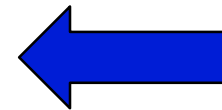Refresher on Computer Fundamentals and Networking

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

Evolution of technology

- 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)

Early computer communication

From mainframe to mainframe
through telephone lines
(point to point connection)
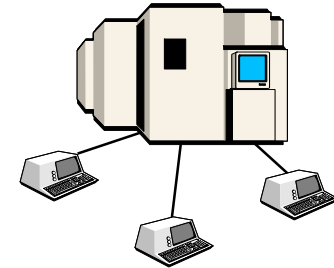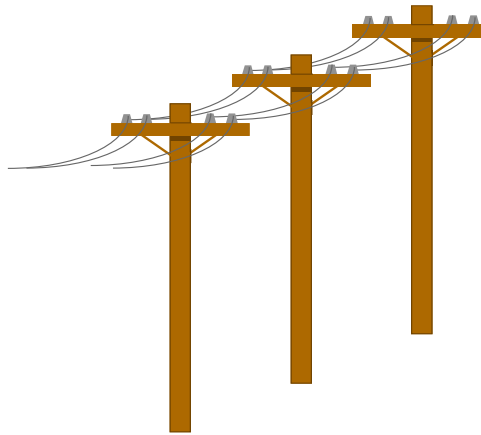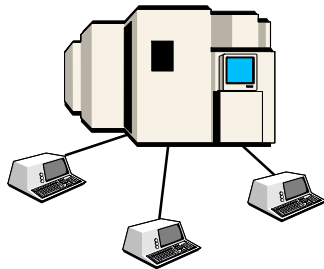
Telephone lines:
slow
expensive
regulated

Networking

- 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

Token ring

Ethernet

LAN switch

Research on networking
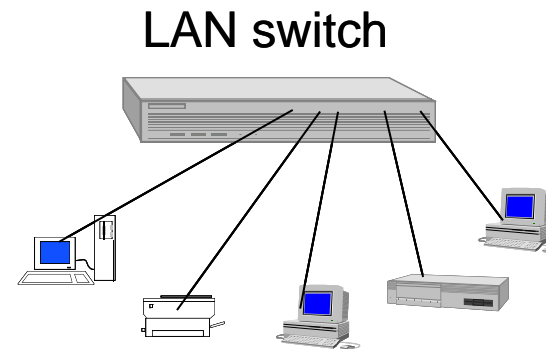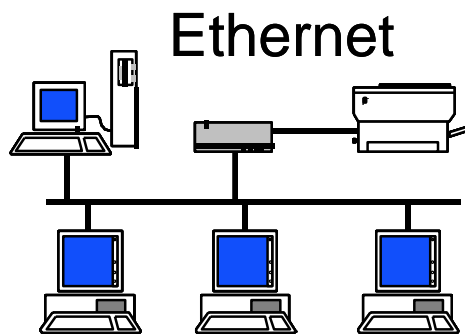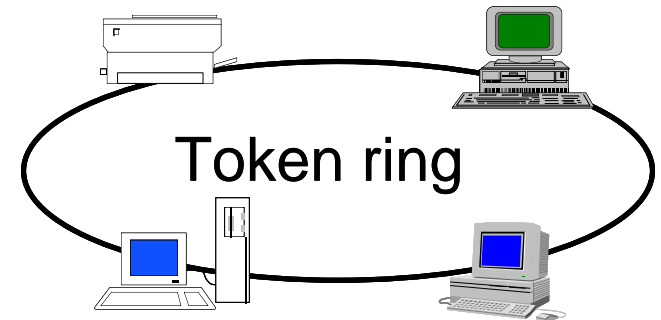
- Starting in the late sixties, many research projects on networking, both from universities and industry
  - Arpanet, Cyclades, SNA (IBM), DECnet
- In the late seventies ISO (International Standard Organization), under pressure of a group of computer manufacturer, started the work for the proposal of 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
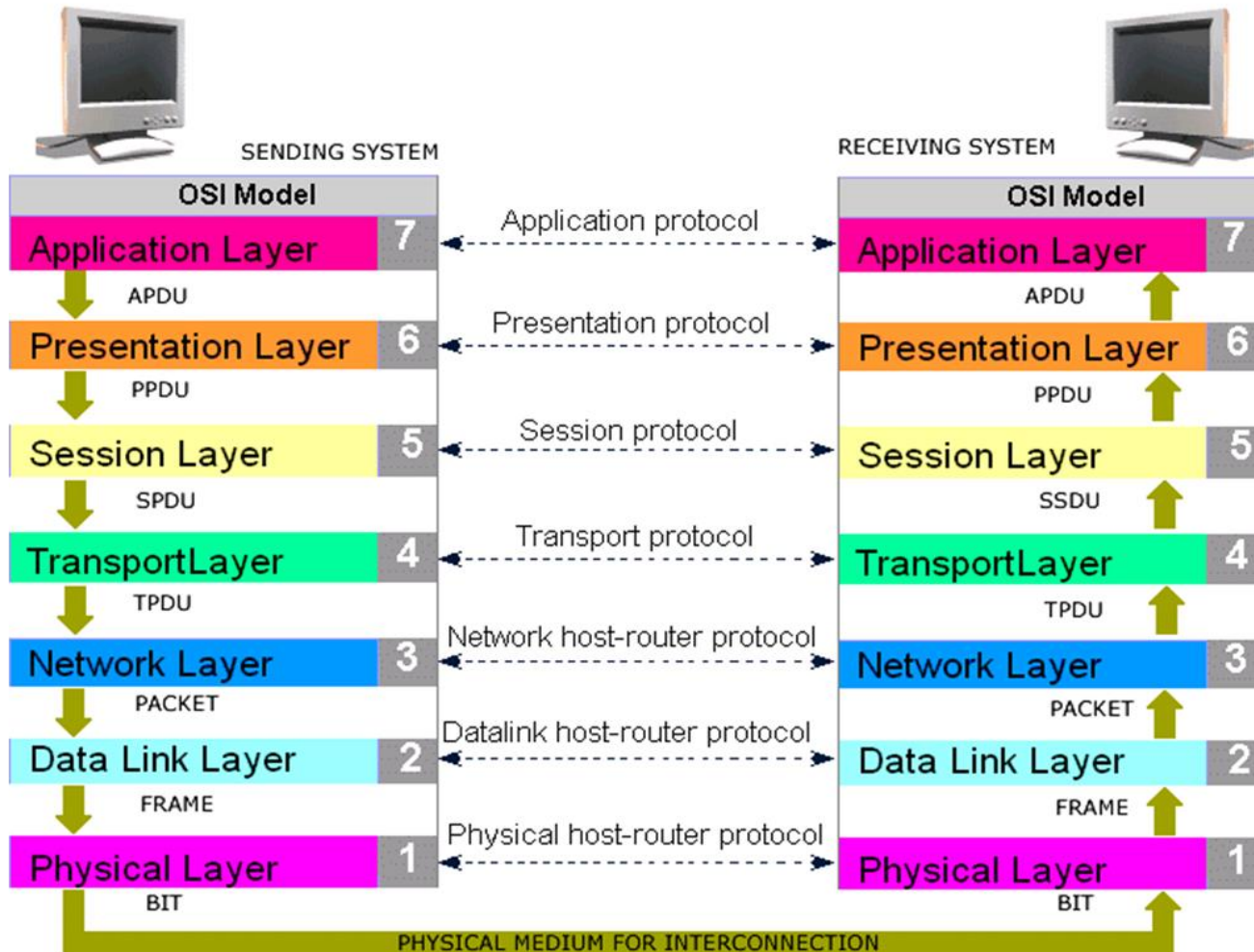
Protocol:
formats and rules for exchanging messages between "partners" (e.g. computers)

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

OSI and Internet

- 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
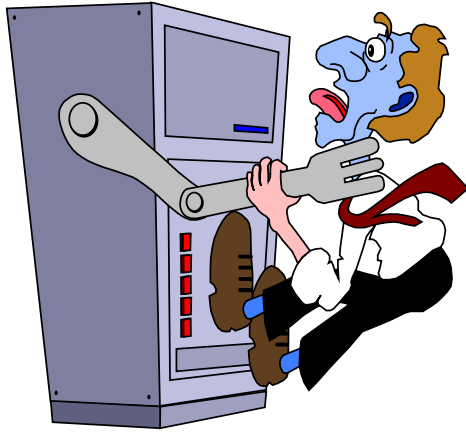
Research
Network
NSF
Internet

Experimental
Network
DARPA
Arpanet
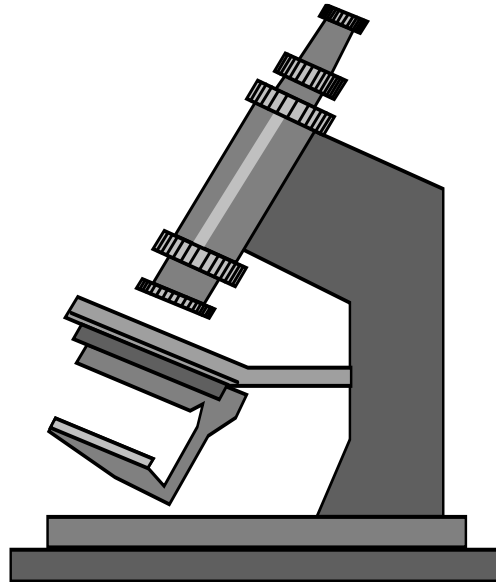
Communication
Infrastructure
Private and public
sectors
The Web

Inter-networking

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

OSI and TCP/IP

OSI Model

7	Application
6	Presentation
5	Session
4	Transport
3	Network
2	Data Link
1	Physical

TCP/IP Model

- Application
- (Host-to-Host) Transport
- Internet
- Network Interface
- (Hardware)
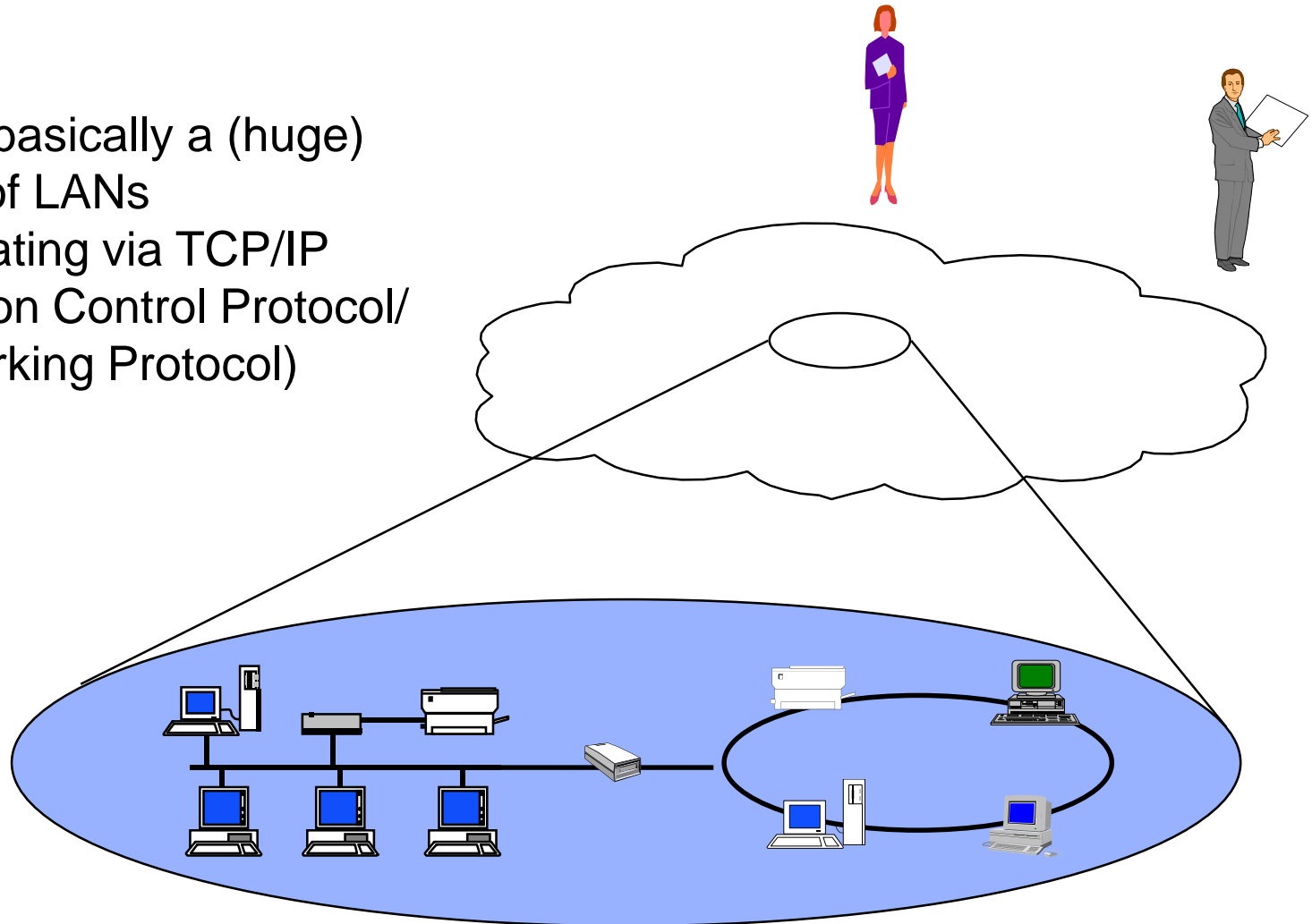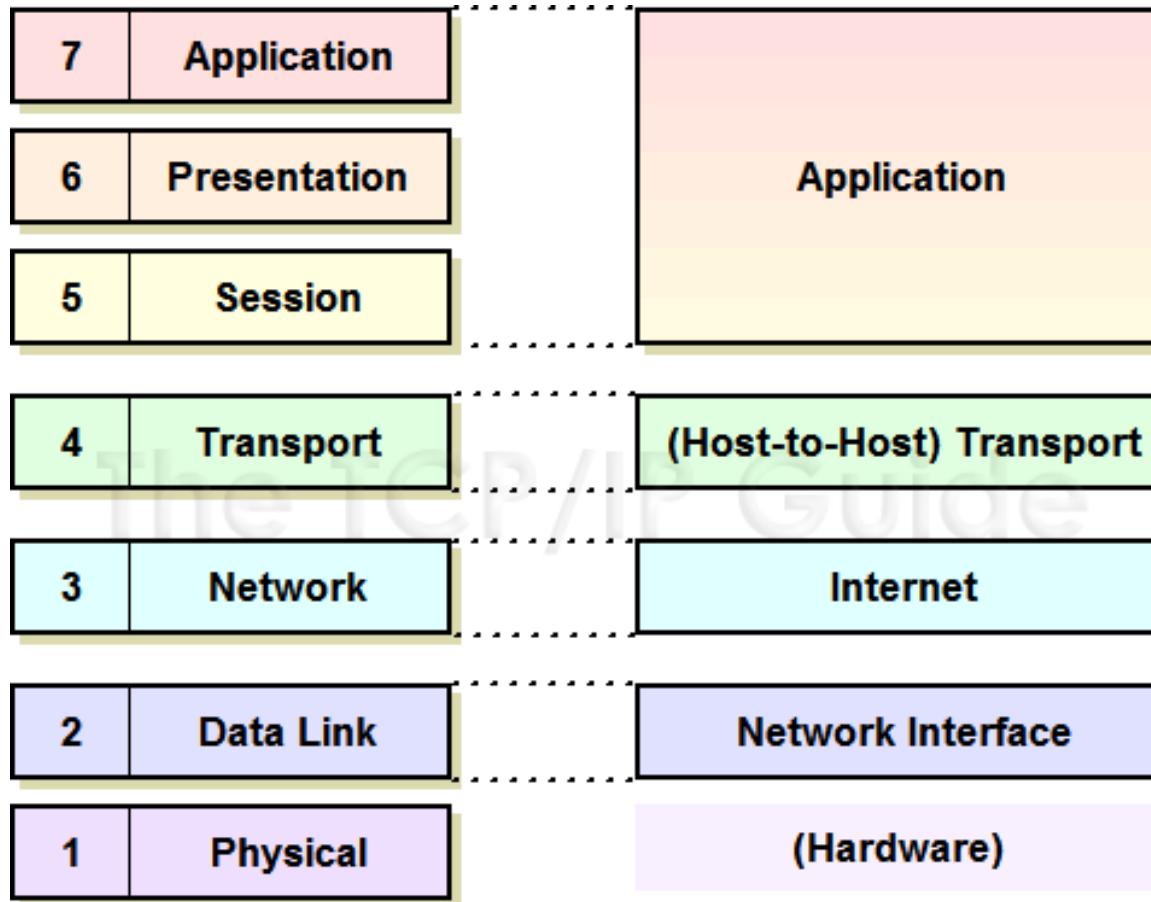
The Internet layers

Layer	Function
Application Services	Email (SMTP), network management (SNMP), WWW (HTTP), and other services
Transport	End-to-end message integrity (UDP, TCP) and error recovery (TCP)
Internet	Connectionless, best-effort packet delivery (IP)
Network Interface	Almost any network technology

Internet protocols

The Internet and the Web

- 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 2020, the total number of web servers is estimated to be in the order of 1200 millions hosts
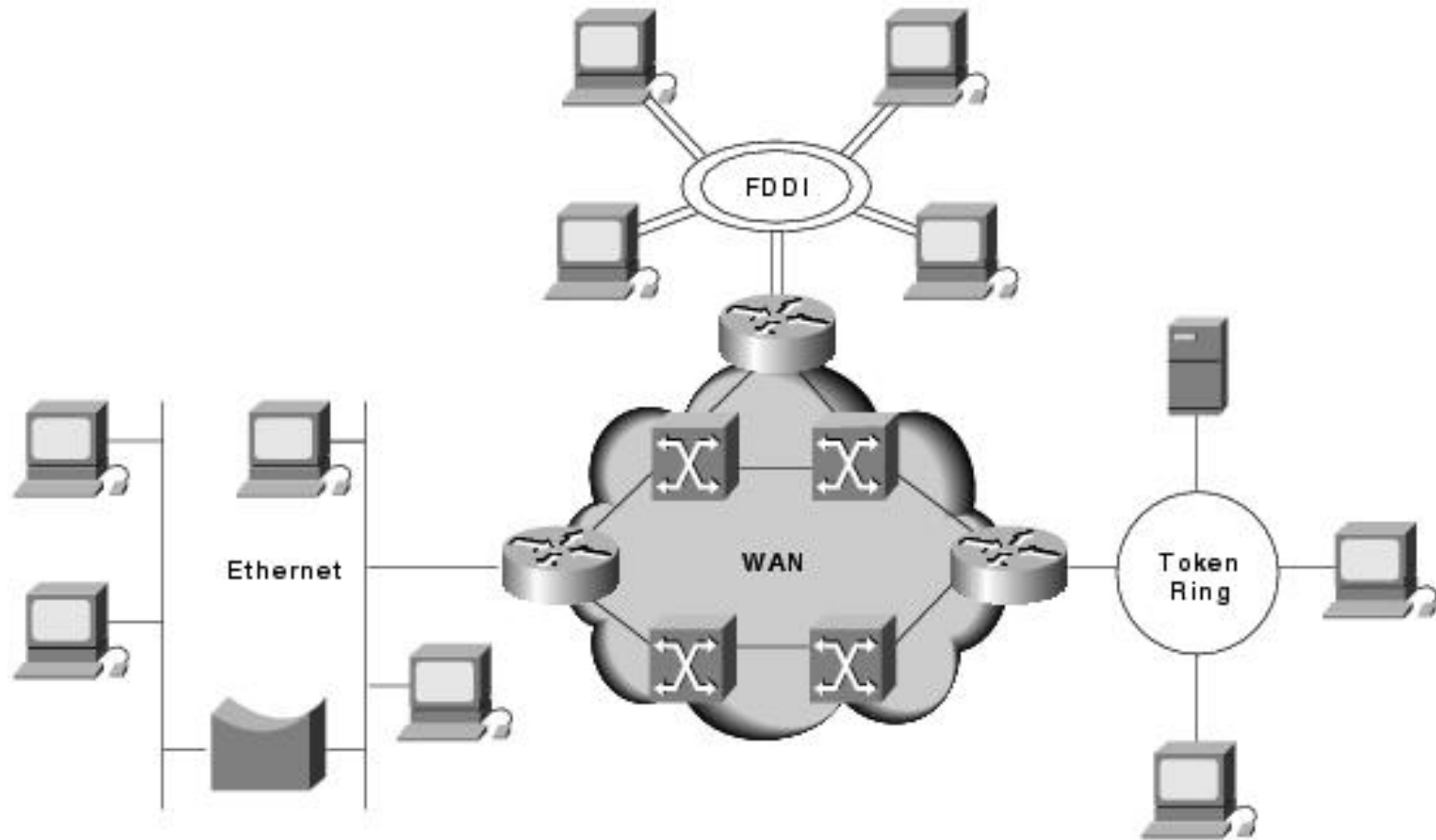 - As of June 2020, the total number of Internet users is estimated to be 4.83 billions users
- The Web is the information space accessible through the Internet
  - As of September 2020, the number of "visible" Web pages (indexed by Google) is estimated to be between 50 and 60 billions pages
  - There 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

Total number of websites (logarithmic scale)

Sep 2020
■ Hostnames: 1,196,298,727
■ Active sites: 192,494,710

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

World Internet usage
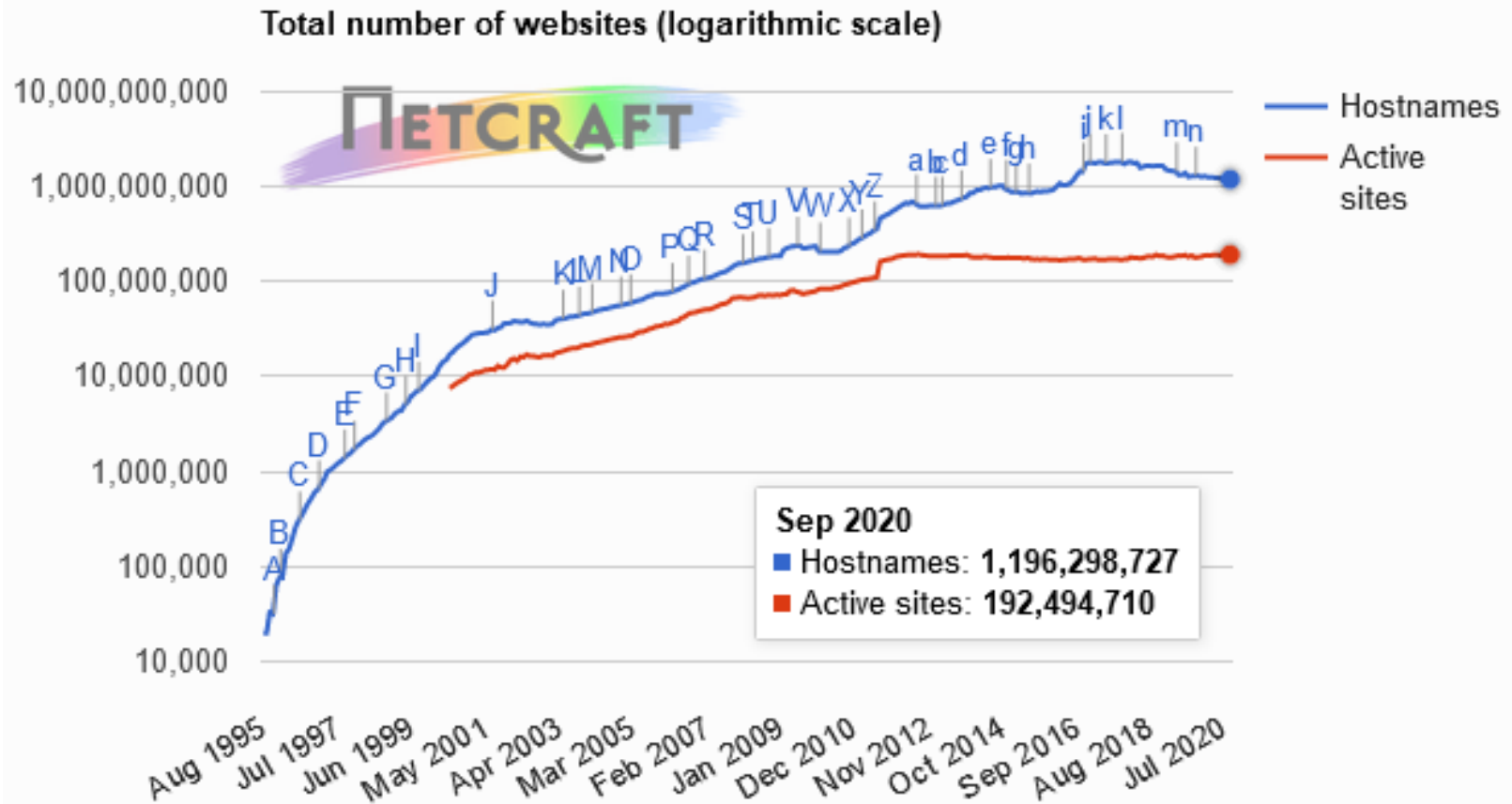
WORLD INTERNET USAGE AND POPULATION STATISTICS
2020 Year-Q2 Estimates

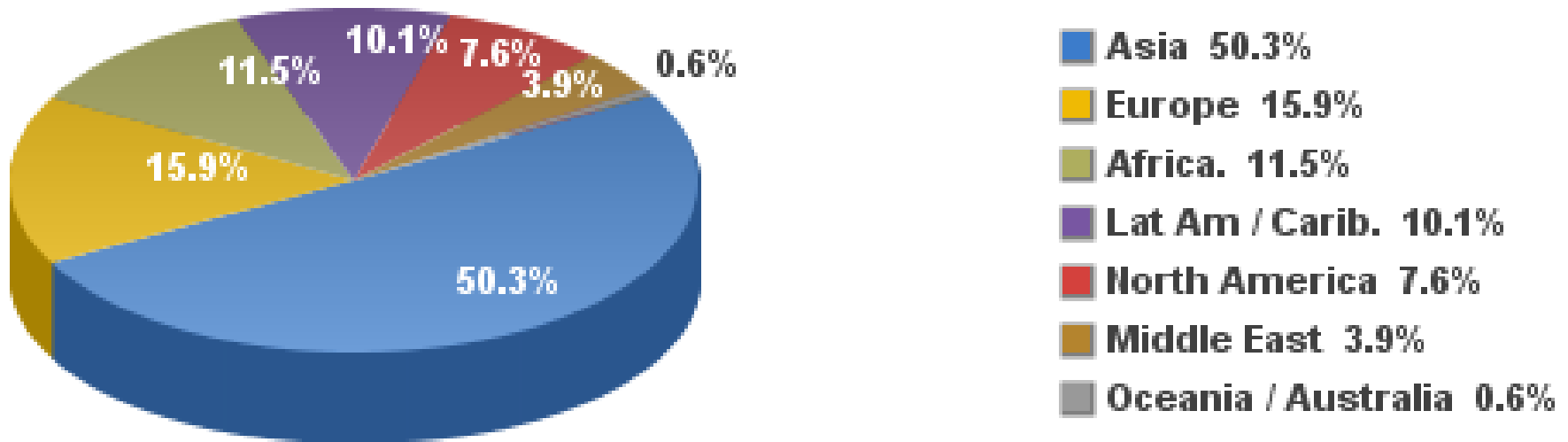| World Regions	Population (2020 Est.)	Population % of World	Internet Users 30 June 2020	Penetration Rate (% Pop.)	Growth 2000-2020	Internet World %
Africa	1,340,598,447	17.2 %	566,138,772	42.2 %	12,441 %	11.7 %
Asia	4,294,516,659	55.1 %	2,525,033,874	58.8 %	2,109 %	52.2 %
Europe	834,995,197	10.7 %	727,848,547	87.2 %	592 %	15.1 %
Latin America / Caribbean	654,287,232	8.4 %	467,817,332	71.5 %	2,489 %	9.7 %
Middle East	260,991,690	3.3 %	184,856,813	70.8 %	5,527 %	3.8 %
North America	368,869,647	4.7 %	332,908,868	90.3 %	208 %	6.9 %
Oceania / Australia	42,690,838	0.5 %	28,917,600	67.7 %	279 %	0.6 %
WORLD TOTAL	7,796,949,710	100.0 %	4,833,521,806	62.0 %	1,239 %	100.0 %

NOTES: (1) Internet Usage and World Population Statistics estimates are for July 20, 2020. (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 and placing a link back to www.internetworldstats.com. Copyright © 2020, Miniwatts Marketing Group. All rights reserved worldwide.

Internet users in the World

Internet Users Distribution in the World - 2020 Q1

Asia	50.3%
Europe	15.9%
Africa.	11.5%
Lat Am / Carib.	10.1%
North America	7.6%
Middle East	3.9%
Oceania / Australia	0.6%

Source: Internet World Stats - www.internetworldstats.com/stats.htm
Basis: 4,574,150,134 Internet users in March 3, 2020
Copyright © 2020, Miniwatts Marketing Group

Internet
World penetration rates

Internet World Penetration Rates by Geographic Regions - 2020 Q1

Region	Penetration Rate
North America	94.6%
Europe	87.2%
Latin America / Caribbean	70.5%
Middle East	69.2%
Oceania	67.4%
World, Avg.	58.7%
Asia	53.6%
Africa	39.3%
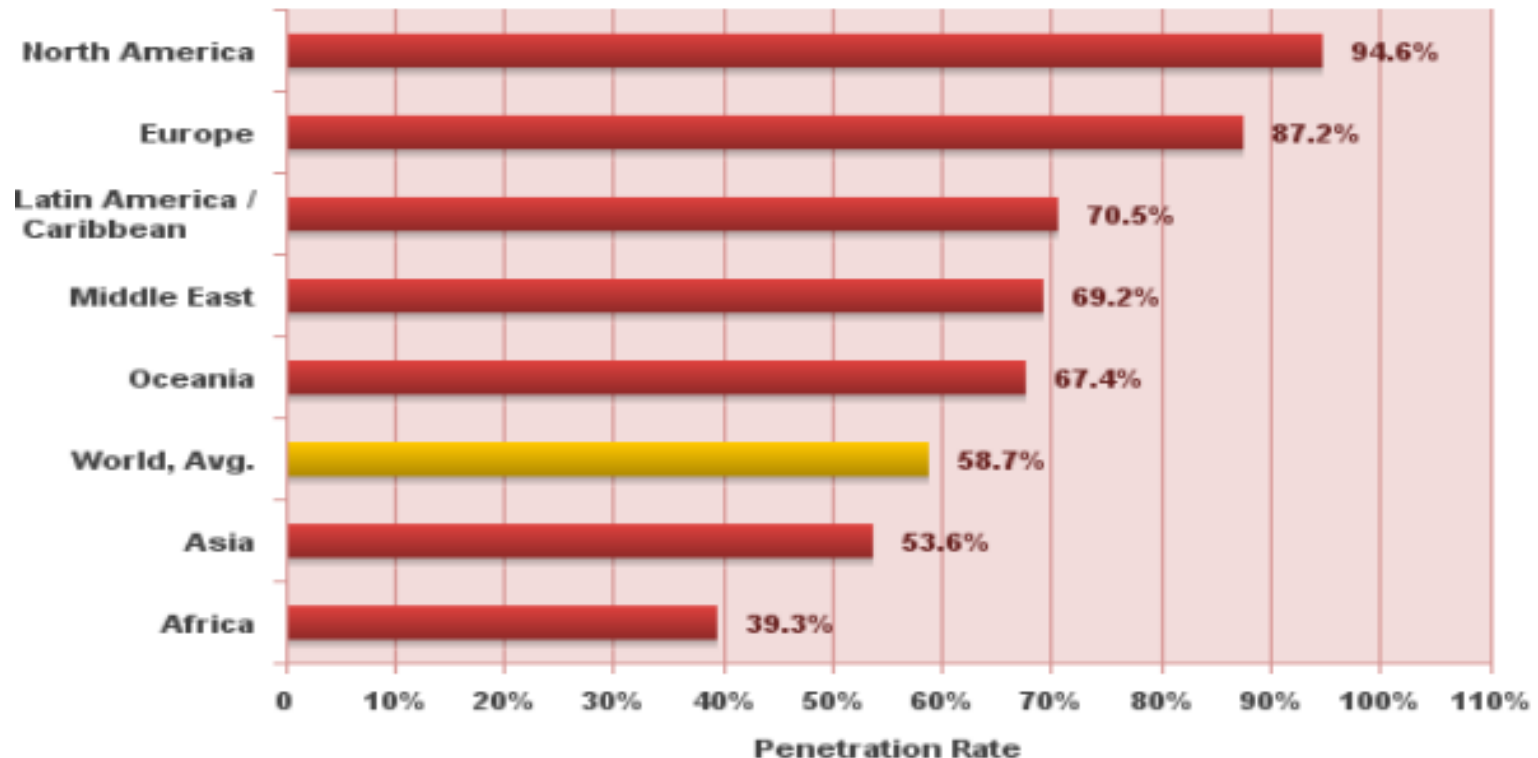
Source: Internet World Stats - www.internetworldstats.com/stats.htm
Penetration Rates are based on a world population of 7,796,615,710
and 4,574,150,134 estimated Internet users in March 3, 2020.
Copyright © 2020, Miniwatts Marketing Group

The Internet and the Web

- 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 2020, the total number of web servers is estimated to be in the order of 1200 millions hosts
  - As of June 2020, the total number of Internet users is estimated to be 4.83 billions users
- The Web is the information space accessible through the Internet
  - As of September 2020, the number of "visible" Web pages (indexed by Google) is 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

The World Wide Web

- 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 (or IoT around 2010)
- Today we have:
  - An estimate of about 1200 million hosts
  - An estimate of 50 to 60 billion pages on line

Evolution of the Web

- 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, 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

The editors

- 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
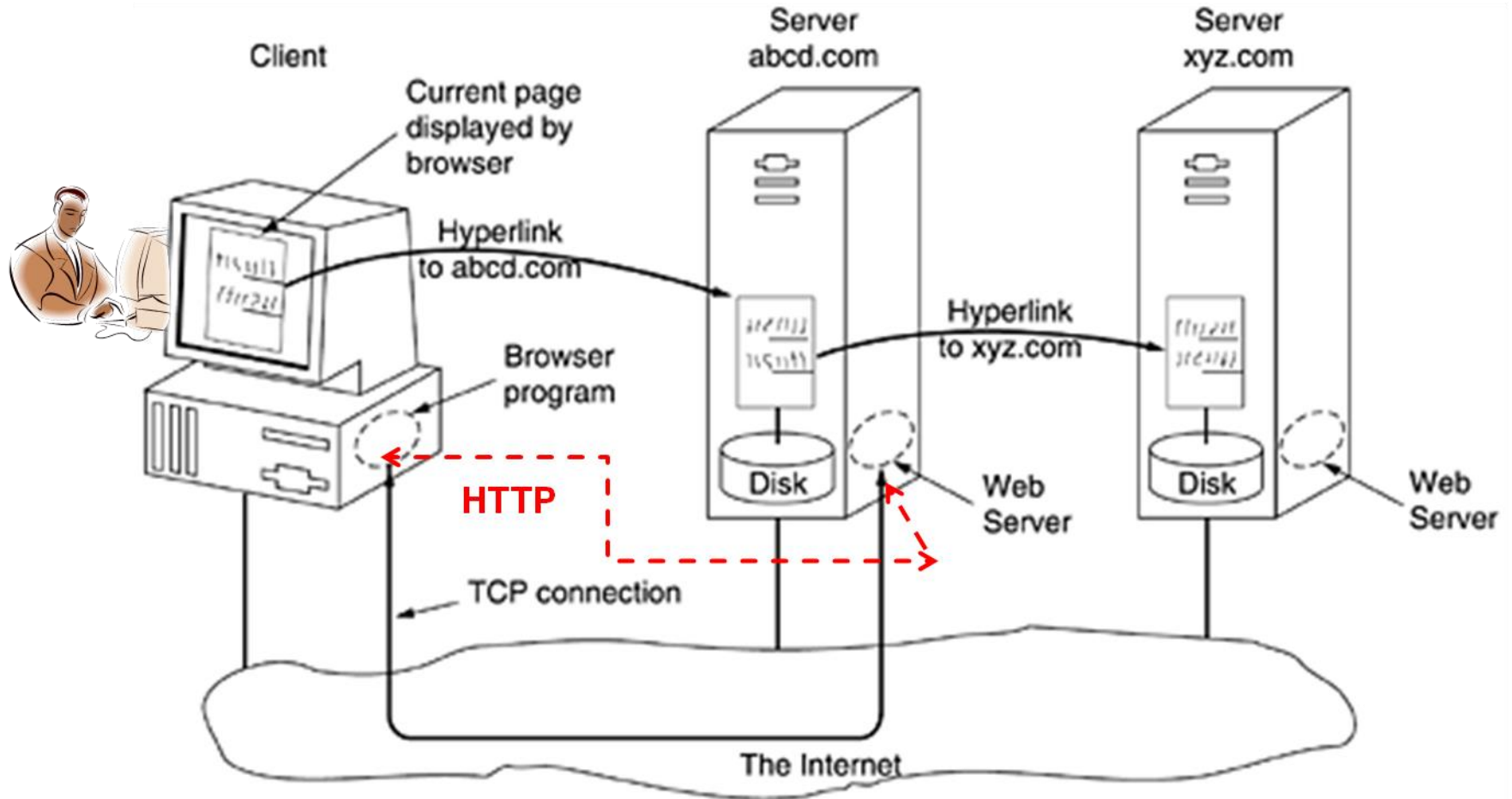
The hyperlink

- 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

The browser

- 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

Vittore Casarosa - University of Pisa and ISTI-CNR

The Web architecture

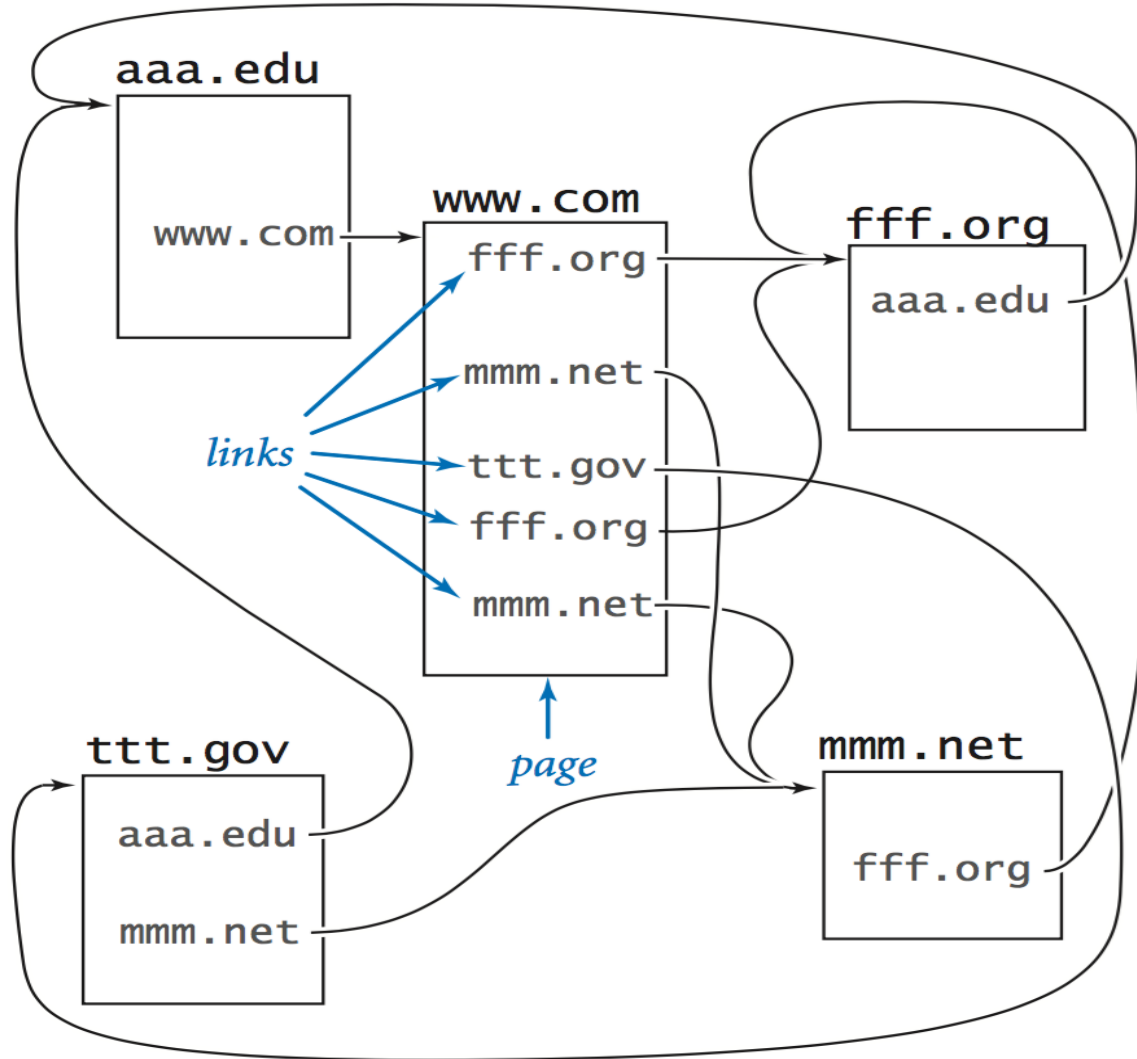# The hyperlinks

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

Refresher

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
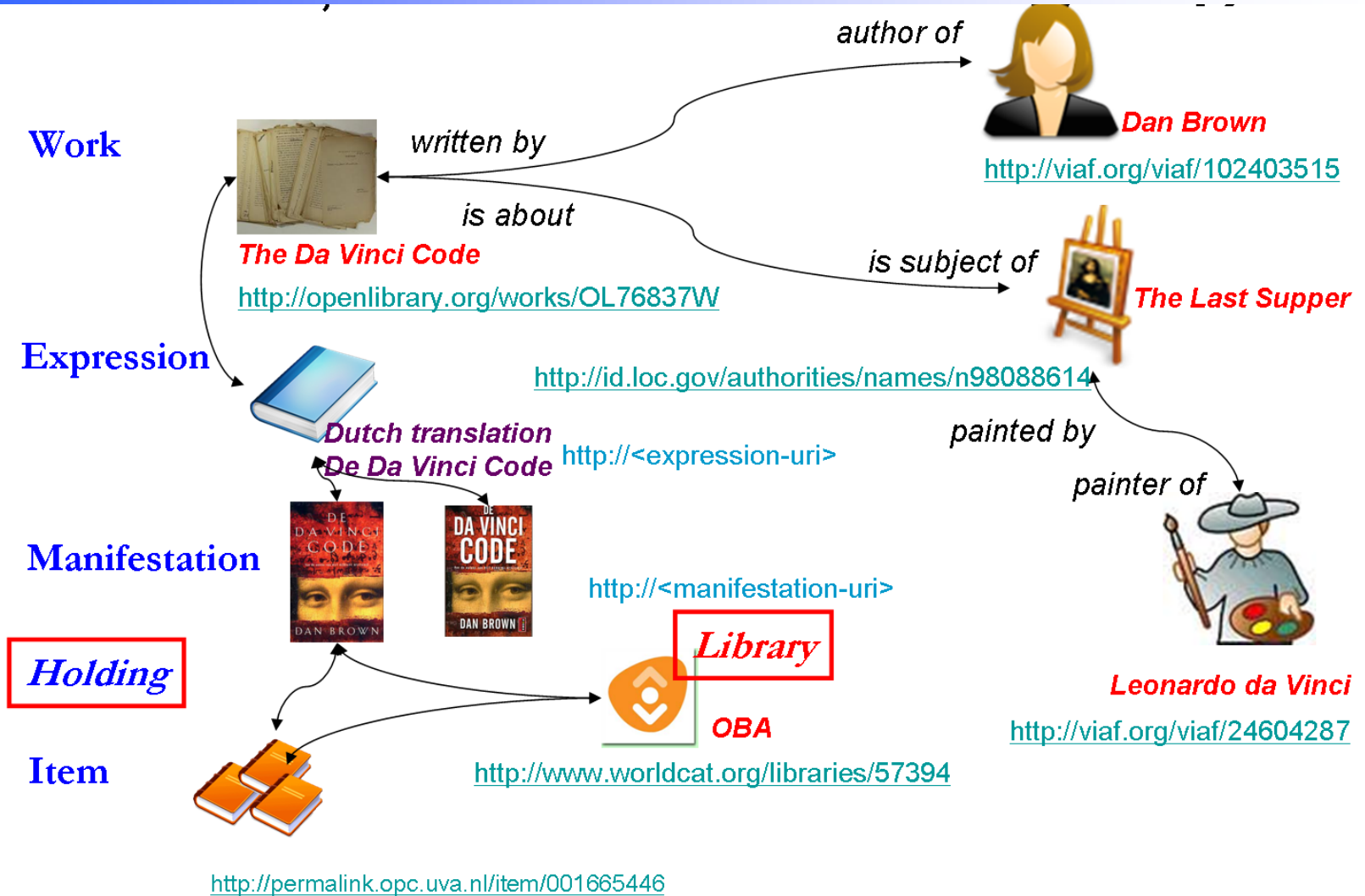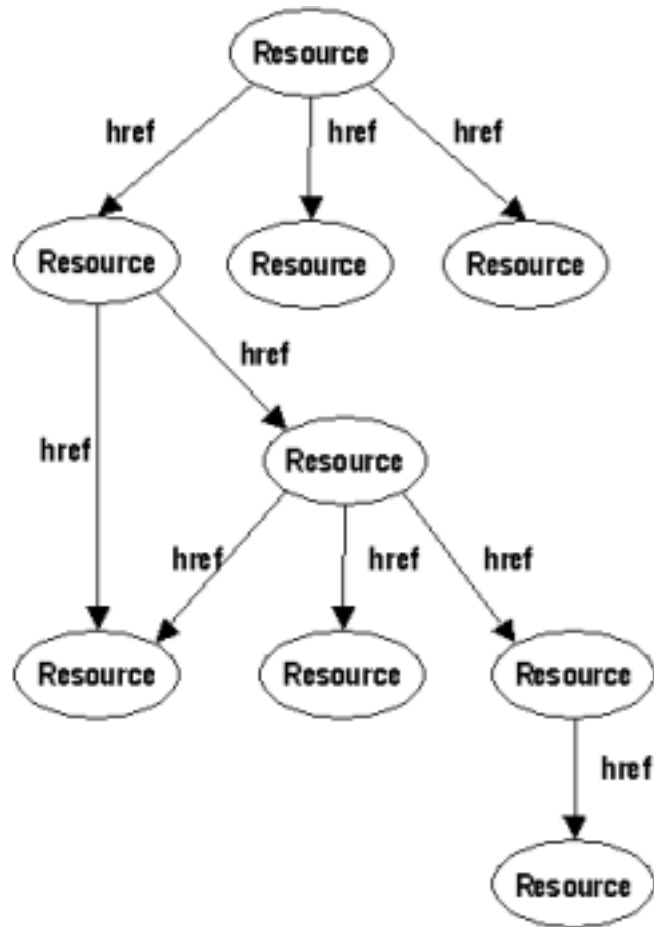
The Semantic Web

a) Current Web
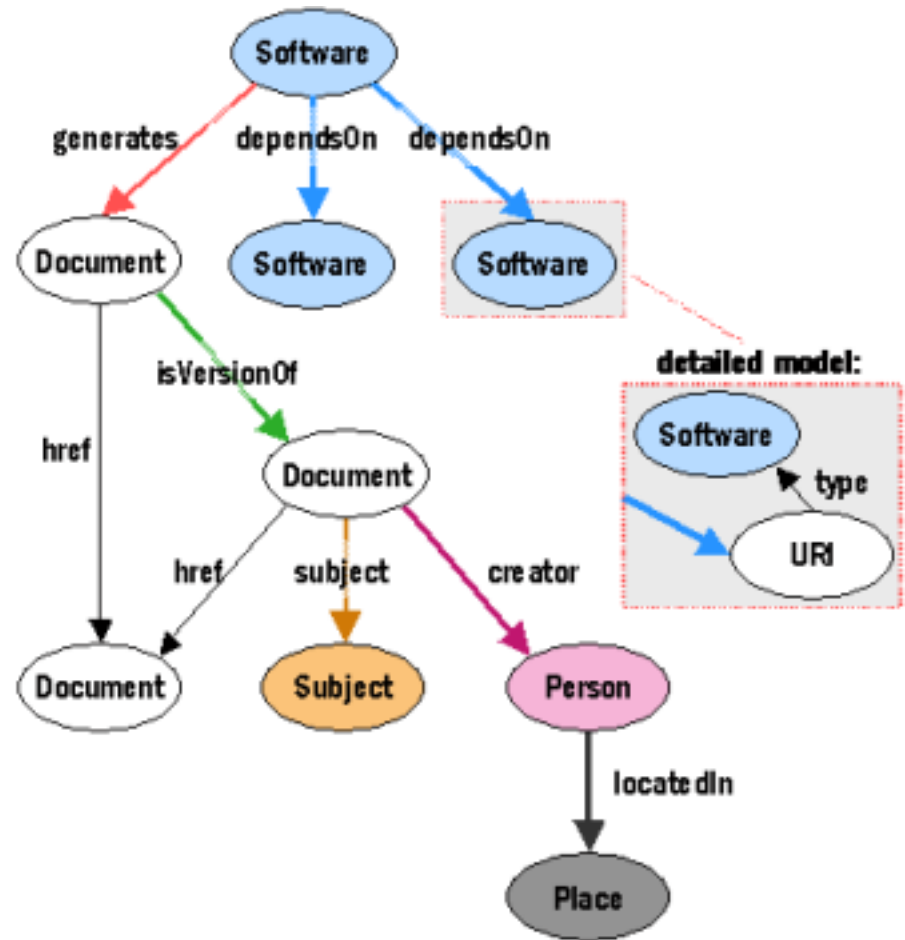
b) Semantic Web

Marja-Riitta Koivunen and Eric Miller w3.org

The Semantic Web

- 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

Resource Description Framework

- 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 Los Angeles"), and 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


```
communication    Web server              folder path        HTML file
protocol         domain name             (optional)         (optional)

http://www.weather.com/weather/us/zips/54701.html
```

RDF statements

- 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 (an URL or any other identifier)
- Values can be resources (for relations) or literals (for attributes)

Simple RDF statement

- 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

RDF statements as triples

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>
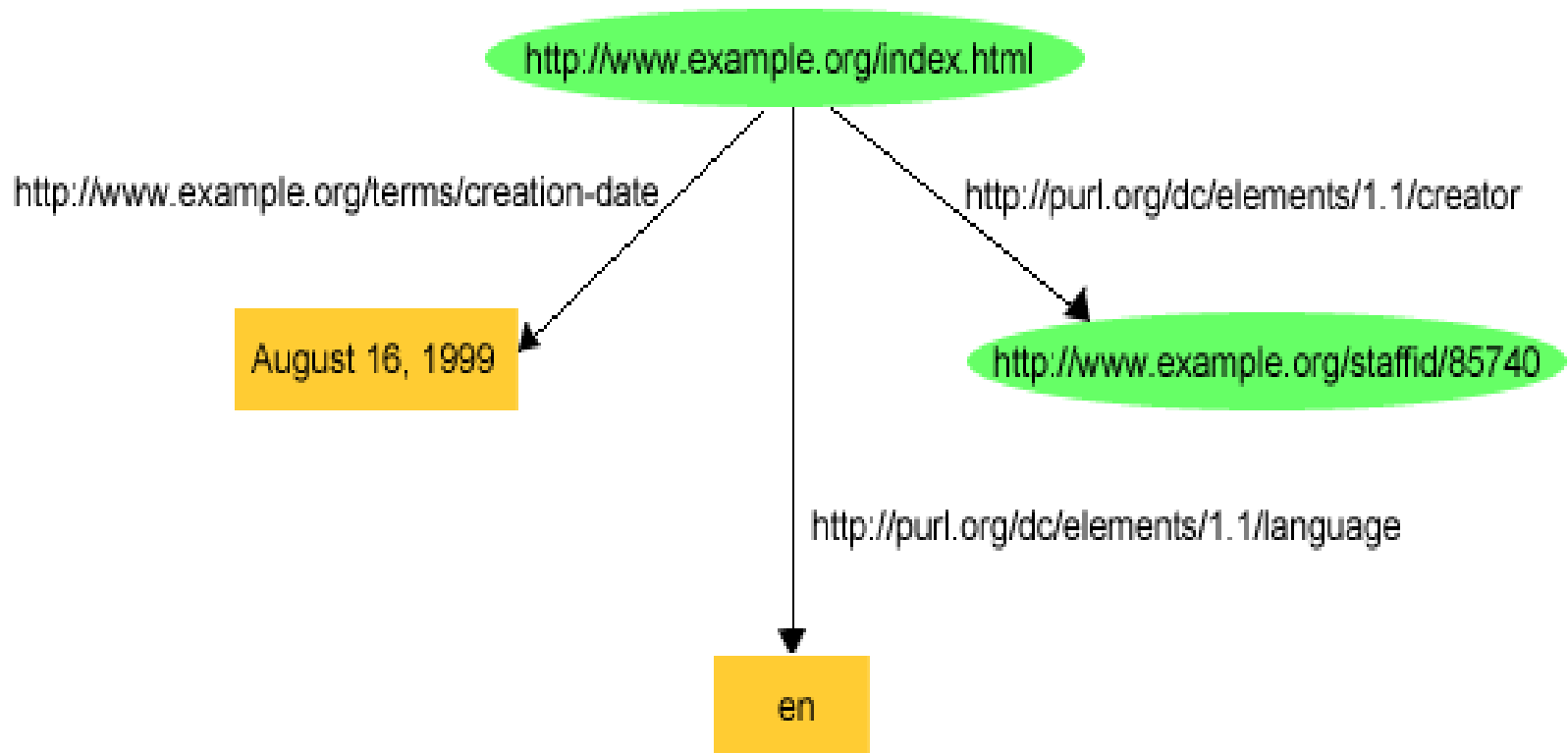- "August 16, 1999" .

- <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

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>
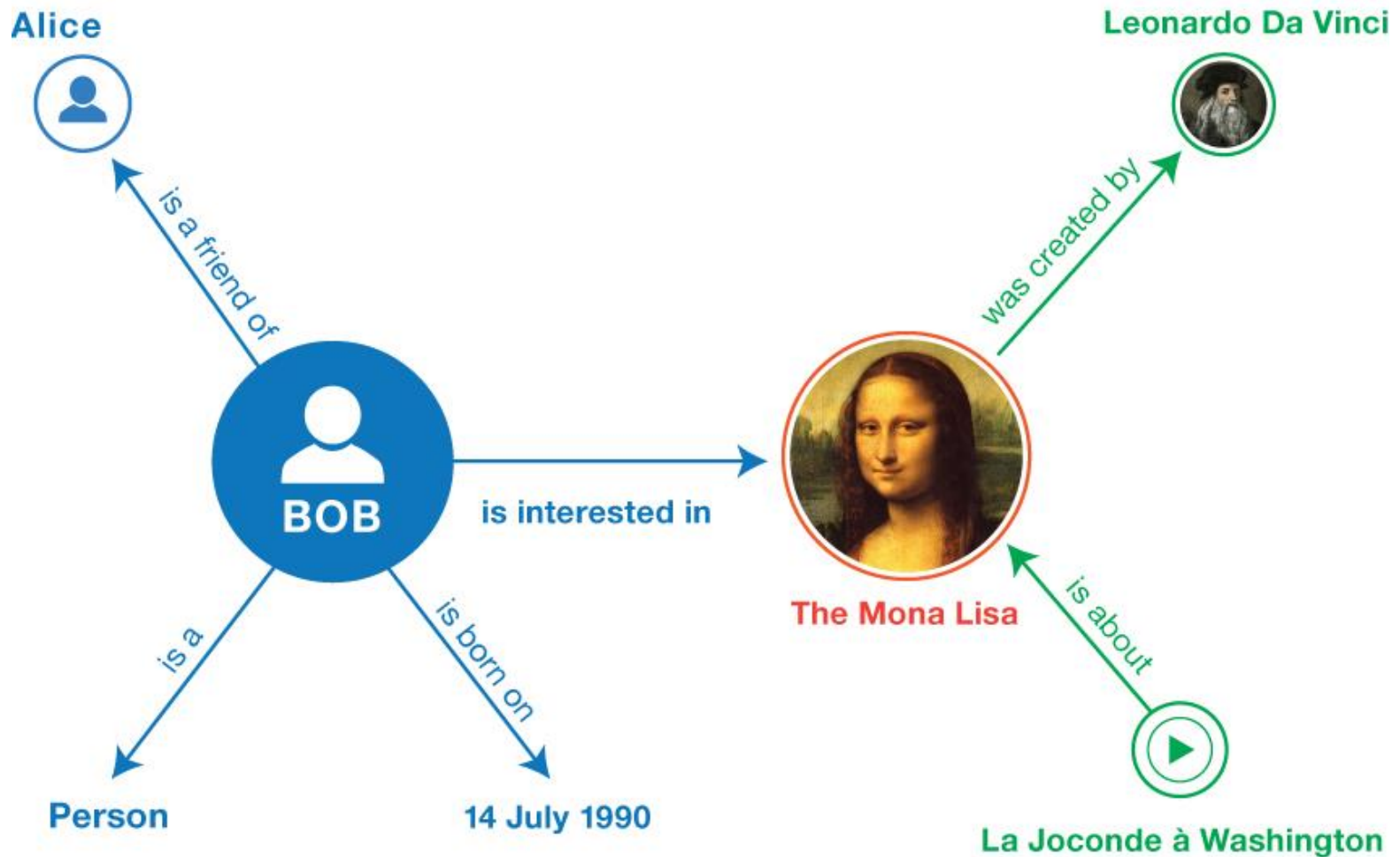
RDF summary

- 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 Schema

- 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
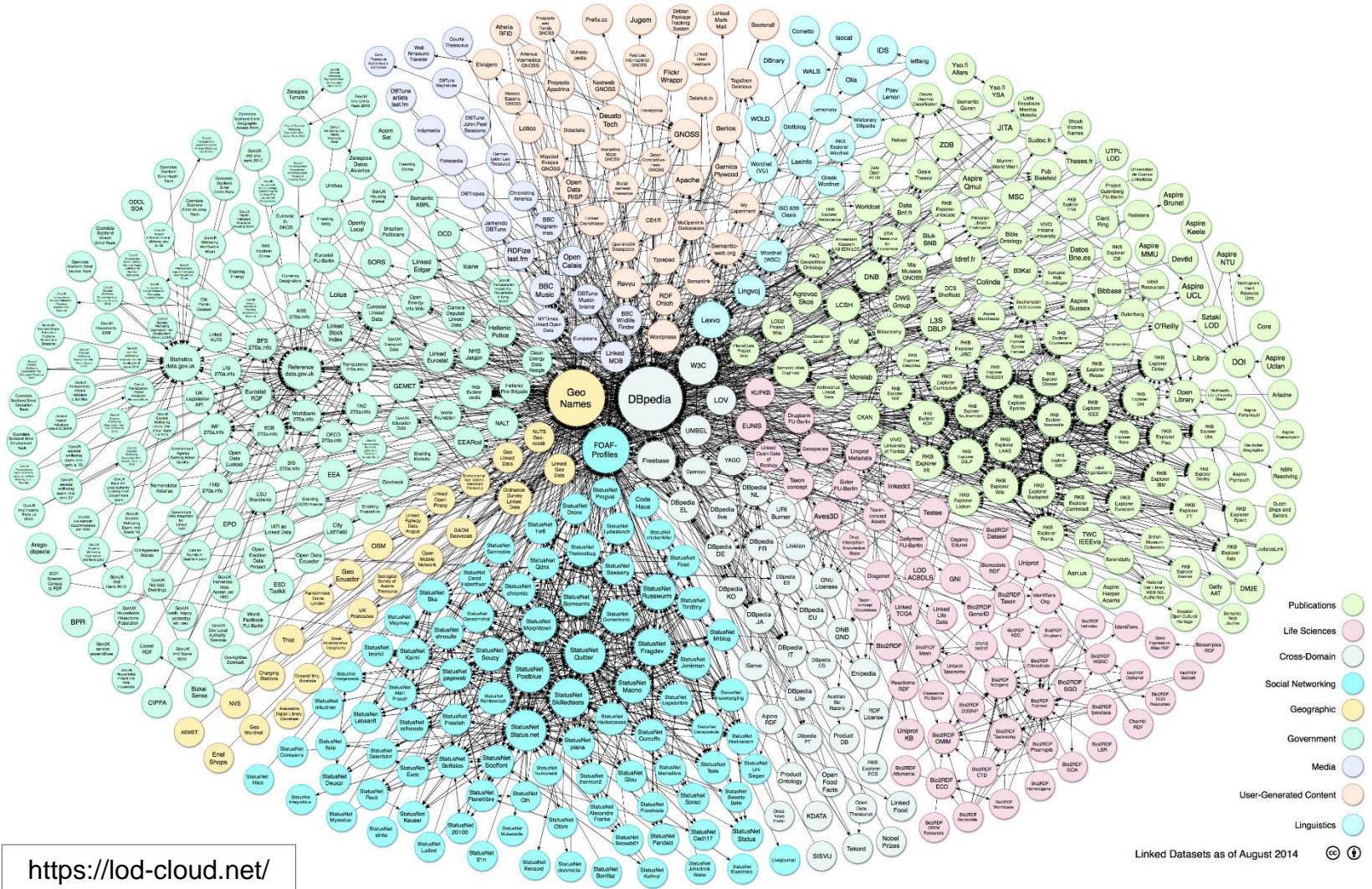
Basics of RDF Schema

- ● In RDF Schema we have a way to express:
  - – that something 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

Main notions of RDF Schema

- 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

Summarizing the RDF Schema

- 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*)
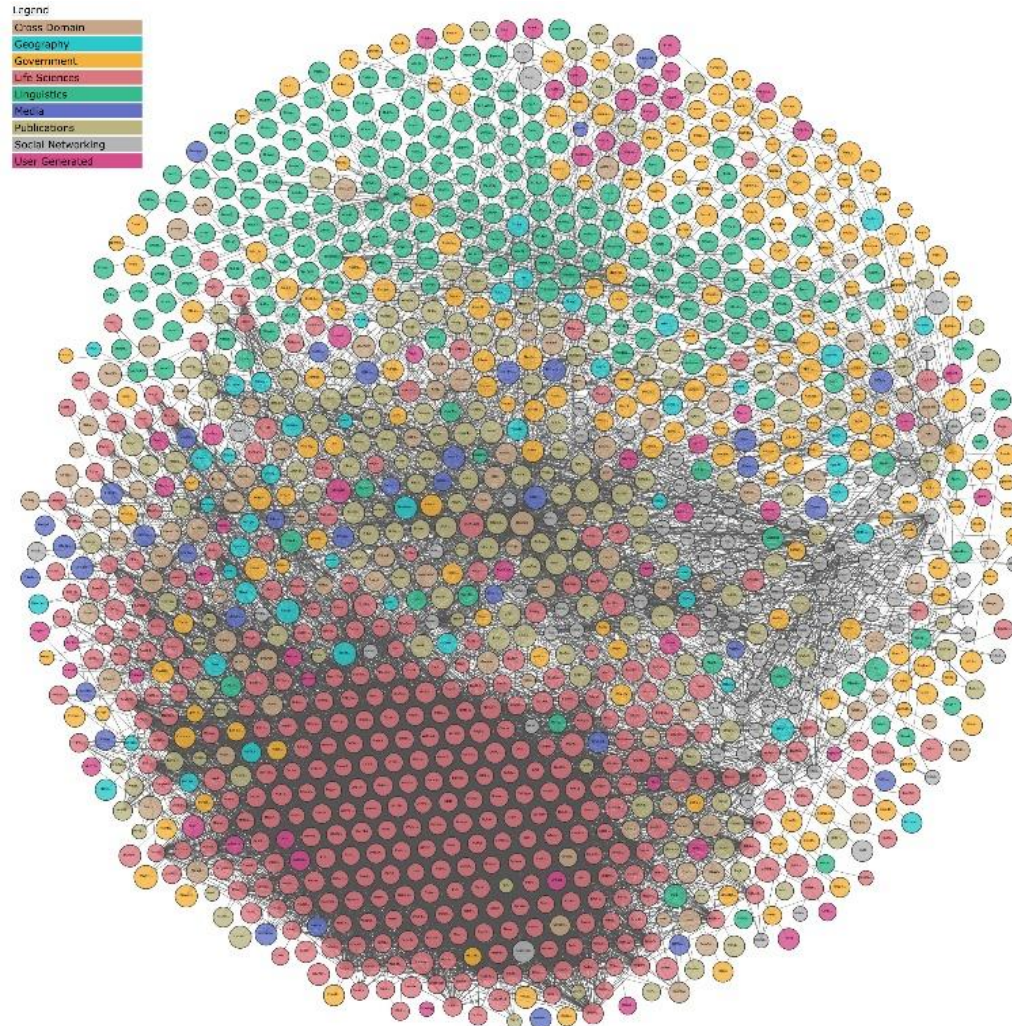   - Person *playsFor* **Team**

LOD and the Semantic Web

- 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)

LOD in the Web (August 2014)

https://lod-cloud.net/

Linked Datasets as of August 2014

Publications
Life Sciences
Cross-Domain
Social Networking
Geographic
Government
Media
User-Generated Content
Linguistics

The dataset currently contains 1269 datasets with 16201 links

https://lod-cloud.net/

Linked Data principles

- Use URIs as names for things

- Use HTTP URIs so that people can look up those names.

- When someone looks up a URI, provide useful information, using the standards (RDF*, SPARQL)

- Include links to other URIs, so that they can discover more things.

Using "de facto standards"

- Different communities have specific preferences on the vocabularies they prefer to use for publishing data on the Web.

- The Web of Data is therefore open to arbitrary vocabularies being used in parallel.

- Despite this general openness, it is considered good practice to reuse terms from well-known RDF vocabularies such as FOAF, SKOS, DOAP, vCard, Dublin Core, or Good Relations wherever possible in order to make it easier for client applications to process Linked Data.

- Only if these vocabularies do not provide the required terms should data publishers define new, data source-specific terminology

SKOS example

animals
cats
    UF domestic cats
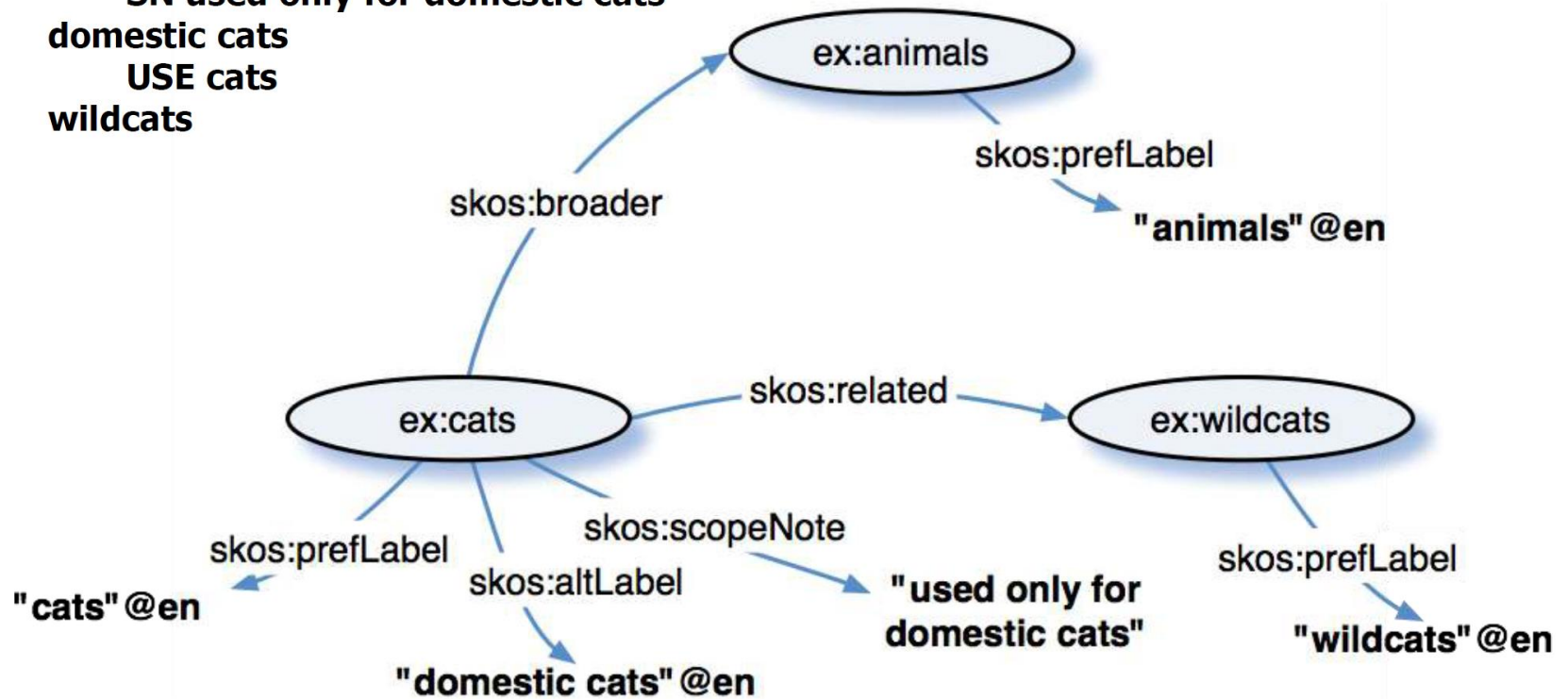   **RT wildcats**
    BT animals
    SN used only for domestic cats
domestic cats
    USE cats
wildcats

Simple Knowledge Organization System

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).

Creative Commons

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