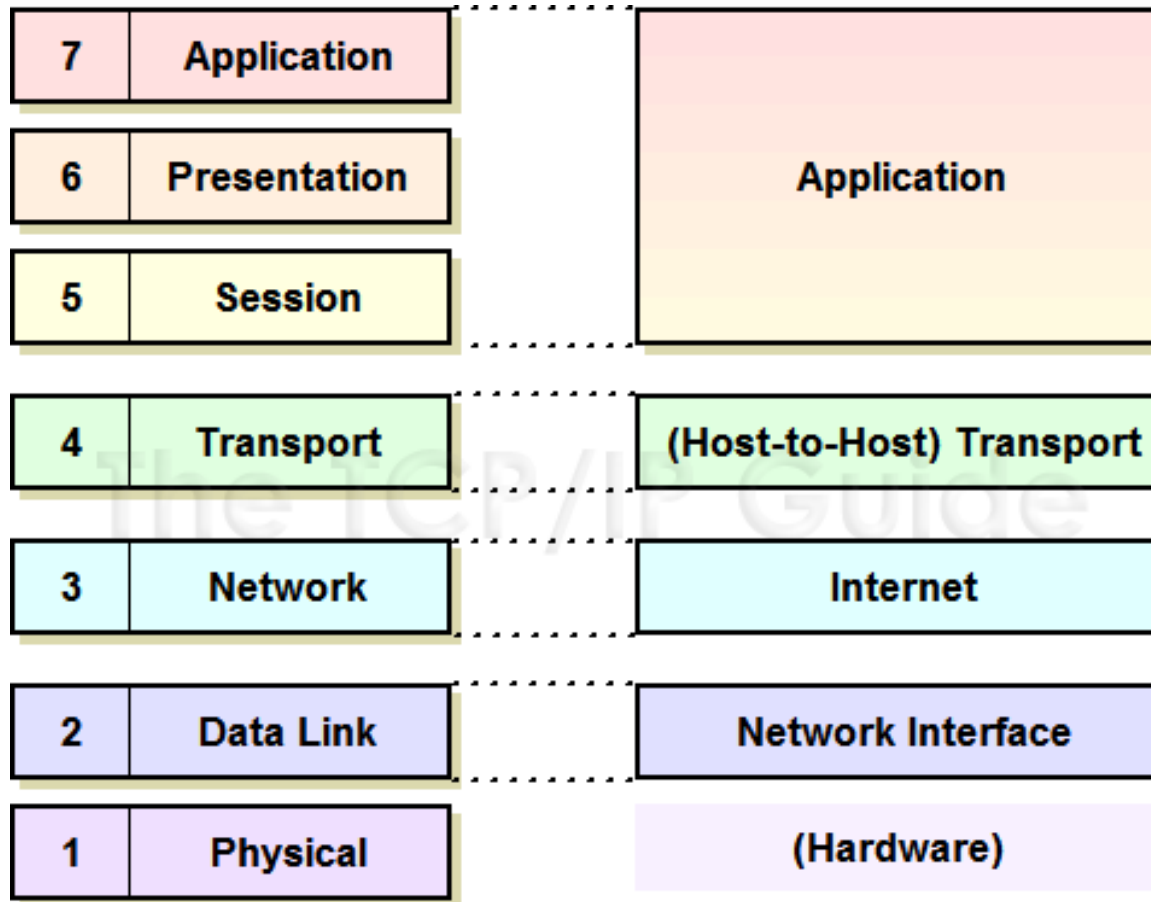
Communication
Infrastructure
Private and public
sectors
The Web

Inter-networking

Internet is basically a (huge) collection of LANs communicating via TCP/IP (Transmission Control Protocol/Internetworking Protocol)



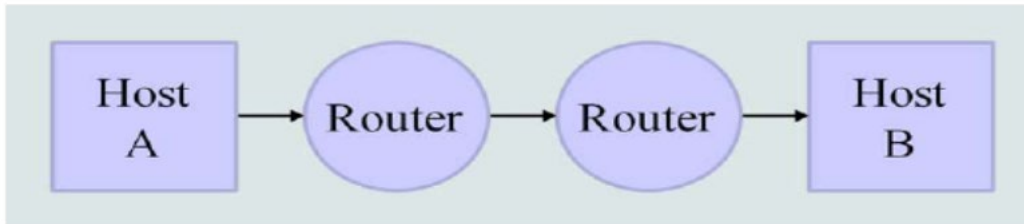
OSI and TCP/IP



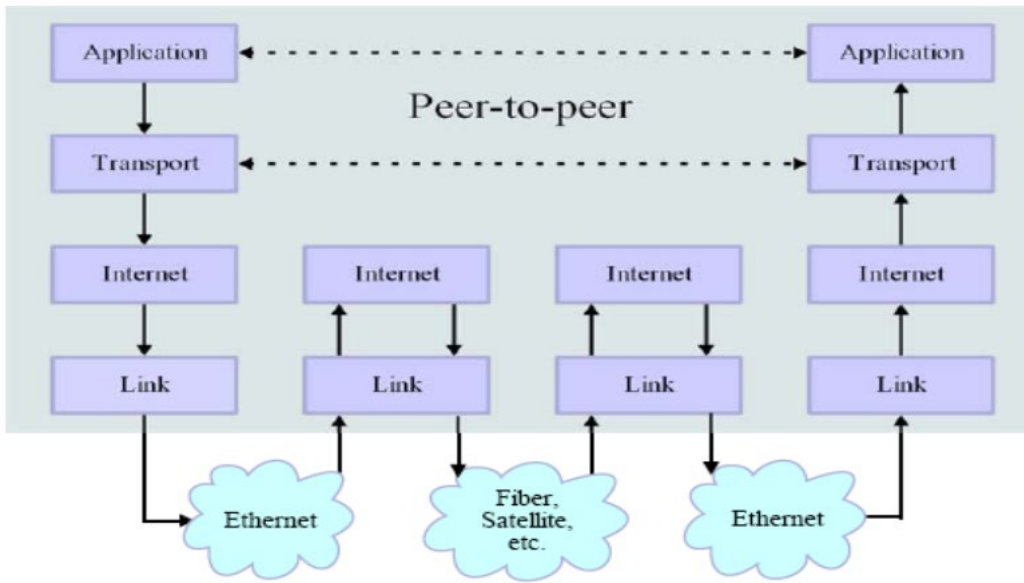
OSI Model

TCP/IP Model

Network Connections



Stack Connections



Application protocols

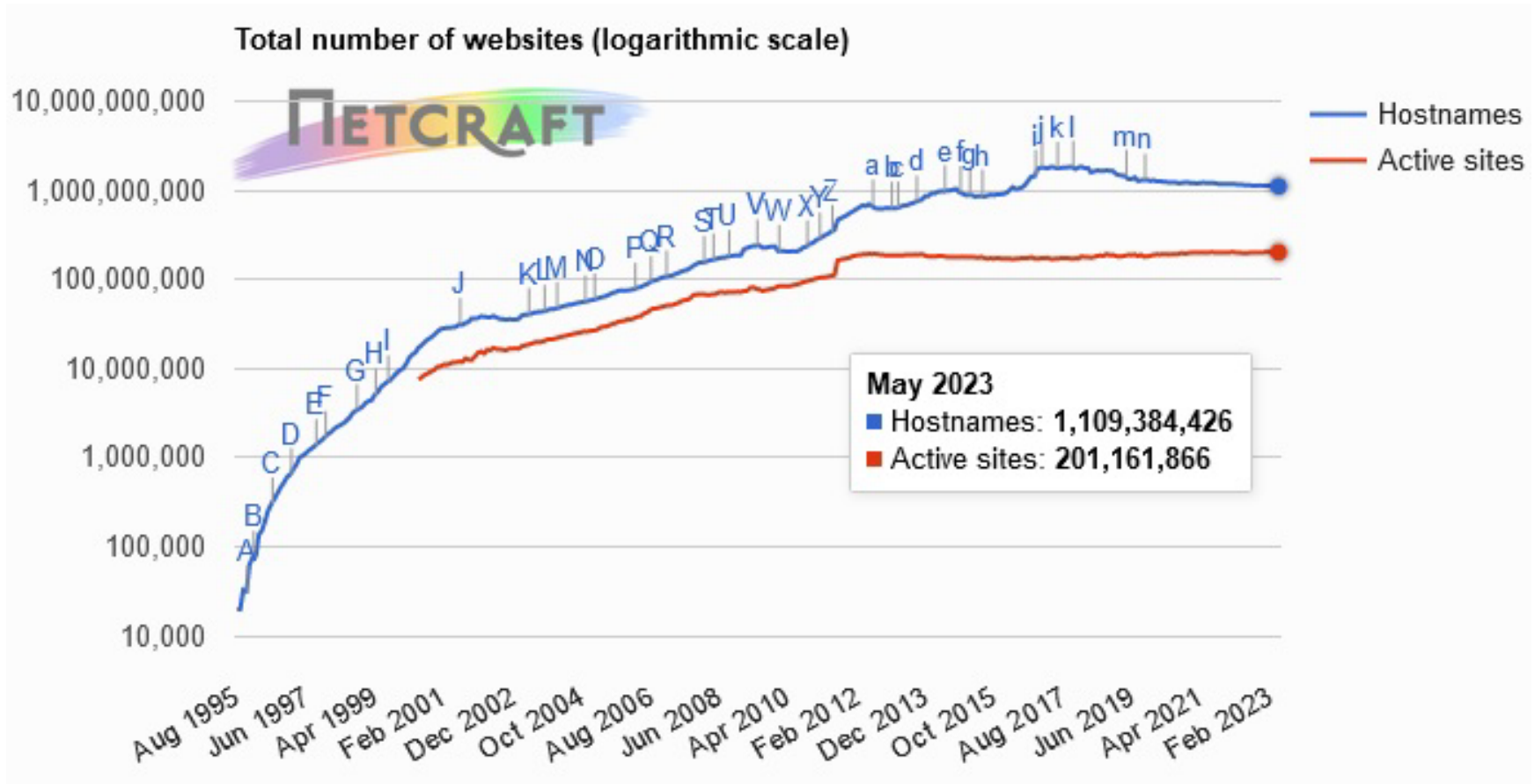
TCP/UDP

IP

ETHERNET

- Internet is a **worldwide network** of computers
 - It started in 1969 as a university research network (funded by DARPA) with 4 computers
 - By the end of the 80's, when it was opened to “the world”, it had more than 20000 hosts in universities and research centers worldwide
 - As of May 2023, the total number of web servers is estimated to be in the order of 1100 millions hosts
 - As of June 2022, the total number of Internet users is estimated to be 5,39 billions users
- The Web is the information space accessible through the Internet
 - As of March 2021, the number of “visible” Web pages (indexed by Google) was estimated to be between 50 and 60 billions pages
 - There is also a Deep Web, whose content is not indexed by any search engine, and whose size is completely unknown
- The Web has been made possible by a combination of computer technology and communication technology

Internet Web servers



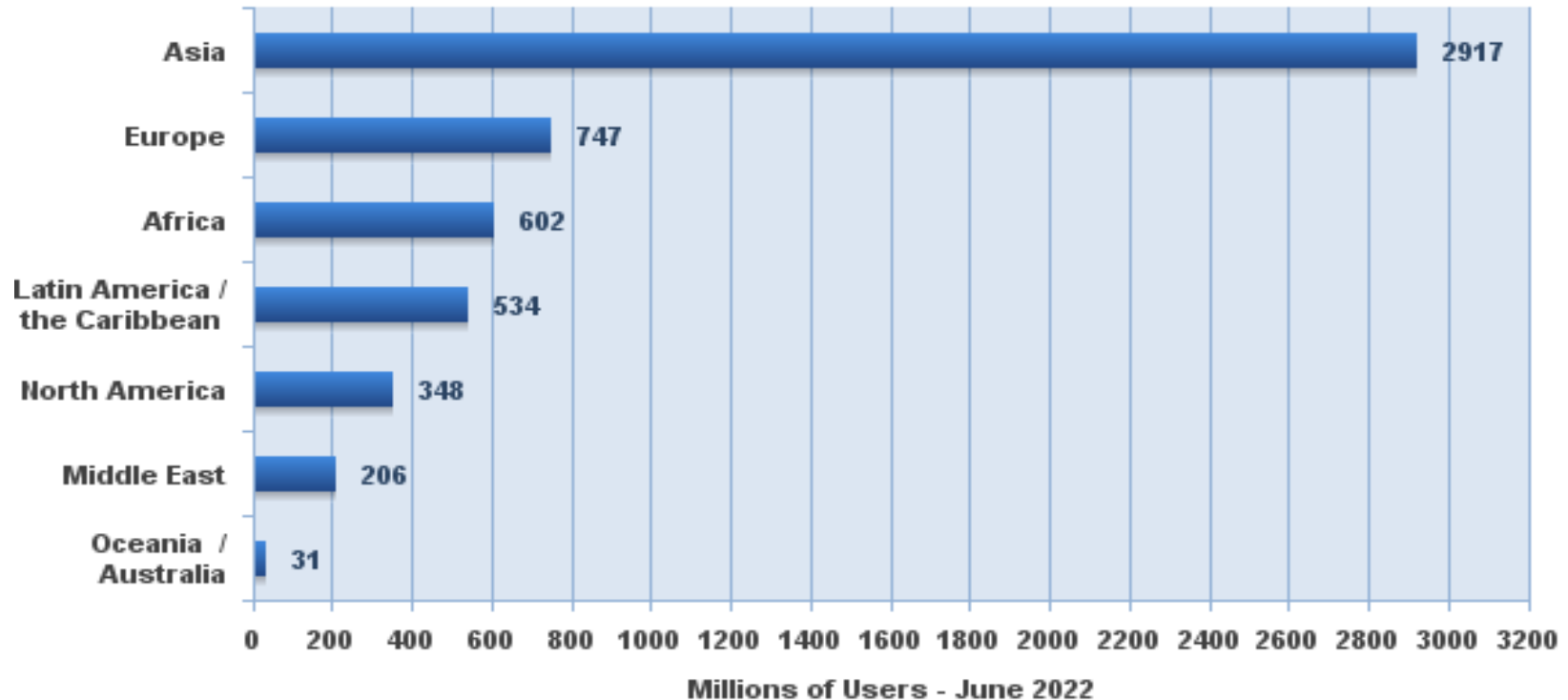
<https://news.netcraft.com/archives/category/web-server-survey/>

WORLD INTERNET USAGE AND POPULATION STATISTICS 2023 Year Estimates

World Regions	Population (2022 Est.)	Population % of World	Internet Users 31 Dec 2021	Penetration Rate (% Pop.)	Growth 2000-2023	Internet World %
Africa	1,394,588,547	17.6 %	601,940,784	43.2 %	13,233 %	11.2 %
Asia	4,352,169,960	54.9 %	2,916,890,209	67.0 %	2,452 %	54.2 %
Europe	837,472,045	10.6 %	747,214,734	89.2 %	611 %	13.9 %
Latin America / Carib.	664,099,841	8.4 %	534,526,057	80.5 %	2,858 %	9.9 %
North America	372,555,585	4.7 %	347,916,694	93.4 %	222 %	6.5 %
Middle East	268,302,801	3.4 %	206,760,743	77.1 %	6,194 %	3.8 %
Oceania / Australia	43,602,955	0.5 %	30,549,185	70.1 %	301 %	0.6 %
WORLD TOTAL	7,932,791,734	100.0 %	5,385,798,406	67.9 %	1,392 %	100.0 %

NOTES: (1) Internet Usage and World Population Statistics estimates are for June 30, 2022. (2) CLICK on each world region name for detailed regional usage information. (3) Demographic (Population) numbers are based on data from the [United Nations Population Division](#). (4) Internet usage information comes from data published by [Nielsen Online](#), by the [International Telecommunications Union](#), by [GfK](#), by local ICT Regulators and other reliable sources. (5) For definitions, navigation help and disclaimers, please refer to the [Website Surfing Guide](#). (6) The information from this website may be cited, giving the due credit to www.internetworldstats.com. Copyright © 2022, Miniwatts Marketing Group. All rights reserved worldwide.

Internet Users in the World by Geographic Regions - 2022

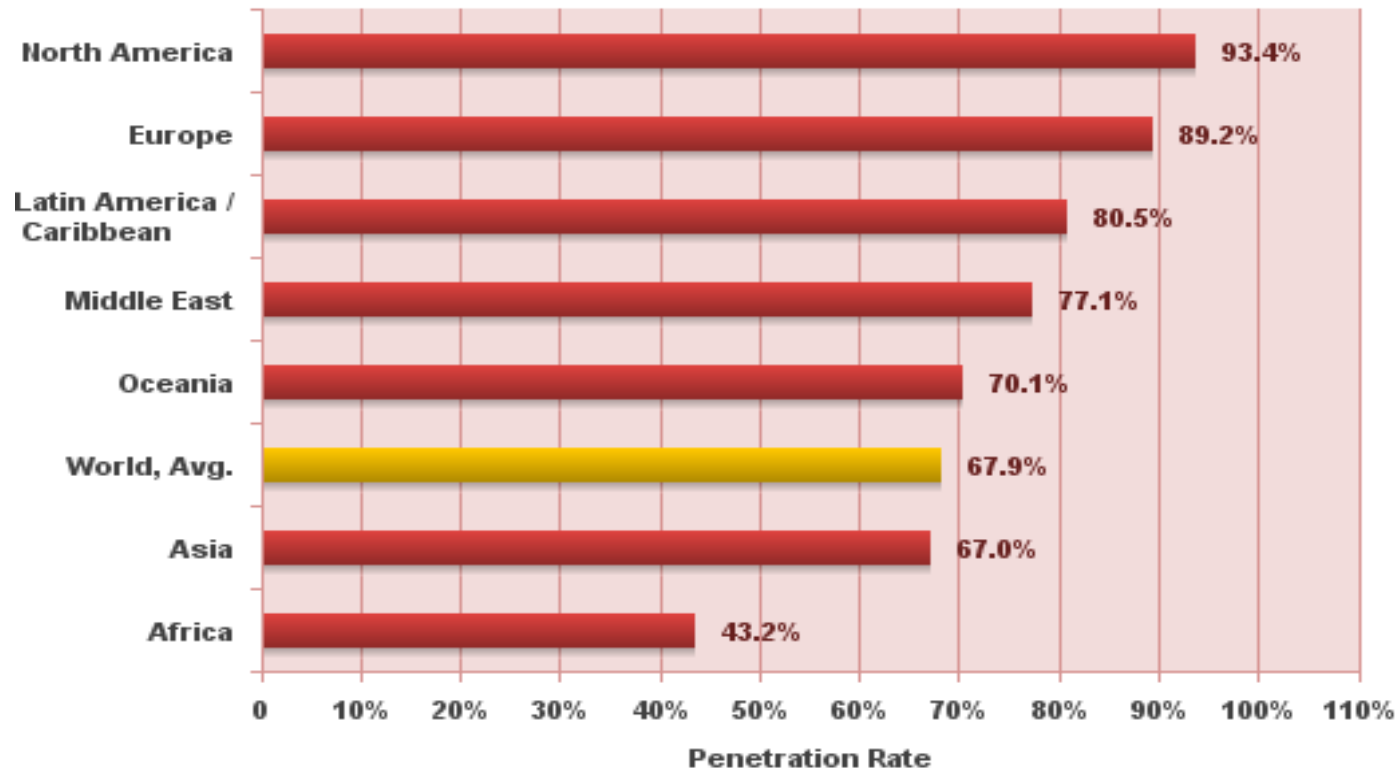


Source: Internet World Stats - www.internetworldstats.com/stats.htm

Basis: 5,385,798,406 Internet users estimated in June 30, 2022

Copyright © 2022, Miniwatts Marketing Group

Internet World Penetration Rates by Geographic Regions - 2022



Source: Internet World Stats - www.internetworldstats.com/stats.htm
 Penetration Rates are based on a world population of 7,932,791,734
 and 5,385,798,406 estimated Internet users in June 30, 2022.
 Copyright © 2022, Miniwatts Marketing Group

- Internet is a worldwide network of computers
 - It started in 1969 as a university research network (funded by DARPA) with 4 computers
 - By the end of the 80's, when it was opened to “the world”, it had more than 20000 hosts in universities and research centers worldwide
 - As of September 2021, the total number of web servers is estimated to be in the order of 1200 millions hosts
 - As of December 2021, the total number of Internet users is estimated to be 5,25 billions users
- The Web is the **information space** accessible through the Internet
 - As of May 2023, the number of “visible” Web pages (indexed by Google) is estimated to be between 40 and 50 billions pages
 - There is also a Deep Web, whose content is not indexed by any search engine, and whose size is completely unknown
- The Web has been made possible by a combination of computer technology and communication technology

The size of the indexed Web

Last Month

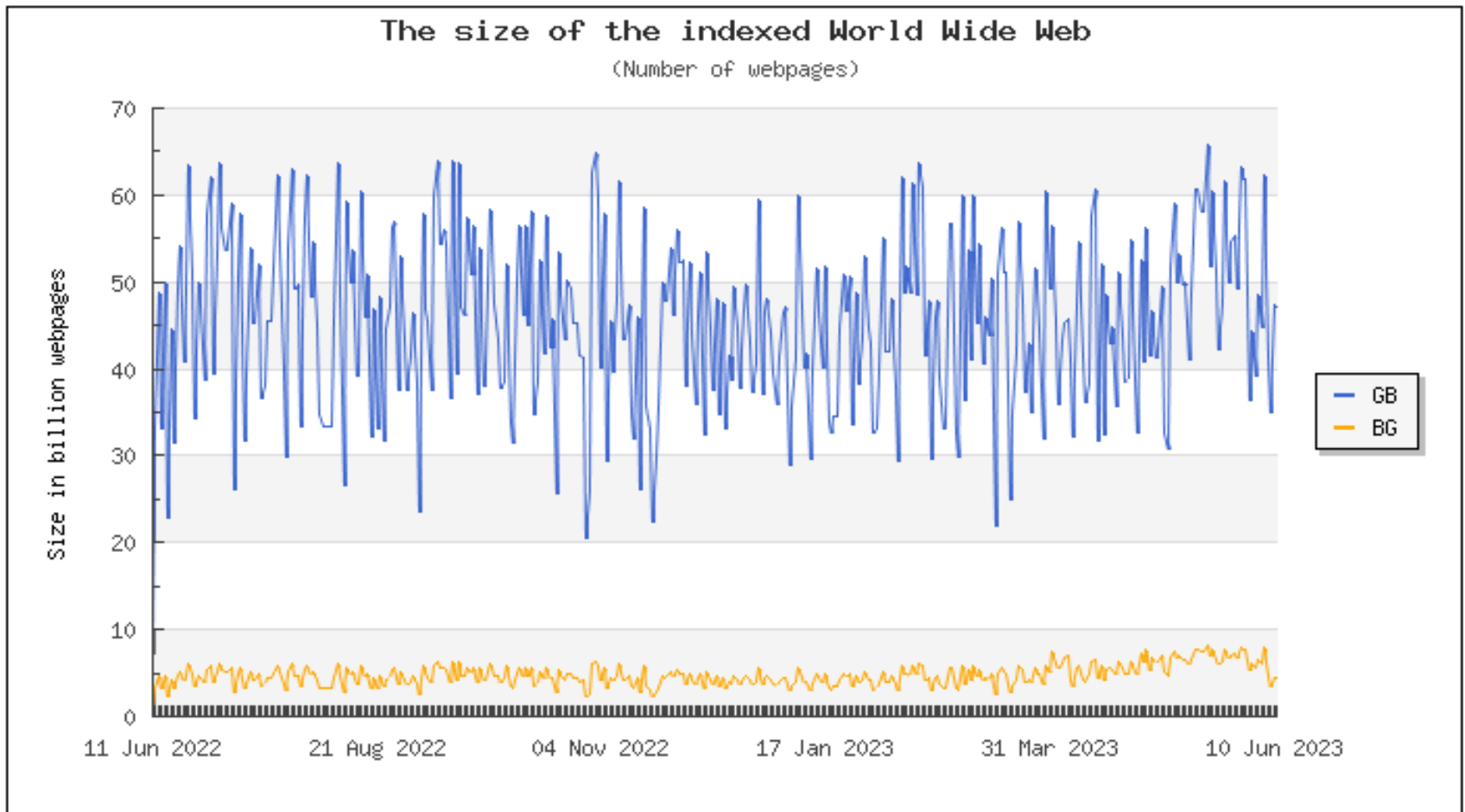
Last Three Months

Last Year

Last Two Years

Last Five Years

Last Ten Years



<https://www.worldwidewebsite.com/>

- Combination of computer technology and communication technology
- It all started with the “hyperlink” (late eighties)
- Then came the “browser” (Mosaic) (early nineties)
- Then came the “information explosion” (mid-nineties)
- Then came the “dot come, dot gone” (late nineties)
- Then came the second wave (early 2000)
- Then came the Web 2.0 (around 2004)
- Then came the Web 3.0 (around 2010)
- Today we have:
 - An estimate of about 1200 million hosts
 - An estimate of 40 to 50 billion pages on line

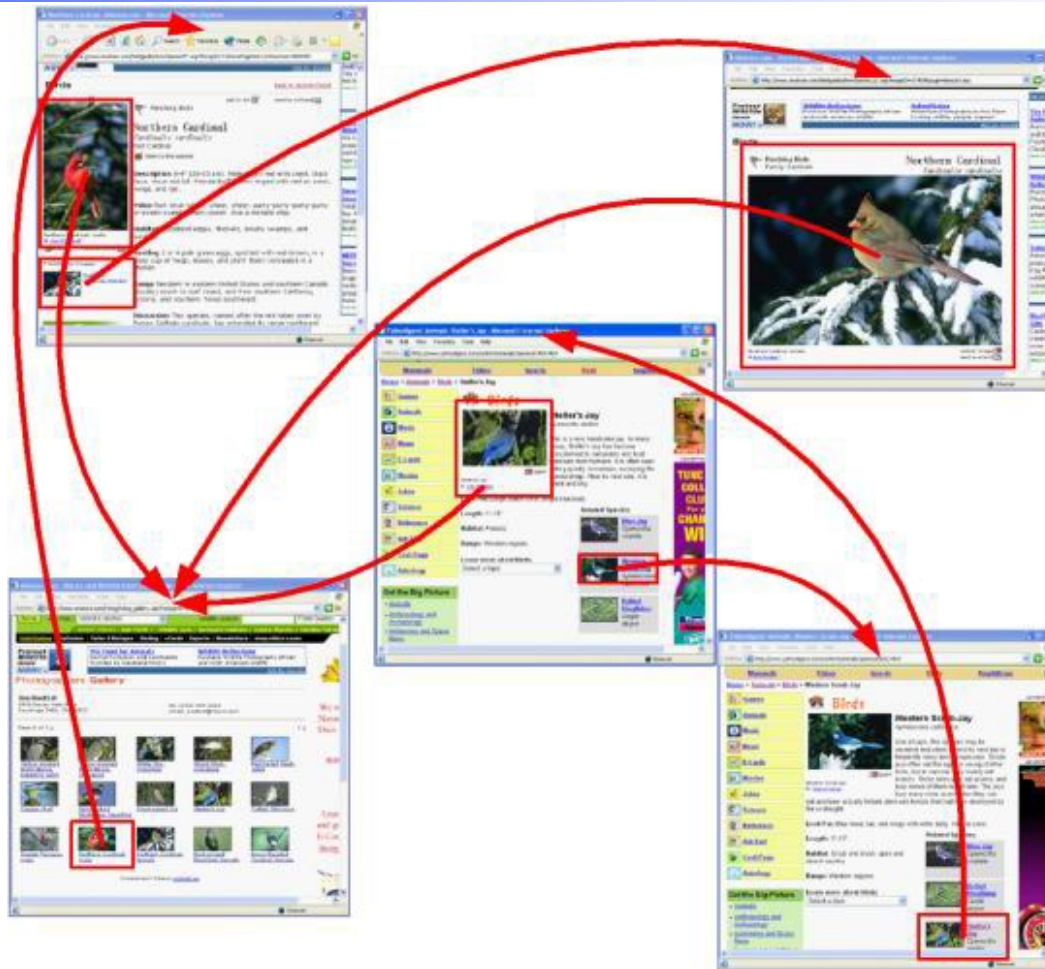
- Text processing applications started already in the early days of the computers (sixties)
- A “text processor” (or editor) has two main functions:
 - processing the text (delete, replace, insert, etc.)
 - specifying the format (bold, center, new line, etc.)
- The first editors were using a “mark up” language (i.e. commands intermixed with the text) to provide formatting instructions (only limited interactivity available through typewriter-like terminals)
- The “second generation” editors (interactivity available with display and mouse) were using the WYSIWYG paradigm: What You See Is What You Get

The hyperlink

- The idea of the “hyperlink” was (experimentally) proposed in the sixties, as a feature of a “**smart editor**”
 - selecting a portion of the text, it was possible to open a second document, in addition to the one being edited (very awkward to use on a typewriter-like terminal)
- With the arrival of display screens and the mouse (eighties) the hyperlink came back in “3D documents”
 - clicking on a portion of the text it was possible to open a second document, which was maintained as a second (virtual) screen behind the first one
- With the arrival of the (fast) internet, it became the “**web hyperlink**”
 - clicking on a portion of the text it was possible to open a second document, **coming from a different computer**

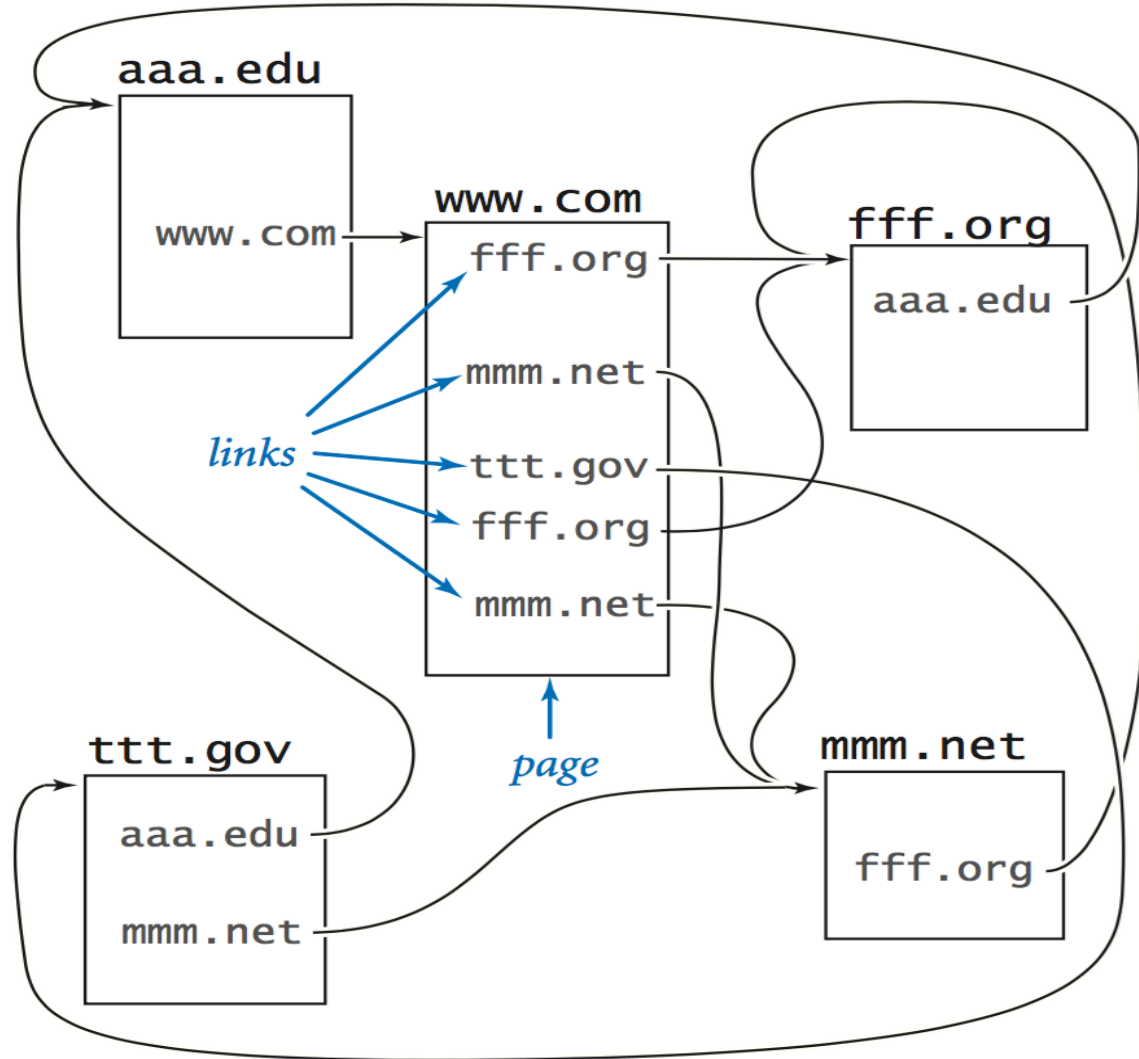
- With the arrival of the (web) hyperlink, the problem was then how to properly display a (web) page that had been generated on a different computer, possibly with a different (wysiwyg) editor
- The solution was the definition of HTML (Hyper Text Markup Language), i.e. a standard mark up language for formatting a page, and the implementation of smart editors (called browsers – the most popular was Mosaic, released in 1993) capable of correctly displaying pages formatted with HTML, regardless of where they were coming from
- At the same time it was defined the HTTP protocol (Hyper Text Transfer Protocol) for the exchange of information between the browser and the Web server

The World Wide Web

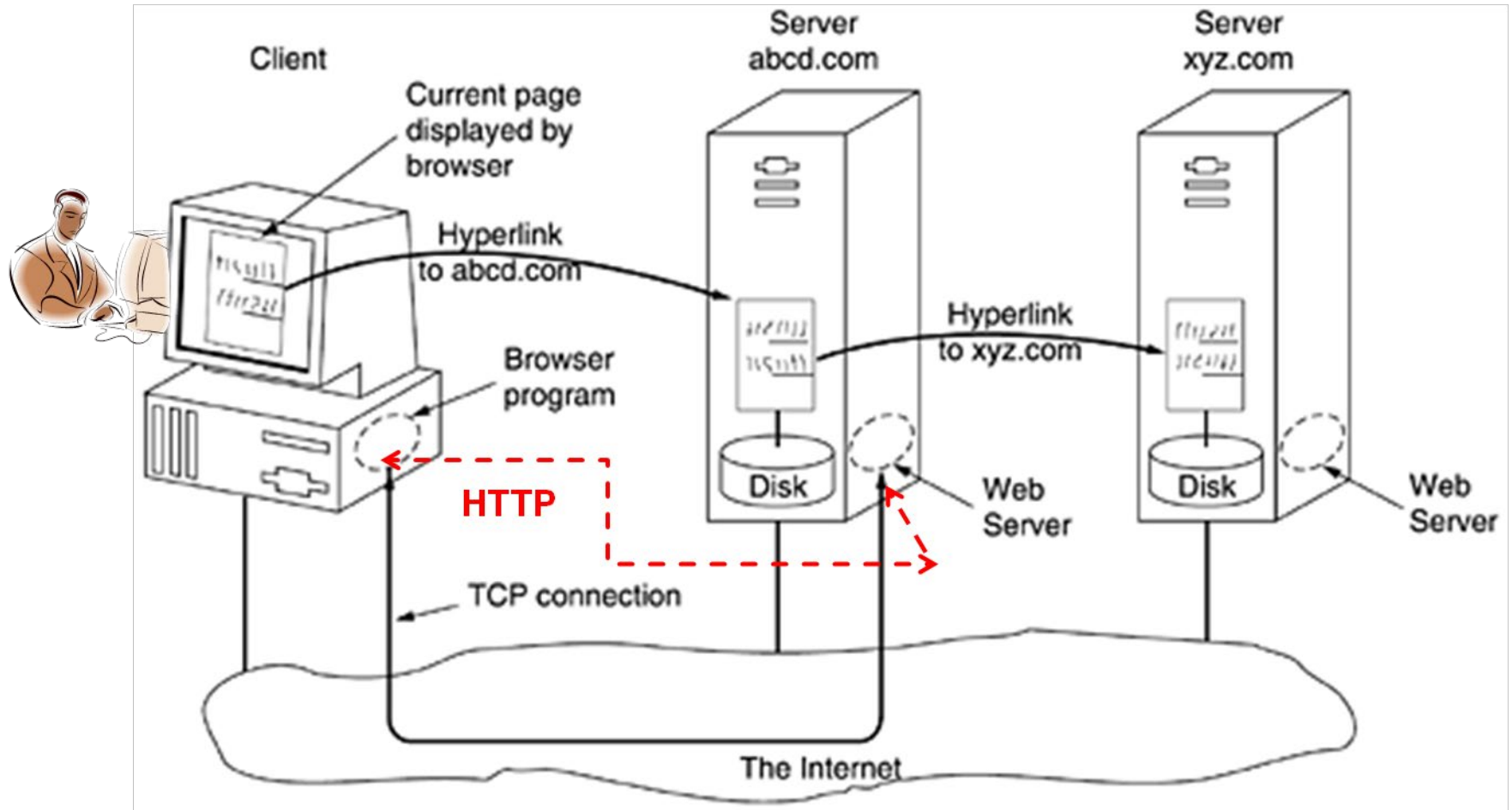


The hyperlinks

A link is made of two parts:
the visible text (or image) and the link to the resource (typically a web page) to be looked for when clicking on the visible text or image

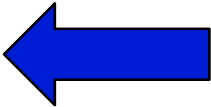


The Web architecture



- Web 1.0 (1993-2003/4)
 - Web is a “publishing medium”
 - Users (humans) can only read
- Web 2.0 (2003/4-today)
 - Web is a “social medium”
 - Users (humans) can publish and interact (e.g. Youtube, Wiki, Flickr, Facebook, etc.)
- Web 3.0 (2010/1-today, more often called IoT - Internet of Things)
 - In addition to humans, users of the Web are “programs” that can interact
 - Users of the Web are “things”, whose programs interact with other things

Refresher on Computer Fundamentals and Networking

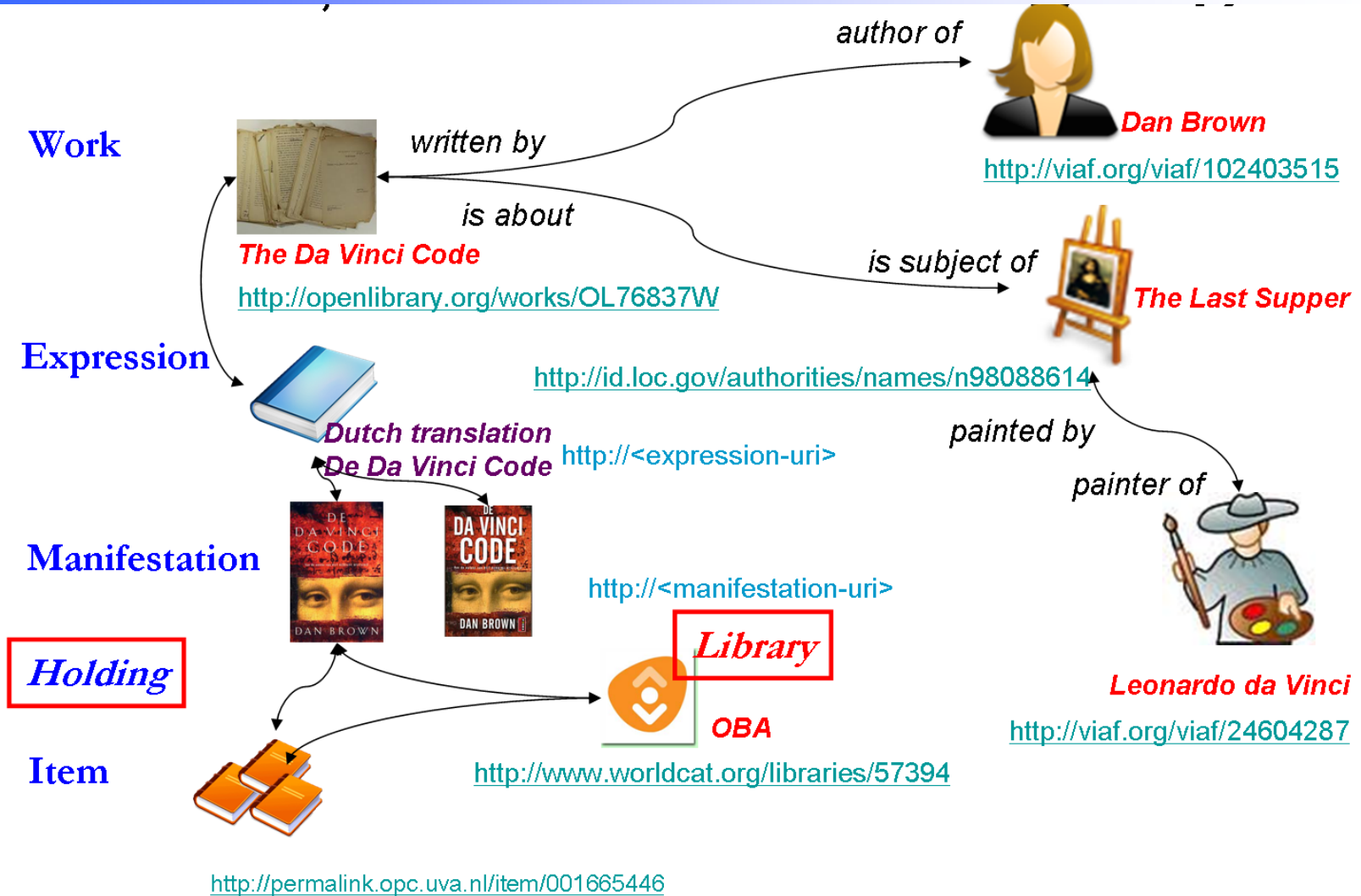
- History of computers
- Architecture of a computer
- Data representation within a computer
- Computer networks and the Internet
- The Semantic Web 

“I have a dream”

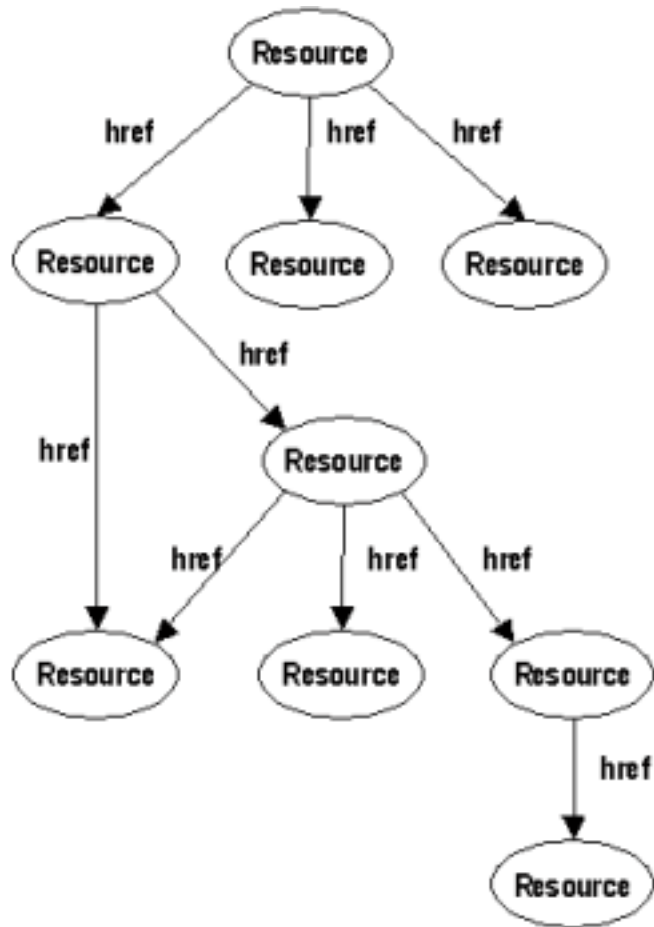
“I have a dream for the Web [in which computers] become capable of analyzing all the data on the Web – the content, links, and transactions between people and computers. A ‘Semantic Web’, which should make this possible, has yet to emerge, but when it does, the day-to-day mechanisms of trade, bureaucracy and our daily lives will be handled by **machines talking to machines**. The ‘intelligent agents’ people have touted for ages will finally materialize”

Tim Berners-Lee, 1999

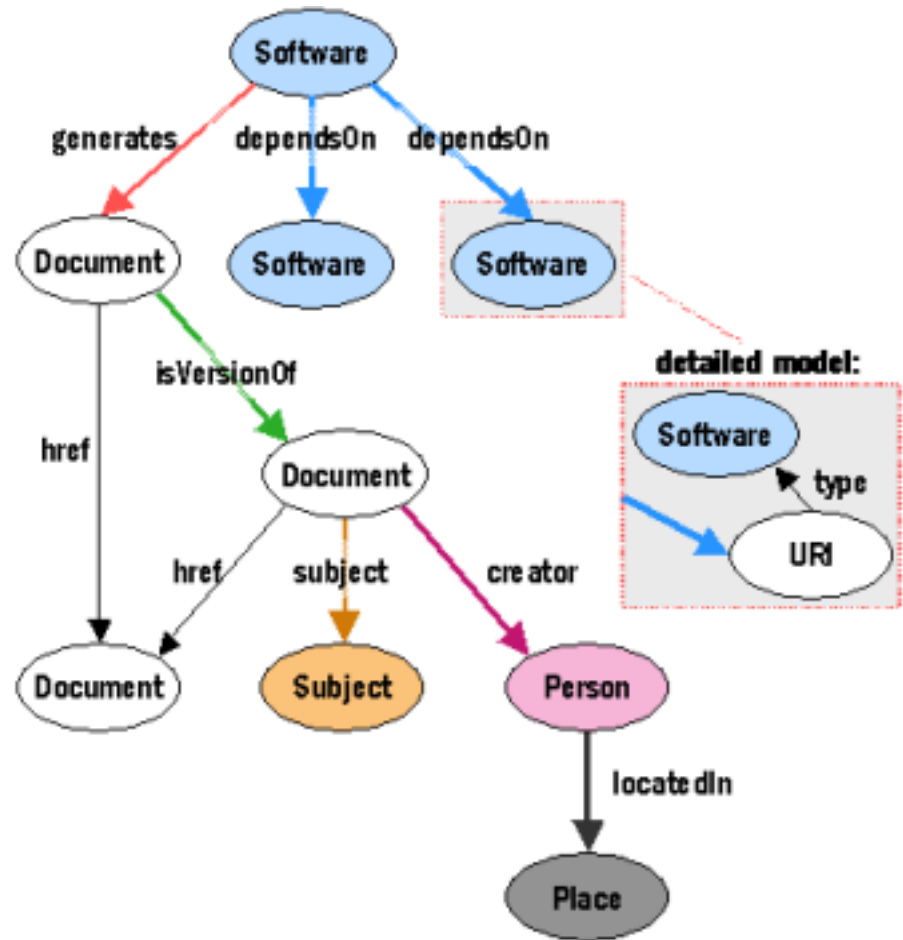
The Web of Linked Data



The Semantic Web



a) Current Web



b) Semantic Web

Marja-Riitta Koivunen and Eric Miller w3.org

- The whole idea of the Semantic Web is to make available (for use in the Web) resources (or resource descriptions) whose “meaning” is understandable by a computer
- This is accomplished by providing descriptions of resources in a “formal way”, so that these descriptions can be “understood” by a computer (i.e. a program running in a computer)
- The first step in approaching this formal description is to define exactly the “portion of the universe” that we want to describe, and then define a “conceptual model” of it
- The conceptual model is then described in a formal notation that can be interpreted by a computer program

- Resource Description Framework (RDF) is a way to represent information about *resources* in the Web (in the World)
- A resource is anything that has identity. For example, a resource may be an electronic document, an image, a service (e.g., "today's weather report for Pisa"), or a collection of other resources. Not all resources are network "retrievable"; e.g., human beings, corporations, and bound books in a library can also be considered resources
- All resources are identified by a URI (Uniform Resource Identifier)
 - a string of characters that unambiguously identifies a particular resource
- *Resources* are described in terms of simple statements specifying properties and property values of resources

Several types of URIs

URI Syntax

<scheme name> : <hierarchical part> [? <query>] [# <fragment>]

any://example.com:8042/over/there?name=ferret#nose

_ / _ / _ / _ / _ /
scheme authority path query fragment

- ftp://ftp.is.co.za/rfc/rfc1808.txt
- <http://www.ietf.org/rfc/rfc2396.txt>
- ldap://[2001:db8::7]/c=GB?objectClass?one
- mailto:John.Doe@example.com
- news:comp.infosystems.www.servers.unix
- tel:+1-816-555-1212
- telnet://192.0.2.16:80/
- urn:oasis:names:specification:docbook:dtd:xml:4.1.2bb



- Resources:
 - An object, an entity or anything we want to talk about (e.g. authors, books, publishers, places, people, facilities)
- Properties:
 - They codify **relations** (e.g. written-by, friend-of, located-in, ...) and **attributes** (e.g. age, date of birth, length ...)
- Statements:
 - Statements assert the properties of resources in form of triples subject-property-value (**subject-predicate-object**)
- Every resource and property has a URI
- Values (the object) can be other resources (for relations) or literals, i.e. terminal values (for attributes)

- A statements is composed of three parts: a subject, a predicate (about the subject), an object (the value of the predicate)
- Example
 - <http://www.example.org/index.html> has a creator whose value is John Smith
- the subject is the resource identified by this URI:
`http://www.example.org/index.html`
- the predicate is the phrase “has a creator”
- the object is the phrase "John Smith“
- To avoid “misunderstandings”, the three components of this statement should be indicated by URIs
 - Subject `http://www.example.org/index.html`
 - Predicate `http://purl.org/dc/elements/1.1/creator`
 - Object `http://www.example.org/staffid/85740`

Additional RDF statements (in natural language)

- <http://www.example.org/index.html>
has a creator
whose value is John Smith
- <http://www.example.org/index.html>
has a creation-date
whose value is August 16, 1999
- <http://www.example.org/index.html>
has a language
whose value is English

Each statement corresponds to a “triple”

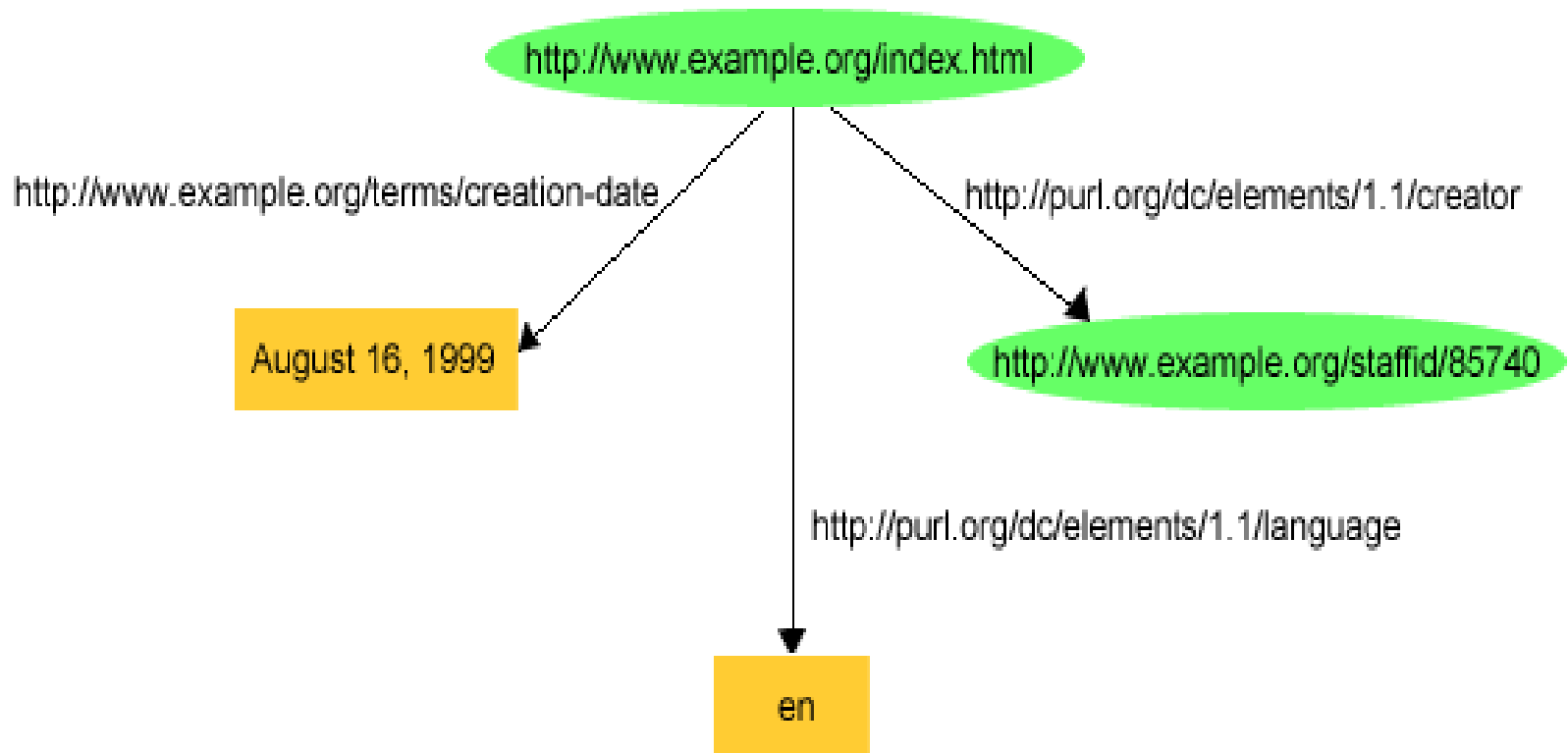
- `<http://www.example.org/index.html>`
- `<http://purl.org/dc/elements/1.1/creator>`
- `<http://www.example.org/staffid/85740>` .

- `<http://www.example.org/index.html>`
- `<http://www.example.org/terms/creation-date>`
- `"August 16, 1999"` .

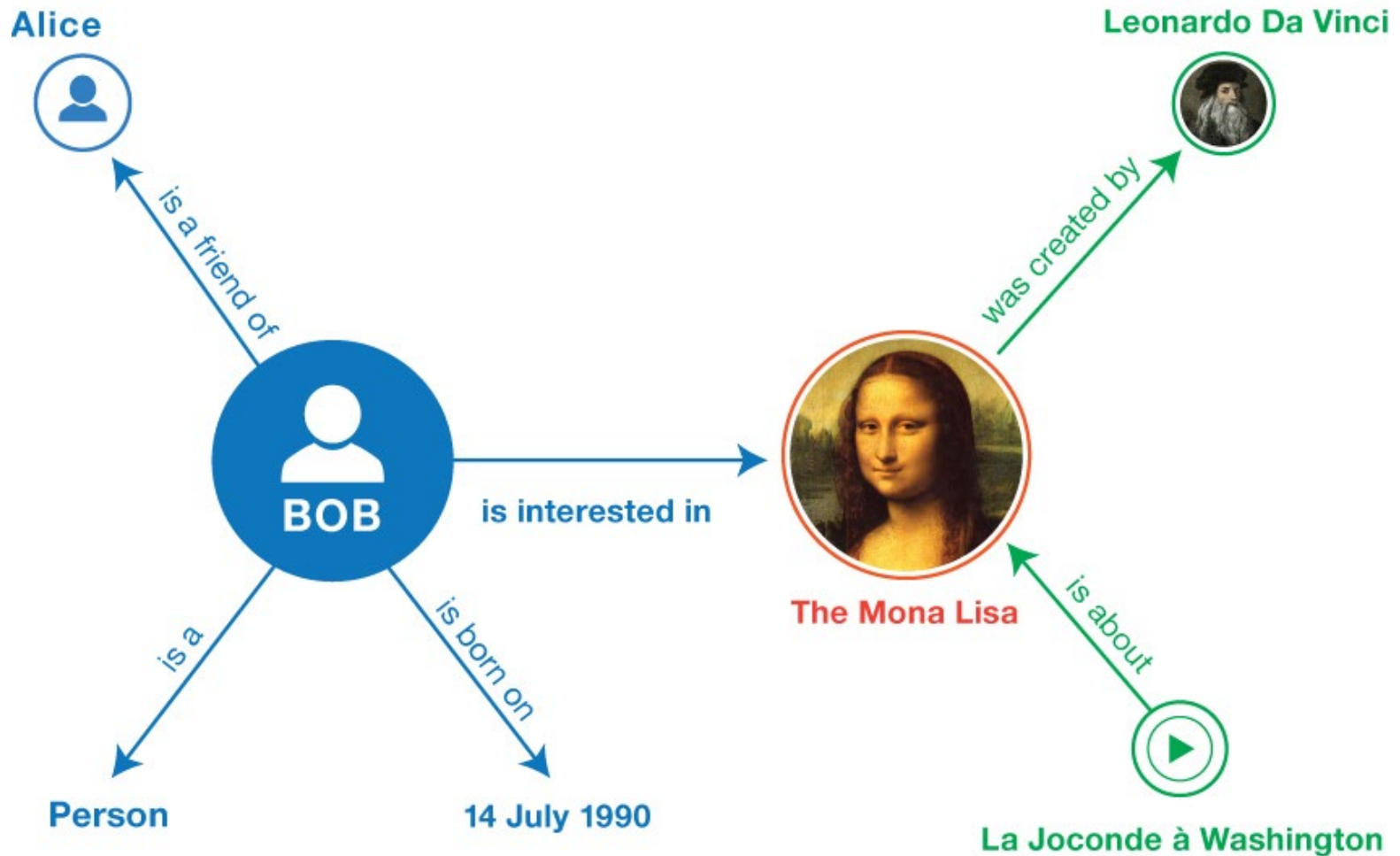
- `<http://www.example.org/index.html>`
- `<http://purl.org/dc/elements/1.1/language>`
- `"en"` .

RDF statements are graphs

Each triple corresponds to an arc in a graph



An RDF graph



Another example of RDF triples

<Bob> <is a> <person>.

<Bob> <is a friend of> <Alice>.

<Bob> <is born on> <the 4th of July 1990>.

<Bob> <is interested in> <the Mona Lisa>.

<the Mona Lisa> <was created by> <Leonardo da Vinci>.

**<the video 'La Joconde à Washington'> <is about>
<the Mona Lisa>**

- A resource can be described by a set of RDF triples
- A set of RDF triples can be represented as a graph
- An RDF triple has three components
 - a **subject**, which is an RDF URI reference
 - **predicate**, which is an RDF URI reference
 - an **object**, which can be:
 - an RDF URI reference
 - an RDF literal

- RDF provides a way to express simple statements about resources, using “named” properties and values
- It is convenient to define the *vocabularies* (terms) that are going to be used in those statements, to indicate that they are describing specific kinds or **classes** of resources, and will use specific **properties** in describing those resources
- For example, to describe bibliographic resources we could define classes such as “Book” or “Journal Article”, and use properties such as “author”, “title”, “borrowedBy” to describe them
- RDF Schema defines the terms used in RDF descriptions by providing a **type system** to be used in the RDF descriptions
- In other words, it provides a way to represent a “conceptual model” of a (small) part of the world, by defining the main “concepts” (classes) in this part of the world, their properties and their relationships

- The main notions of the RDF Schema are:
 - Classes, which can be organized in sub-classes, to any level (defining a taxonomy)
 - Properties, which also can be organized in sub-properties, to any level (defining another taxonomy)
- Vocabulary descriptions (schemas) written in the RDF Schema language are valid RDF graphs
- There is a close analogy with XML documents and XML schemas

- In RDF Schema we have a way to express:
 - that something (i.e. a term in a vocabulary) is a **class** or a **property**
 - that a class is a sub-class of another class
 - that a property is a sub-property of another property
 - that a class is the **domain** of a property
 - that a class is the **range** of a property

- An RDF Schema is a simple “meta” vocabulary used to describe **ontologies**
 - Class, subclassOf, type
 - e.g., Person, Team
 - Property, subPropertyOf
 - e.g., playsFor
 - Domain (the **class for the subjects** of a particular *property*)
 - **Person** *playsFor* Team
 - Range (the **class for the values** of a particular *property*)
 - Person *playsFor* **Team**

<code>ex:Person</code>	<code>rdf:type</code>	<code>rdfs:Class</code>
<code>ex:Book</code>	<code>rdf:type</code>	<code>rdfs:Class</code>
<code>ex:hasAuthor</code>	<code>rdf:type</code>	<code>rdfs:Property</code>
<code>ex:hasAuthor</code>	<code>rdfs:domain</code>	<code>rdfs:Book</code>
<code>ex:hasAuthor</code>	<code>rdfs:range</code>	<code>rdfs:Person</code>
<code>ex:isAuthorOf</code>	<code>rdf:type</code>	<code>rdfs:Property</code>
<code>ex:isAuthorOf</code>	<code>rdfs:domain</code>	<code>rdfs:Person</code>
<code>ex:isAuthorOf</code>	<code>rdfs:range</code>	<code>rdfs:Book</code>

With these definitions of domain and range we are saying that in our (simple) “model of the world” books can only be written by a person, and a person can only write books (unless there are other triples in the schema defining other objects that can be written by a Person)

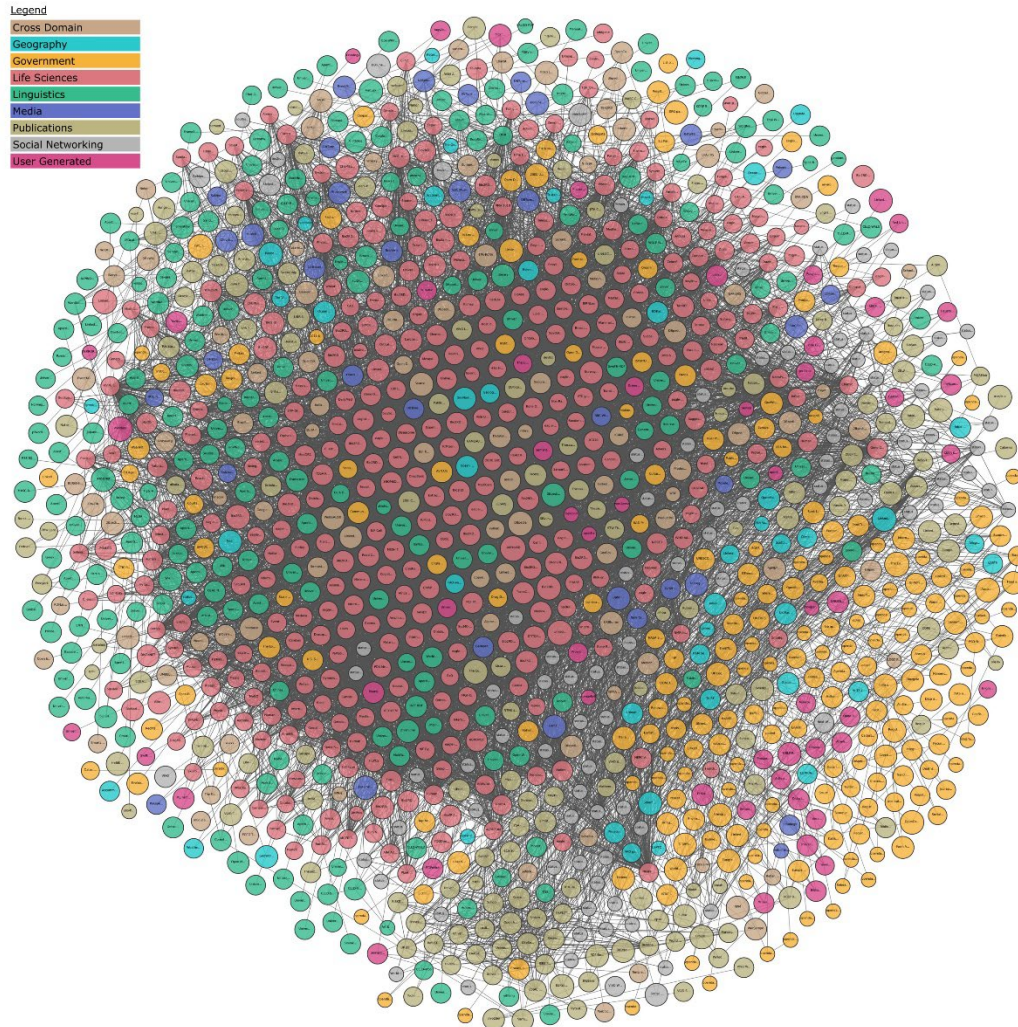

```
ex:hasMother rdf:type      rdfs:Property
ex:hasMother rdfs:range    ex:Female
ex:hasMother rdfs:range    ex:Person
```

```
exstaff:Frank ex:hasMother  exstaff:Mary
```

Mary (the mother of Frank) must be at the same time a person and a female

It is therefore possible to answer queries like:
"List the names of all the females"
and Mary will be in the list, without having ever provided the information that Mary is a female

- The main formalism used today for describing resources is **RDF** – Resource Description Framework
- The RDF descriptions are based on **RDF schemas** (often called **vocabularies or ontologies**), which are also described in RDF (they are the “conceptual models”)
- One of the main initiatives in the Semantic Web is “**Linked Open Data**” (**LOD**), where the resources (or their descriptions) to be made freely available on the Web must be described in RDF and must be linked one to another with “typed links” (i.e. RDF predicates)
- The term Linked Data refers to a set of best practices for publishing and connecting structured data on the Web
- An increasing number of data providers over the last years have contributed to the creation of a global data space containing billions of statements (RDF triples)



The LOD cloud currently contains 1255 datasets

<https://lod-cloud.net/>

Five Star Open Data



- make your stuff available on the Web (whatever format) under an open license



- make it available as structured data (e.g., Excel instead of image scan of a table)



- use non-proprietary formats (e.g., CSV instead of Excel)



- use URIs to denote things, so that people can point at your stuff



- link your data to other data to provide context

Usually Open Data is available under a [CC-BY-SA license](#). This means you can include it in any other work (Creative Commons) under the condition that you give proper attribution (created BY). If you create derivative works (such as modified or extended versions of the Open Data), then you must also license them as CC-BY-SA (Share Alike).



Icon	Description	Acronym	Attribution Required	Allows Remix culture	Allows commercial use	Allows Free Cultural Works	Meets 'Open Definition'
	Freeing content globally without restrictions	CC0	No	Yes	Yes	Yes	Yes
	Attribution alone	BY	Yes	Yes	Yes	Yes	Yes
	Attribution + ShareAlike	BY-SA	Yes	Yes	Yes	Yes	Yes
	Attribution + Noncommercial	BY-NC	Yes	Yes	No	No	No
	Attribution + Noncommercial + ShareAlike	BY-NC-SA	Yes	Yes	No	No	No
	Attribution + NoDerivatives	BY-ND	Yes	No	Yes	No	No
	Attribution + Noncommercial + NoDerivatives	BY-NC-ND	Yes	No	No	No	No