

Libraries


- Description
 - Bibliographic records
 - MARC
- Interoperability
 - Z39-50
- Conceptual models
 - FRBR /LRM for Works, Expr., Manif.
- Information Retrieval
 - Full text of catalogues

The Web

- Description
 - Metadata
 - Dublin Core
- Interoperability
 - OAI-PMH
- Conceptual models
 - RDF and RDF Schema for all resources (ontologies)
- Information Retrieval
 - Full text of resources

Diachronic (but similar) evolution of the libraries and the Web

- Bibliographic records and metadata
 - MARC vs Dublin Core
- Interoperability (exchange of information)
 - Z39.50 vs OAI-PMH
- Representation of knowledge
 - FRBR/LRM vs RDF/RDF Schema
- Information Retrieval and Web search engines
 - Free text search vs Google

- Bibliographic records and metadata 
- Interoperability - Exchange of information
- Conceptual models - Representation of knowledge
- Information Retrieval - Web search engines

Centuries and centuries of history as mediators between information and users

- Selection
 - Definition of collections
- Acquisition
 - Physical objects
- Description
 - Catalogs
- Access
 - Shelves
- Preservation
 - Controlled environment

- Classification means to bring related items together. Conventional libraries, in order to stack books on related subjects together, have used library classification. This facilitates the browsing approach of the information seekers
- Cataloguing creates “**document surrogates**”, i.e. a description of a document (a catalog record), to be used (to a certain extent) in the place of the document. Catalog records provide searching facility by Authors, Titles, Subjects, Series and other elements
- The catalogue was initially used as an “inventory” of the library to aid the librarian to maintain a “list” of the content of the library content, but very soon became a “tool” for the library users to find information in the library and to facilitate access to the library content

“What can be more easy (those lacking understanding say), having looked at the title pages than to write down the titles ? But these inexperienced people, who think making an index of their own few private books a pleasant task of a week or two, have no conception of the difficulties that arise or realize how carefully each book must be examined when the library numbers myriad of volumes.

In the colossal labor, which exhausts both body and soul, of making into an alphabetical catalog a multitude of books gathered from every corner of the earth there are many intricate and difficult problems that torture the mind.”

(Thomas Hyde, Bodleian Library, Oxford, 1674)

- Over the centuries, the “art of librarianship” has evolved, identifying the major tasks of a library user, and the catalogue has become the main “tool” to facilitate them
 - **find** entities that correspond to the user’s stated search criteria (i.e., to locate either a single entity or a set of entities in a file or database as the result of a search using an attribute or relationship of the entity)
 - **identify** an entity (i.e., to confirm that the entity described corresponds to the entity sought, or to distinguish between two or more entities with similar characteristics)
 - **select** an entity that is appropriate to the user’s needs (i.e., to choose an entity that meets the user’s requirements with respect to content, physical format, etc., or to reject an entity as being inappropriate to the user’s needs)
 - acquire or **obtain** access to the entity described (i.e., to acquire an entity through purchase, loan, etc., or to access an entity electronically through an online connection to a remote computer)

| 971. | Liberat. class. Roman. Poesis | 18 ^{to} . |
|-------------------------|---|--|
| Horatius A. Flaccus. | Quinti Horatii Flacci Opera, interpre- tatione et notis illustravit Ludovicus Desprez, jussu Christianissimi Regis, in usum Serenissimi Delphini, ac Serenissi- morum Principum Burgundiae, Andium, Biturigum. | Passani, 1777. Sed prostant Venetiis apud Remondini |
| II 127 10714 | jussu omnium | I cont Lib I M. c. c. b |

II
66054

L E P R O N I, F(ederico) :

Sulla tubercolosi miliare del polmone. Osservazioni Cliniche.

Foligno. 1934. 4o.

Estr. da: Rinnovamento Medico, Il Giornale di Tisiologia, 1934

(Lavori... 225.)

adligat n 66031

Alphabetically
ordered by
access point

- Author
- Title
- Subject



On-line Public Access Catalog (beginning of the seventies)

- **Images** of the catalog cards and/or **text** contained in the catalog cards
- Many advantages over traditional access via physical catalog cards
 - More than one access point
 - Author=Salton, Gerald AND Title=Modern Library Services
 - Author=Salton AND Title=Library
 - Any=Library
- WorldCat (maintained by OCLC) is presently the biggest OPAC
 - started in 1971
 - about 10 thousand libraries from more than 90 countries
 - more than 90 million records
 - 1200 million physical and digital assets
 - 360 languages

MAchine Readable Cataloging

- Started in the late sixties and developed by the Library of Congress to facilitate catalog sharing
- Provides a machine readable representation of a catalog card
- Based on a system of numbers, letters and symbols to identify fields in the record
- Provides a precise (sharable) description of the object
- Many “national” versions (UKMARC, CANMARC, AUSMARC, DANMARC, ANNAMARC, INTERMARC, etc)
- UNIMARC (Universal MARC) as standard format for exchange of information
 - e.g. USMARC to UNIMARC to AUSMARC

Actual MARC record

```

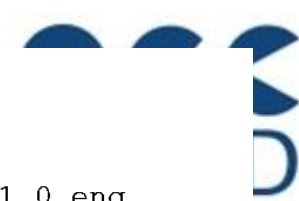
01041cam 2200265 a 4500001002000000000300040002000
50017000240080041000410100024000820200025001060200
04400131040001800175050002400193082001800217100003
20023524500870026724600360035425000120039026000370
04023000029004395000042004685200220005106500033007
30650001200763^###89048230#/AC/r91^DLC^19911106082
810.9^891101s1990####maua###j#####000#0#eng##^##$
a###89048230#/AC/r91^##$a0316107514 :$c$12.95^##$a
0316107506 (pbk.) :$c$5.95 ($6.95 Can.)^##$aDLC$cD
LC$dDLC^00$aGV943.25$b.B74 1990^00$a796.334/2$220^
10$aBrenner, Richard J.,$d1941-^10$aMake the team.
$pSoccer :$ba heads up guide to super soccer! /$cR
ichard J. Brenner.^30$aHeads up guide to super soc
cer.^##$alst ed.^##$aBoston :$bLittle, Brown,$cc19
90.^##$a127 p. :$bill. ;$c19 cm.^##$a"A Sports ill
ustrated for kids book."^##$aInstructions for impr
oving soccer skills. Discusses dribbling, heading,
playmaking, defense, conditioning, mental attitud
e, how to handle problems with coaches, parents, a
nd other players, and the history of soccer.^#0$aS
occer$vJuvenile literature.^#1$aSoccer.^
  
```

Leader record
(24 digits)

ID of the tag field
(3 digits)

Length of the field
(4 digits)

Position of the starting
character of the field
(5 digits)



HURC

HUman Readable Catalog record

| | | | | |
|----------------|----------|--|--------|-----------|
| Leader | 01041cam | 2200265 | a | 4500 |
| Control No. | 001 | ###89048230 | | |
| Control No. ID | 003 | DLC | | |
| DTLT | 005 | 19911106082810.9 | | |
| Fixed Data | 008 | 891101s1990 | maua j | 001 0 eng |
| LCCN | 010 | ## \$a ###89048230 | | |
| ISBN | 020 | ## \$a 0316107514 : | | |
| | | \$c \$12.95 | | |
| ISBN | 020 | ## \$a 0316107506 (pbk.) : | | |
| | | \$c \$5.95 (\$6.95 Can.) | | |
| Cat. Source | 040 | ## \$a DLC | | |
| | | \$c DLC | | |
| | | \$d DLC | | |
| LC Call No. | 050 | 00 \$a GV943.25 | | |
| | | \$b .B74 1990 | | |
| Dewey No. | 082 | 00 \$a 796.334/2 | | |
| | | \$2 20 | | |
| ME:Pers Name | 100 | 1# \$a Brenner, Richard J., | | |
| | | \$d 1941- | | |
| Title | 245 | 10 \$a Make the team. | | |
| | | \$p Soccer : | | |
| | | \$b a heads up guide to super soccer! / | | |
| | | \$c Richard J. Brenner. | | |
| Variant Title | 246 | 30 \$a Heads up guide to super soccer | | |
| Edition | 250 | ## \$a 1st ed. | | |
| Publication | 260 | ## \$a Boston : | | |
| | | \$b Little, Brown, | | |
| | | \$c c1990. | | |
| Phys Desc | 300 | ## \$a 127 p. : | | |
| | | \$b ill. ; | | |
| | | \$c 19 cm. | | |
| Note: General | 500 | ## \$a "A Sports illustrated for kids | | |
| | | book." | | |
| Note: Summary | 520 | ## \$a Instructions for improving soccer | | |
| | | skills. Discusses dribbling, heading, | | |
| | | playmaking, defense, conditioning, | | |
| | | mental attitude, how to handle | | |
| | | problems with coaches, parents, | | |
| | | and other players, and the history | | |
| | | of soccer. | | |
| Subj: Topical | 650 | #0 \$a Soccer | | |

- Increase in the amount of information available on-line (data bases, repositories, the Web, etc)
- Increase in the variety of information available on-line (text, sound, images, video, 3D, etc)
- Self-publishing of papers (sent to journals) into Institutional Repositories
- Description of information not always done by “specialists”
- Need to describe **resources** available online through **metadata**

- An Institutional repository is a centrally managed collection of institutionally generated digital objects designed to be maintained “for ever”
- Established and maintained by universities and research institutions (initially) for “self-publishing”
- An *e-print* is an author self-archived document. The content of an e-print is usually the result of scientific or other scholarly research.
- Repositories contain scholarly publications
 - Reports
 - Working papers
 - Pre- and post-prints of articles and books
 - Doctoral thesis
 - Data supporting research
 - References and professional databases related to research topics

- Machine-understandable information about Web resources or other things (Tim Berners-Lee 1997)
- Data associated with objects which relieves their potential users of having to have full advance knowledge of their existence or characteristics; a user might be a program or a person (Lorcan Dempsey 1998)
- Structured data about resources that can be used to help support a wide range of operations (Michael Day, 2001)
- Structured data about data (DCMI 2003)
- Structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource (NISO 2004)

Metadata can be associated with any “entity”: physical, digital, abstract, fantasy, etc.

Not simply a cataloguing record

- An important reason for creating descriptive metadata is to facilitate discovery of relevant information, as it serves the same functions in resource discovery as good cataloging does by:
 - allowing resources to be found by relevant criteria
 - identifying resources
 - bringing similar resources together
 - distinguishing dissimilar resources
 - giving location information
- In addition to resource discovery, metadata can
 - help organize electronic resources
 - facilitate interoperability and legacy resource integration
 - provide digital identification
 - support archiving and preservation

Another point of view

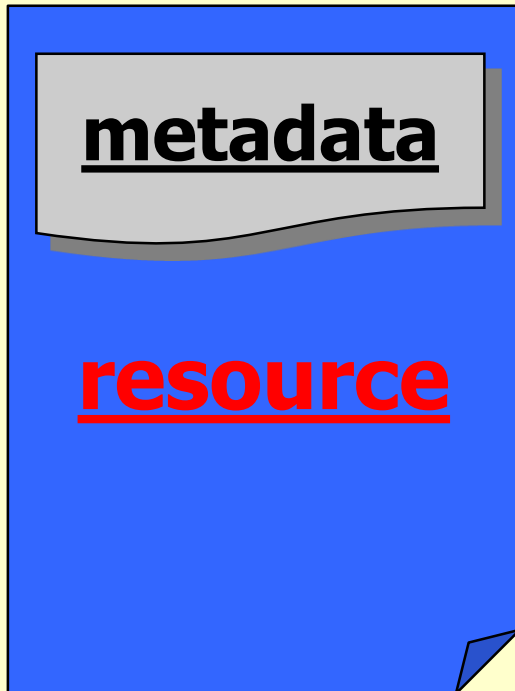
“Then there is the question of cataloguing and metadata. My view of the latter is that it is an ill-considered attempt to find some kind of Third Way between the wilderness of search engines and free text searching and the grand architecture of bibliographic control that librarians have developed over the last 150 years. I think that metadata is the product of those with no knowledge of, or regard for, cataloguing; they are bibliographic alchemists seeking the philosopher’s stone that will offer us effective cataloguing without expense and effective access without controlled vocabularies.

There is no such thing and the sooner that notion is disposed of, the better”

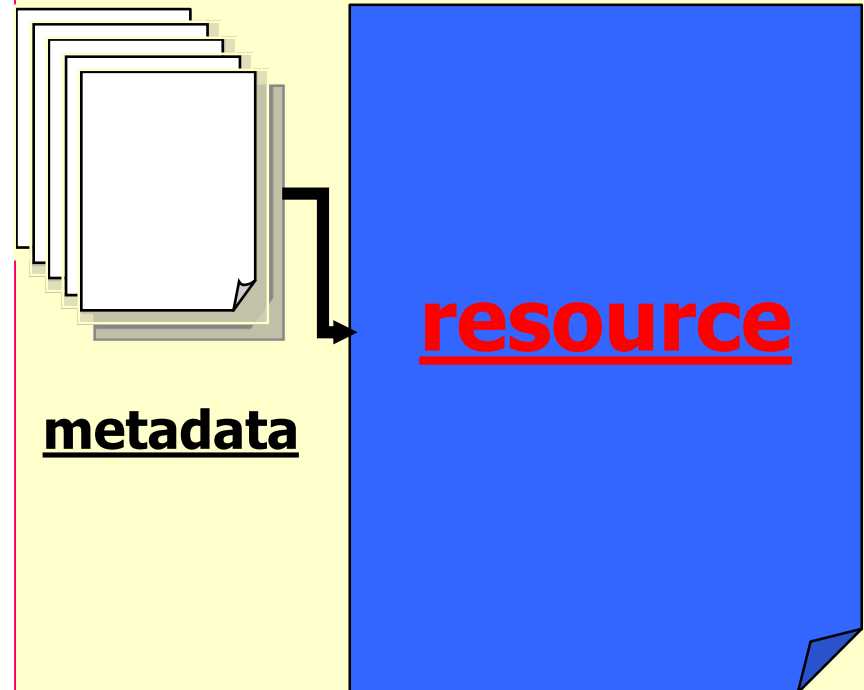
(Michael Gorman, Dean of Library Services at California State University, past President of the American Library Association, November 2000)

Storing Metadata

embedded metadata



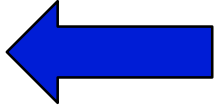
stand-alone metadata



- Dublin Core
 - Dublin: Dublin, Ohio, 1995
 - Core: minimal set of broad and generic elements
- Dublin Core was originally developed with an eye to describing document-like objects
 - Descriptions easy to create (unlike MARC)
- Despite initial focus, has proved to be general enough to describe “any” type of objects
 - unlike catalog records, often tied to specific application fields
- It is now a widely used international standard
 - ISO Standard 15836-2003
 - NISO Standard Z39.85-2007
 - IETF RFC 5013

Definition of elements (or terms) to describe resources

| Content | Intellectual Property | Instantiation |
|----------------|------------------------------|----------------------|
| Title | Creator | Date |
| Subject | Contributor | Format |
| Description | Publisher | Identifier |
| Type | Rights | Language |
| Source | | |
| Relation | | |
| Coverage | | |

- Bibliographic records and metadata
- Interoperability - Exchange of information 
- Conceptual models - Representation of knowledge
- Information Retrieval - Web search engines

Interoperability in digital libraries

- Need for interoperability between libraries
 - Union catalogs (OPACs)
 - Interlibrary loan
- Interoperability between different organizations
 - Eg. using different library formats
- Interoperability between groups of users
 - Eg. Public libraries/Academic libraries
 - Eg. libraries in different countries
- Interoperability between communities
 - Eg. libraries, publishers, archives, museums

The Z39.50 protocol (the 80')

- "Information Retrieval (Z39.50); Application Service Definition and Protocol Specification, ANSI/NISO Z39.50-1995"
- Current version (Version 3) was adopted in 1995, superceding earlier versions adopted in 1992 and 1988 (1984 version was rejected)
 - Another revision, initiated in 2001, is still “work in progress”
- Z39.50 was heavily influenced by OSI, and was an “application layer” protocol; in Version 3 it runs over TCP/IP
- It is a wide ranging protocol for information retrieval between a client and a database server, which attempts to standardize shared semantic knowledge
- A server houses one or more databases containing records; associated with each database are a set of **access points** (indexes) that can be used for searching
- A search (sent from the client/origin to the server/target) produces a set of records, called a "result set", that are **maintained on the server**



Z39.50 : THE BASICS

Fay Turner

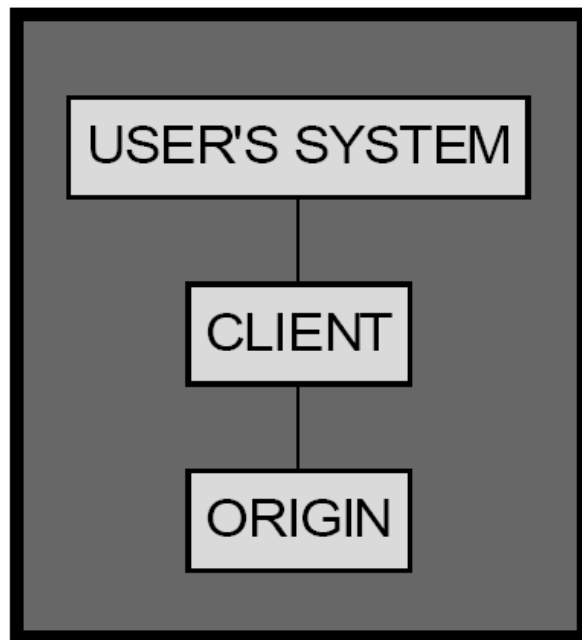
National Library of Canada

fay.turner@nlc-bnc.ca

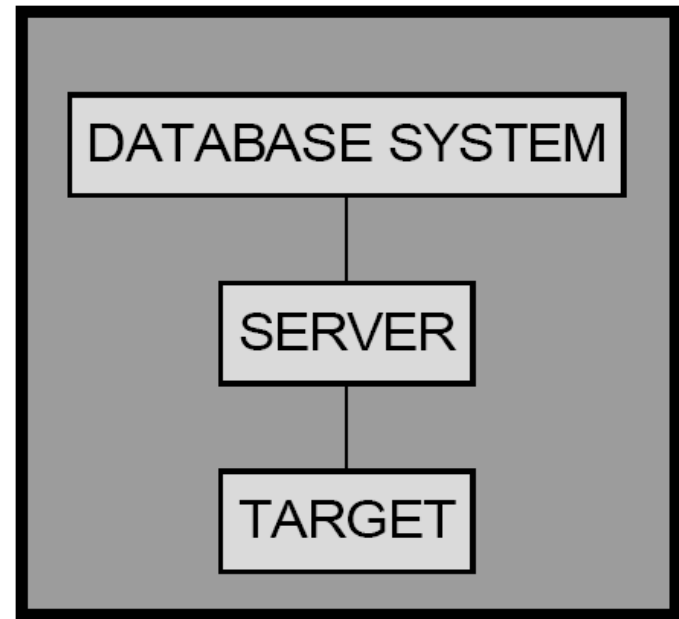
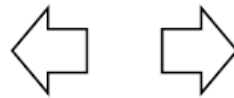
1997 IFLA Conference

Session - Z39.50: Information Retrieval in an
Open Networked Environment

Z39.50 MODEL



DRA, NOTIS, VTLS



GEAC, OCLC, LC

ORIGIN SYSTEM (CLIENT)

- λ SOFTWARE ON LOCAL SYSTEM
TRANSLATES SEARCH QUERY INTO
FORMAT OF Z39.50 STANDARD
- λ CONNECTS TO AND SENDS QUERY TO
SYSTEM HOUSING THE DATABASE
- λ PRESENTS RECORDS/RESULTS OF
QUERY TO SEARCHER

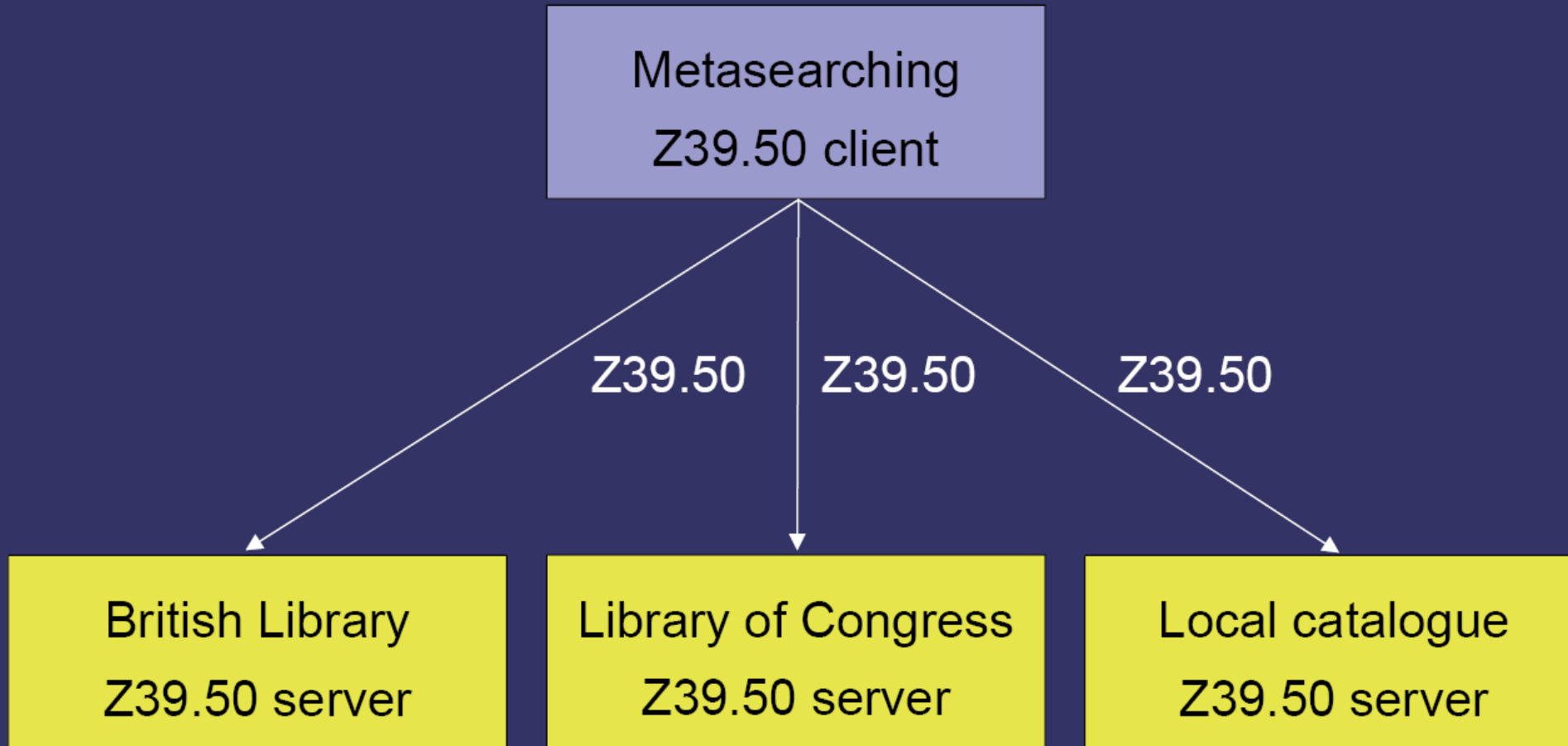
**SEARCHER OF ORIGIN SYSTEM NEVER
INTERACTS DIRECTLY WITH TARGET SYSTEM**

TARGET SYSTEM (SERVER)

- λ COMPUTER HOUSING THE DATABASE(S)
- λ TRANSLATES THE Z39.50 QUERY TO SEARCH LOGIC OF DATABASE SYSTEM
- λ OBTAINS INFO FROM DATABASE, RETURNS IT TO ORIGIN SYSTEM
- λ RETURNS RECORDS OR REPORTS A RESULT SET

**CLIENT AND TARGET ROLES CAN BE
CONTAINED IN SAME SYSTEM**

Z39.50 for searching multiple catalogues



OAI – Open Archives Initiative (the 90’)

- The roots of OAI lie in the development of eprint archives (i.e. Institutional Repositories) such as arXiv, CogPrints, NACA (NASA), RePEc, NDLTD, NCSTRL, etc.
- Each repository offered a web interface for deposit of articles and for end-user searches
- It was difficult for end-users to work across archives without having to learn multiple different interfaces
- Initial experiments for single search interface to all archives
- Universal Pre-print Service (UPS) renamed OAI at the Santa Fe Convention (1999)

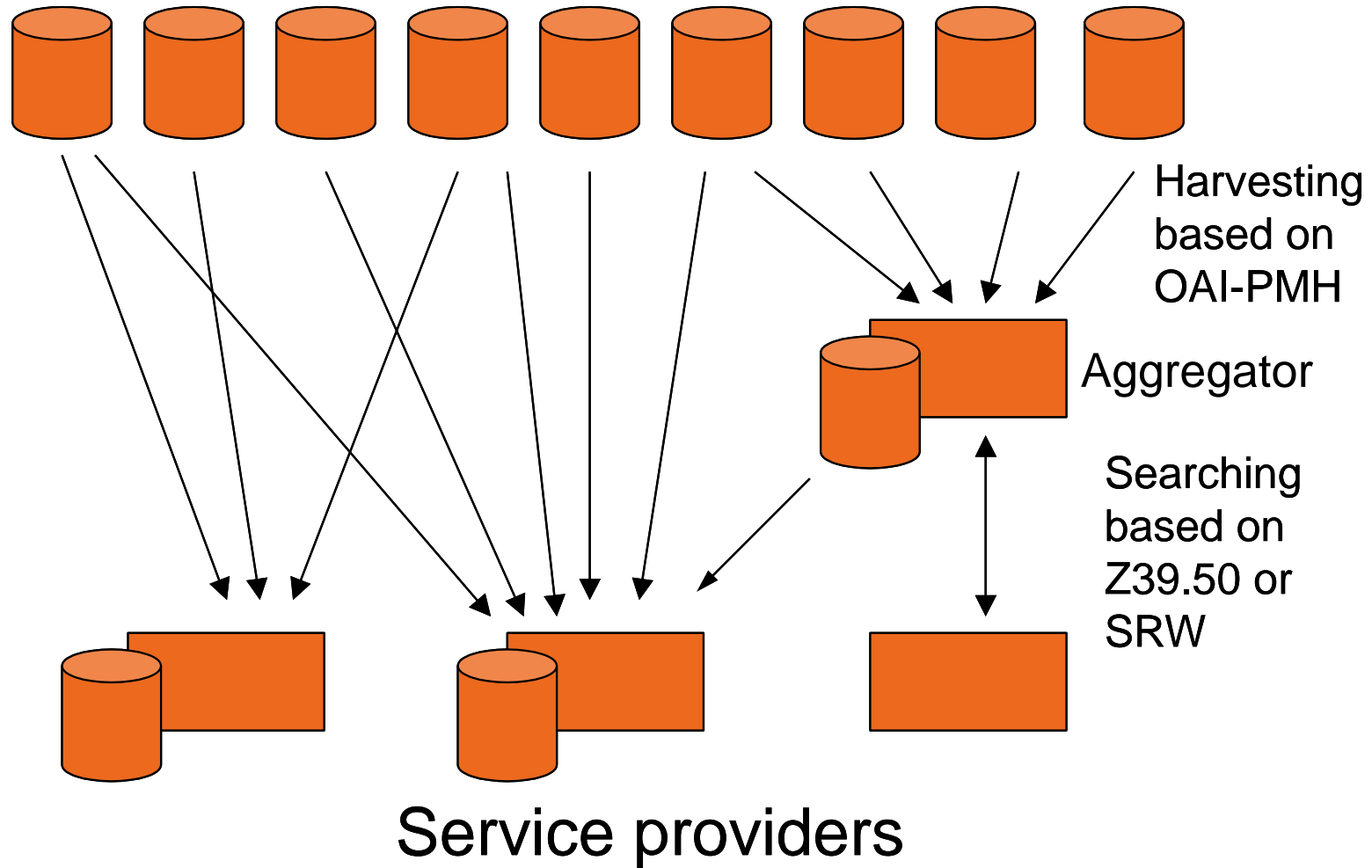
- It is interesting to see another protocol for “resource discovery”, namely the Open Archive Initiative (OAI) protocol
- Historical separation from Z39.50
 - OAI appears about 15 years after Z39.50
- Cultural separation from Z39.50
 - Z39.50 originated in the traditional library community
 - OAI originated in the “Web Community”
- Conceptual separation from Z39.50
 - Z39.50 based on solid (but heavy and bulky) foundations
 - OAI based on simple and pragmatic ideas

- Two possible approaches for single search interface to all archives
 - cross searching multiple archives based on protocol like Z39.50 (possibly lighter)
 - harvesting metadata into one or more ‘central’ services
- Problems with cross searching
 - Not scalable (overall performance determined by slowest server)
 - Problems of deciding which servers to target (collection descriptions not consistent)
 - Differences in interfaces and query languages
 - Problems in the ranked merging of results (different types and size of targets can skew results)
 - Browse interface very difficult to build
- Decision was to go with harvesting

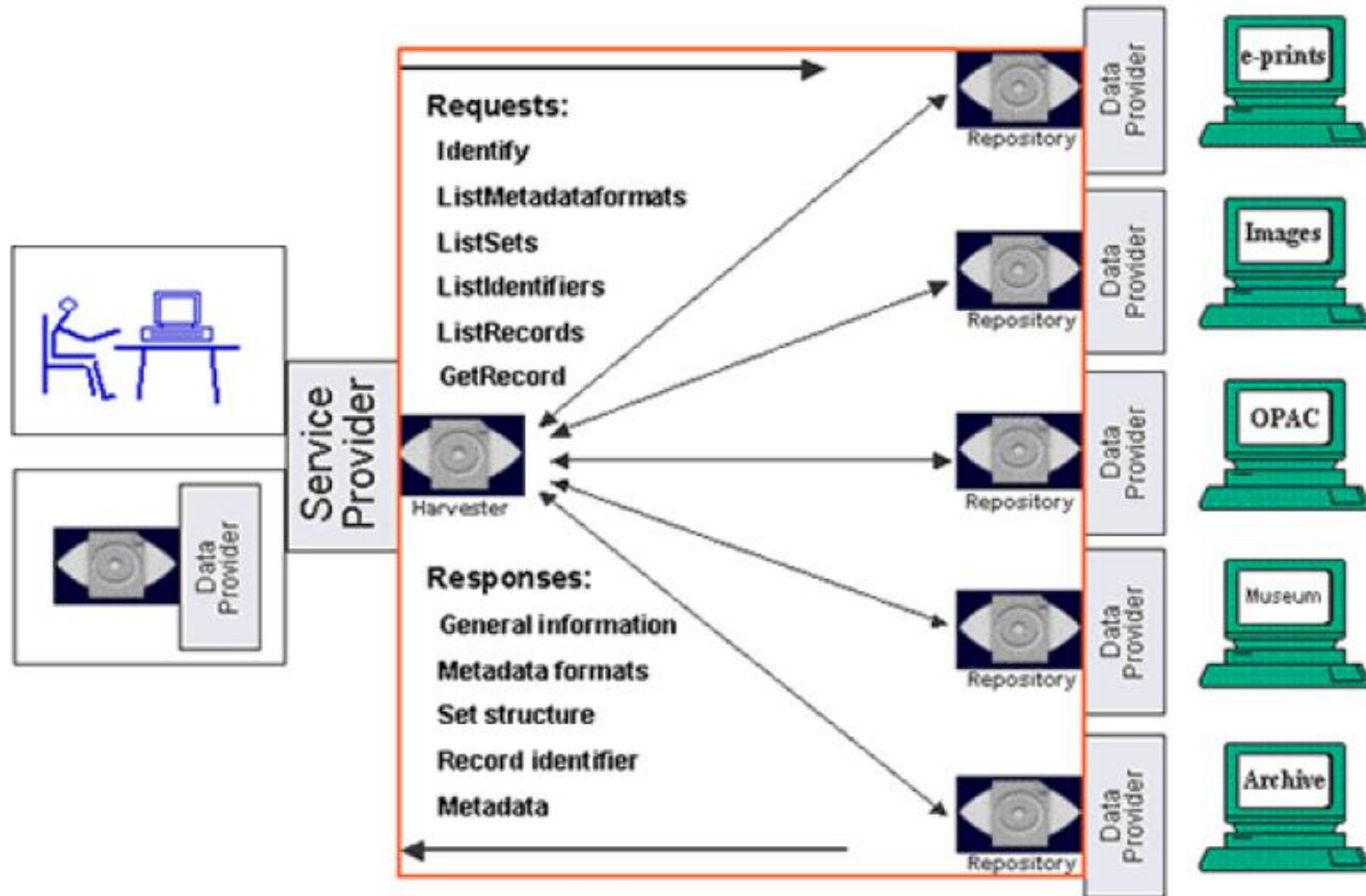
OAI Protocol for Metadata Harvesting

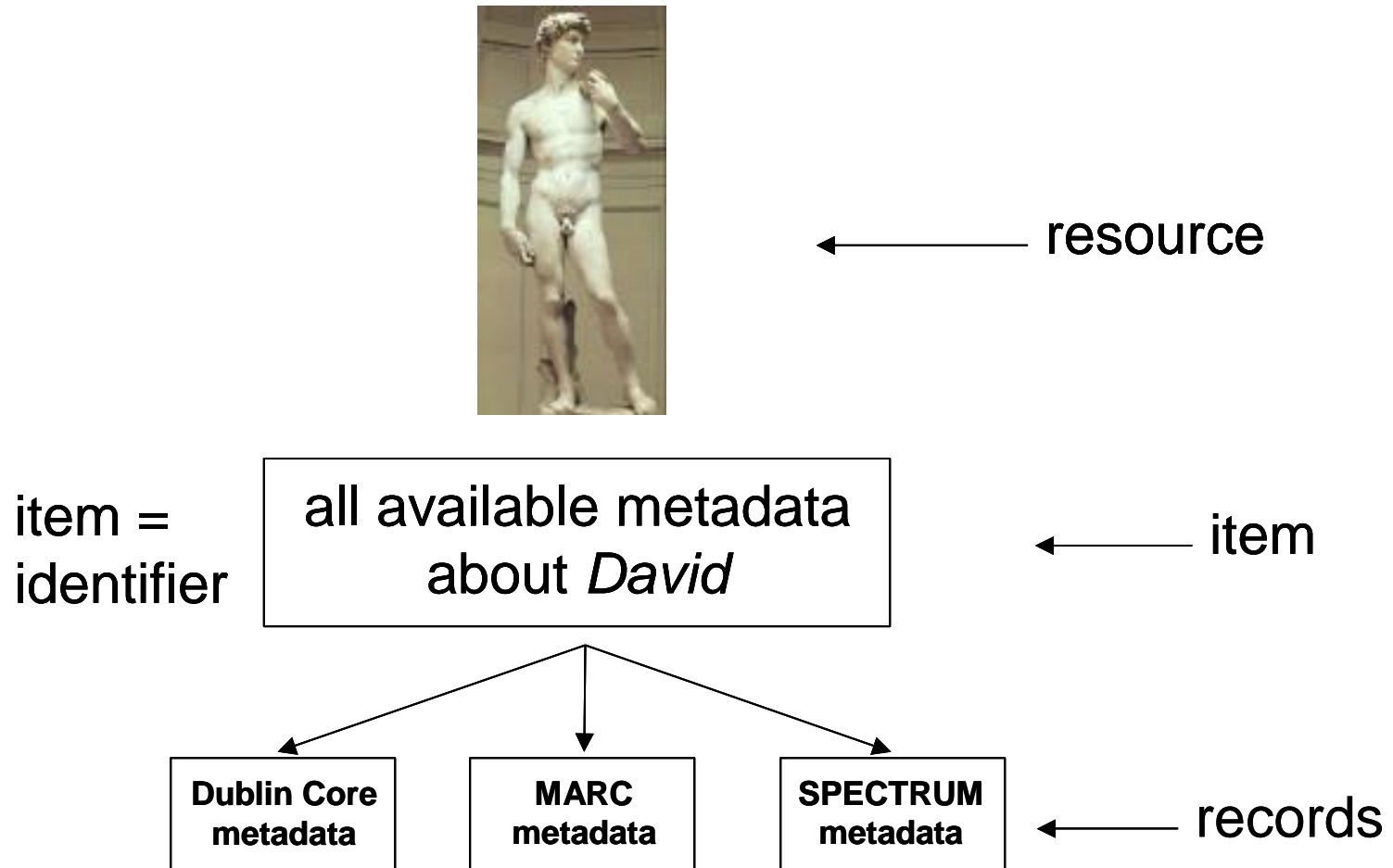
- Data providers make metadata available for harvesting
- Service Providers harvest metadata
- Metadata can be centrally collected or “aggregated”
- Data Providers
 - Are creators and keepers of the metadata for objects (repositories) and (possibly but not necessarily) archives of resources
 - Handle deposit and publishing
- Service Providers
 - Are harvesters of metadata for the purpose of providing a service such as a search interface, peer-review system, etc.

OAI – PMH overview

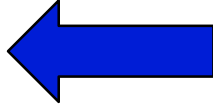


Overview of OAI - PMH





- Identify
 - description of an archive
- ListMetadataFormats
 - retrieve available metadata formats from archive
- ListSets
 - retrieve set structure of a repository
- ListIdentifiers
 - abbreviated form of ListRecords, retrieving only headers
- ListRecords
 - harvest records from a repository
- GetRecord
 - retrieve individual metadata record from a repository

- Bibliographic records and metadata
- Interoperability - Exchange of information
- Conceptual models
Representation of knowledge 
- Information Retrieval - Web search engines

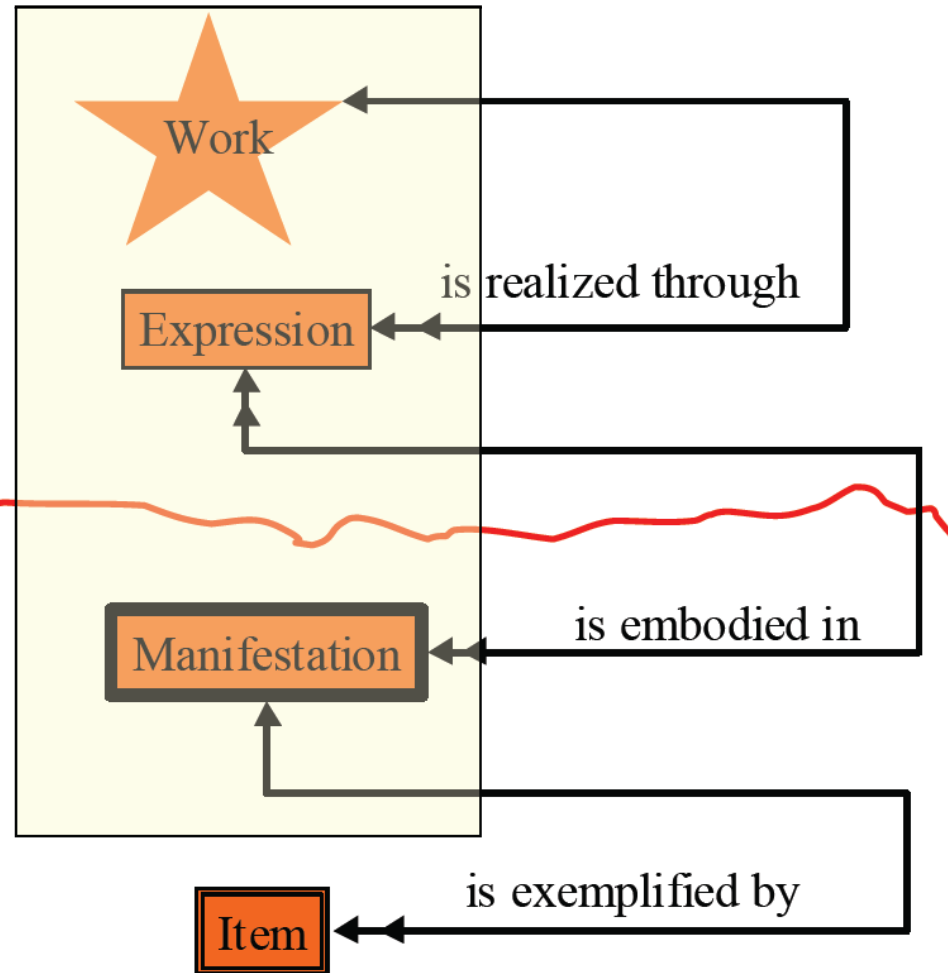
| | |
|---------|--|
| Group 1 | Work, Expression, Manifestation, Item |
| Group 2 | Person, Corporate body |
| Group 3 | Concept, Object, Event, Place |

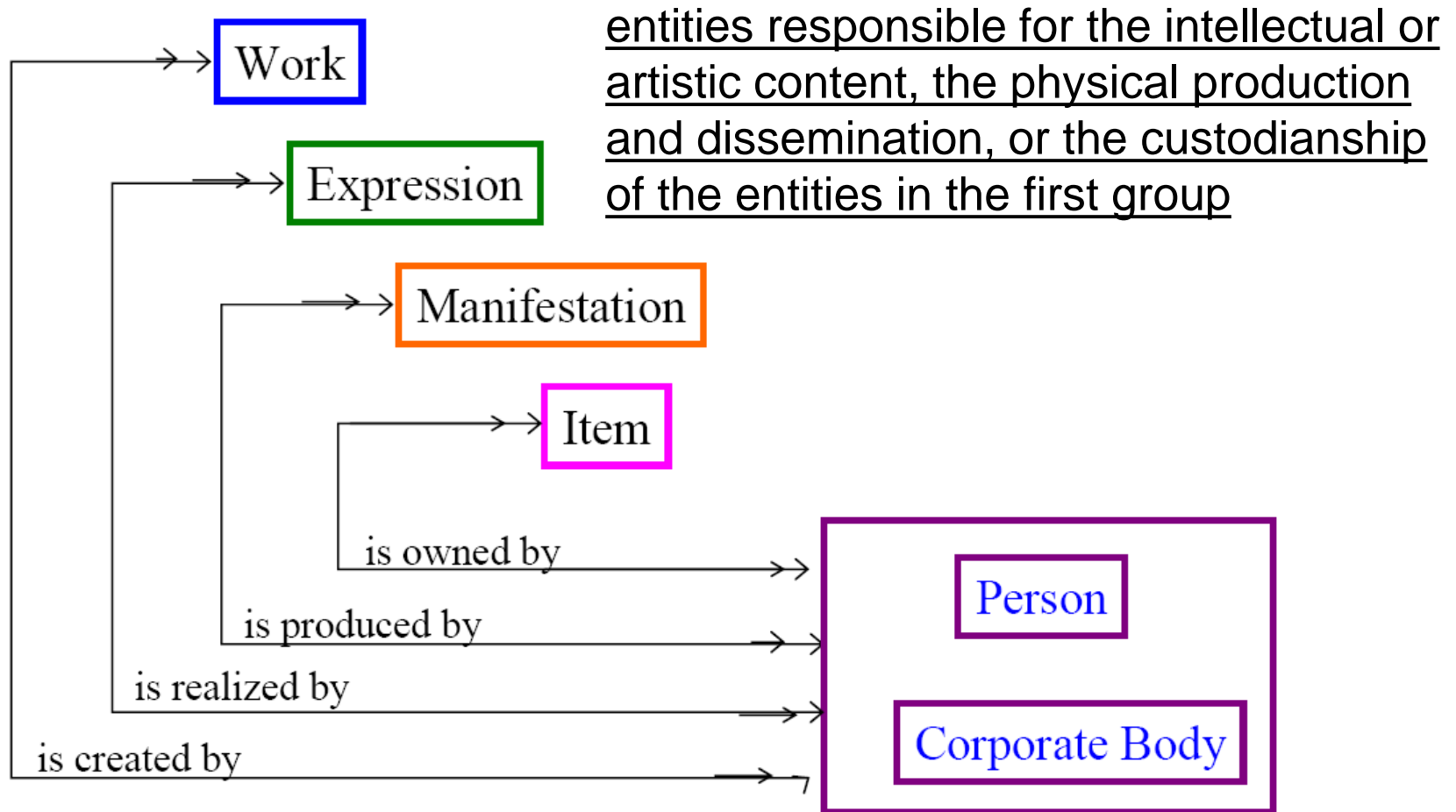


products of
intellectual
or artistic
endeavour

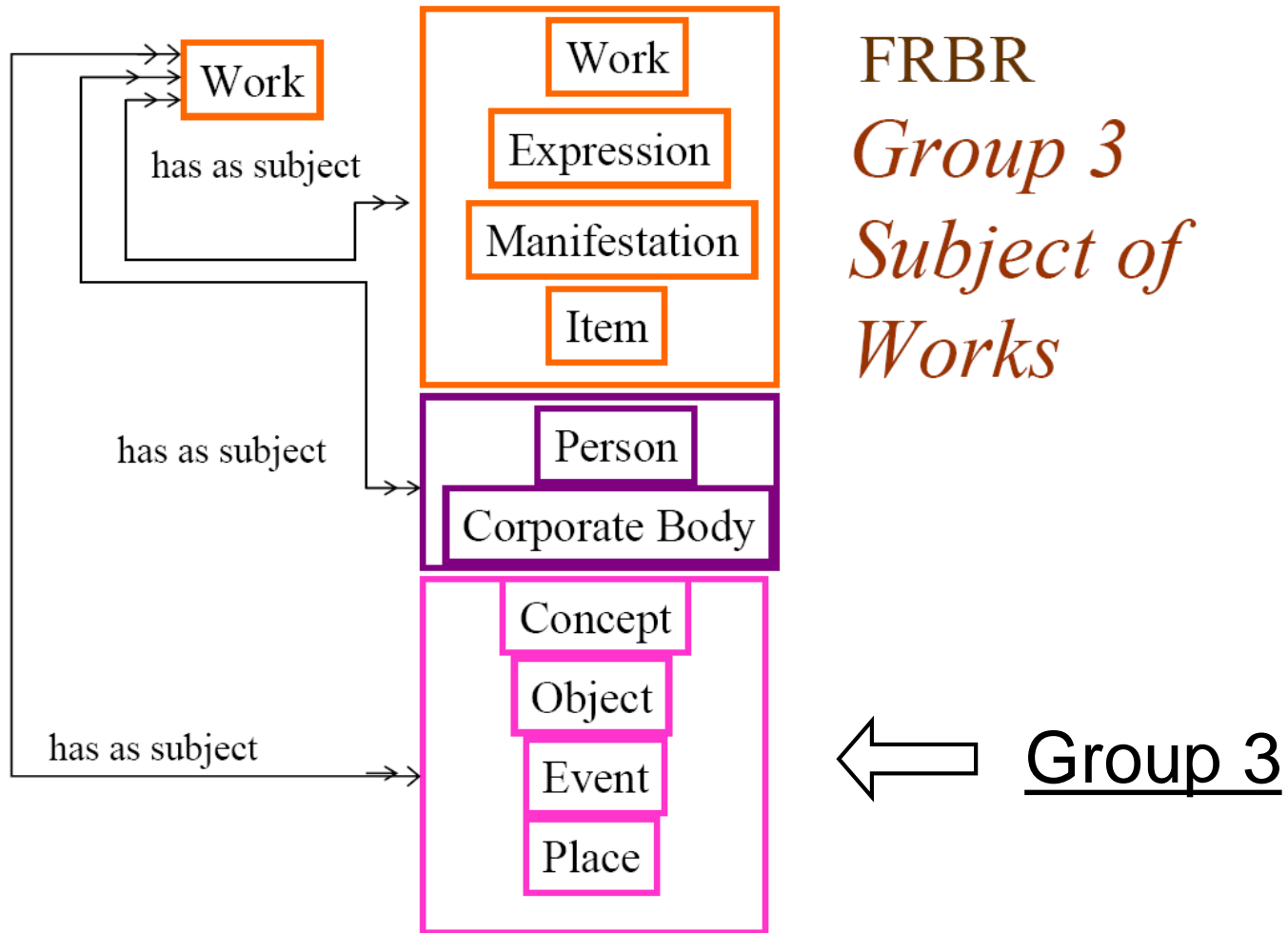
Intellectual/
artistic content

Physical -
recording of
content

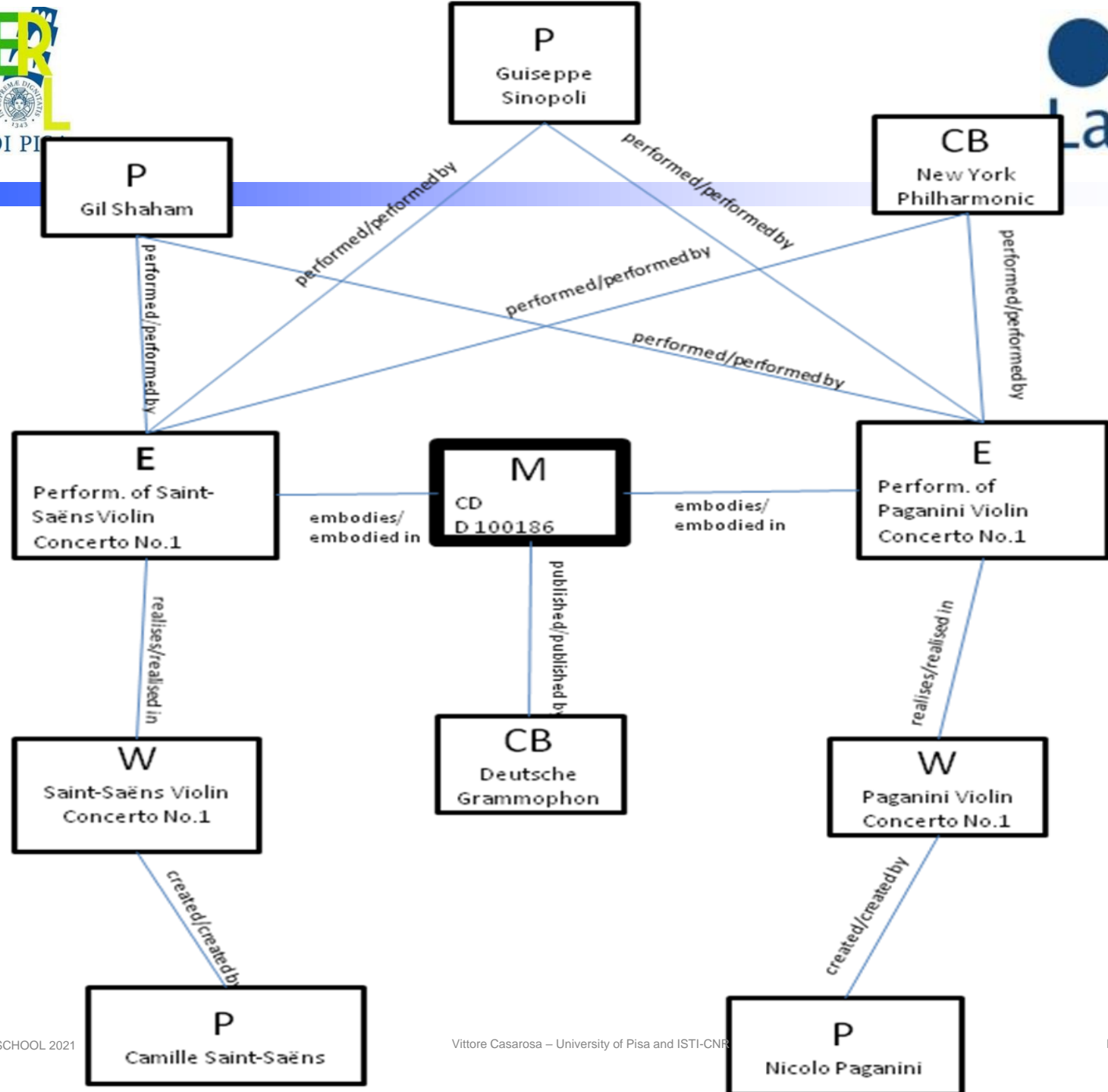




FRBR – Group 3 entities



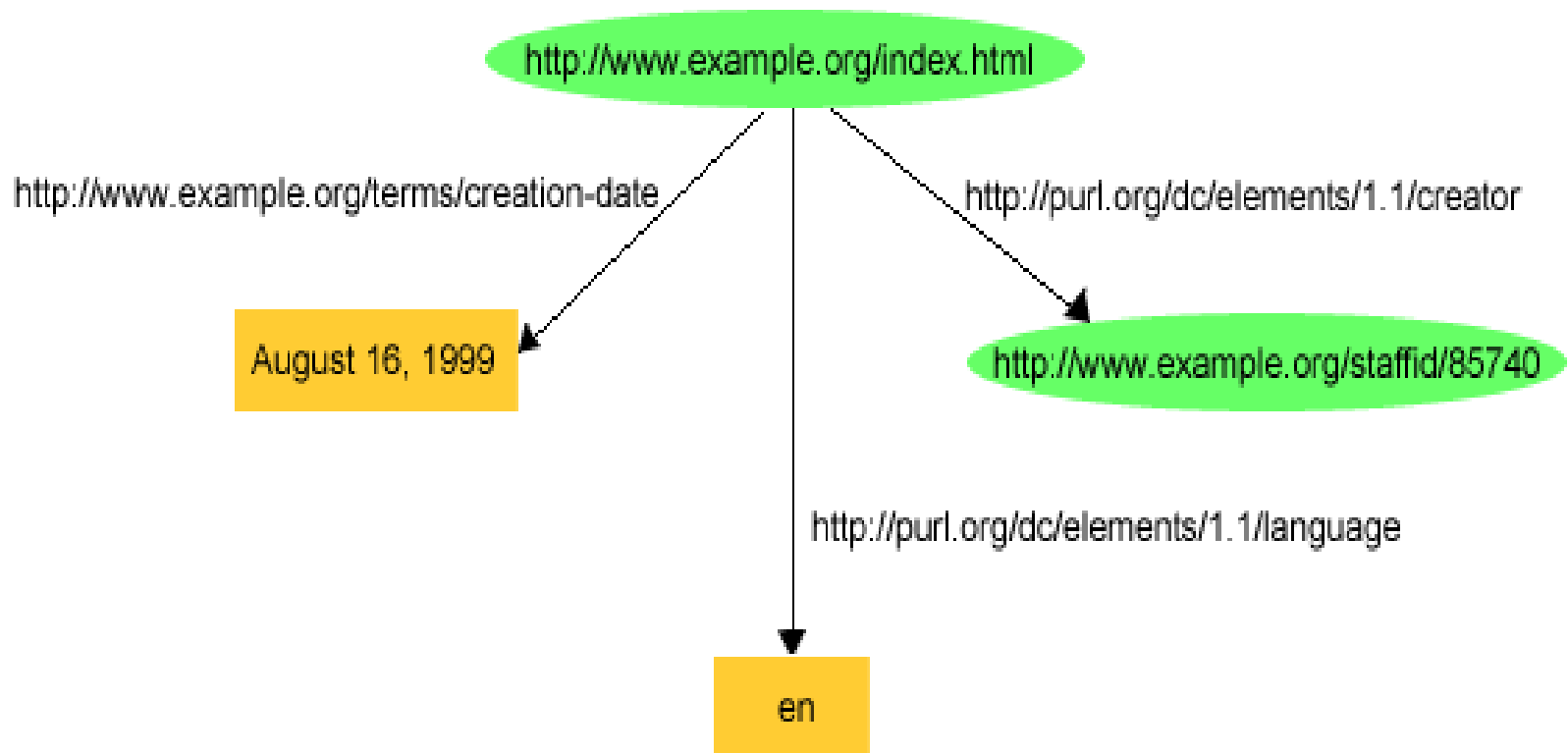


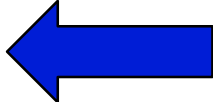


- Resource Description Framework (RDF) is a language for representing information about *resources* in the WWW
- All *resources* are identified by a URI. A resource is anything that has identity. For example, a resource may be an electronic document, an image, a service (e.g., "today's weather report 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
- *Resources* are described in terms of simple statements specifying properties and property values
- A statements is: (subject and predicate are indicated by a URI)
 - A subject
 - A predicate (about the subject)
 - An object (the value of the predicate, a URI or a terminal string)

- <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 is represented as a graph



- Bibliographic records and metadata
 - Interoperability - Exchange of information
 - Conceptual models - Representation of knowledge
 - Information Retrieval
Web search engines
- 

- Information Retrieval (IR) is finding material (usually documents) of an **unstructured** nature (usually text) that satisfies an information need, from within large collections (usually stored on computers).
- Research in Information Retrieval started in the seventies, as a field complementary to data base querying (retrieval of **structured** data)
- Today, search engines have made full text retrieval the normal way to query for information
- The model of IR is:
 - There is a collection of **digital documents**
 - The user enters a query (usually a few words)
 - As quickly as possible, the system returns a list of documents **ranked** in order of **relevance** to the query

In order to do that quickly, the IR system has to do two things, before receiving any query

- it is necessary to build an **index**, i.e. a list of all the different words contained in all the documents of the collection
 - each word (term) is associated with the list of documents in which it appears
- it is necessary to represent the documents in a way suitable for an algorithm to compute the relevance between the query and a document
 - each document must have a **“mathematical” representation**

Sample “collection”

Document

Text

-
- | | |
|---|--|
| 1 | Pease porridge hot, pease porridge cold, |
| 2 | Pease porridge in the pot, |
| 3 | Nine days old. |
| 4 | Some like it hot, some like it cold, |
| 5 | Some like it in the pot, |
| 6 | Nine days old. |
-

The index

| Number | Term | Documents | (Document; Words) |
|--------|----------|---------------------------|--|
| 1 | cold | $\langle 2; 1, 4 \rangle$ | $\langle 2; (1; 6), (4; 8) \rangle$ |
| 2 | days | $\langle 2; 3, 6 \rangle$ | $\langle 2; (3; 2), (6; 2) \rangle$ |
| 3 | hot | $\langle 2; 1, 4 \rangle$ | $\langle 2; (1; 3), (4; 4) \rangle$ |
| 4 | in | $\langle 2; 2, 5 \rangle$ | $\langle 2; (2; 3), (5; 4) \rangle$ |
| 5 | it | $\langle 2; 4, 5 \rangle$ | $\langle 2; (4; 3, 7), (5; 3) \rangle$ |
| 6 | like | $\langle 2; 4, 5 \rangle$ | $\langle 2; (4; 2, 6), (5; 2) \rangle$ |
| 7 | nine | $\langle 2; 3, 6 \rangle$ | $\langle 2; (3; 1), (6; 1) \rangle$ |
| 8 | old | $\langle 2; 3, 6 \rangle$ | $\langle 2; (3; 3), (6; 3) \rangle$ |
| 9 | pease | $\langle 2; 1, 2 \rangle$ | $\langle 2; (1; 1, 4), (2; 1) \rangle$ |
| 10 | porridge | $\langle 2; 1, 2 \rangle$ | $\langle 2; (1; 2, 5), (2; 2) \rangle$ |
| 11 | pot | $\langle 2; 2, 5 \rangle$ | $\langle 2; (2; 5), (5; 6) \rangle$ |
| 12 | some | $\langle 2; 4, 5 \rangle$ | $\langle 2; (4; 1, 5), (5; 1) \rangle$ |
| 13 | the | $\langle 2; 2, 5 \rangle$ | $\langle 2; (2; 4), (5; 5) \rangle$ |

document frequency

lexicon or vocabulary

inverted list (postings list)

- The basic idea is to represent a document with the list of all the distinct words it contains, in alphabetical order (**bag of words**)
- Given the existence of the lexicon, the most “natural” (and easy) way to represent “the bag of words” in a mathematical form is a **vector** (a list of numbers), with as many elements as there are terms in the lexicon
- The elements of the vector are:
 - **zero** for those words of the lexicon that are not in the document
 - **one** for those words of the lexicon that are contained in the document
- It is easy to represent also the query (at query time) as a binary vector
- The similarity between the query and a document (the relevance) can now be defined as the “inner product” of the two vectors (an easy mathematical computation)

Document as binary vectors

(a)

| <i>d</i> | Document vectors $\langle w_{d,t} \rangle$ | | | | | | | | | |
|----------|--|------------|------------|------------|------------|------------|------------|------------|------------|------------|
| lexicon | <i>col</i> | <i>day</i> | <i>eat</i> | <i>hot</i> | <i>lot</i> | <i>nin</i> | <i>old</i> | <i>pea</i> | <i>por</i> | <i>pot</i> |
| 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0 |
| 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 |
| 3 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 |
| 4 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 |
| 6 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |

(b)

| | | | | | | | | | | |
|---------------------|---|---|---|---|---|---|---|---|---|---|
| <i>eat</i> | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| <i>hot porridge</i> | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |

each document (and each query) is represented as a vector (a sequence) of 0's and 1's
the number of components of the vectors is equal to the size of the lexicon

- Three drawbacks of this approach to compute relevance
 - No account of term frequency in the document (i.e. how many times a term appears in the document)
 - No account of term scarcity (in how many documents the term appears)
 - Long documents with many terms are favoured
- To overcome these drawbacks, for each distinct term in a document we compute its **weight** in representing the document

Component of the weight

- What is easy to measure when building the index is the number of times that a term appear in a document; this number is called **term frequency**, usually indicated with *tf*
- The way to take into account the scarcity (or abundance) of a term in a collection is to count the number of documents in which a term appears (called the term **document frequency**) and then to consider its inverse, i.e. the inverse of the document frequency, usually indicated with *idf*)

Final weight: $tf \times idf$ (or $tf.idf$)

- In conclusion, the weight of each term i in each document d ($w_{i,d}$) is usually given by the following formula (or very similar variations), called the ***tf.idf*** weight

$$w_{i,d} = tf_{i,d} \times \log(n / df_i)$$

$tf_{i,d}$ = frequency of term i in document d

n = total number of documents

df_i = the number of documents that contain term i

- Increases with the number of occurrences of a term *within* a doc
- Increases with the rarity of the term *across* the whole corpus

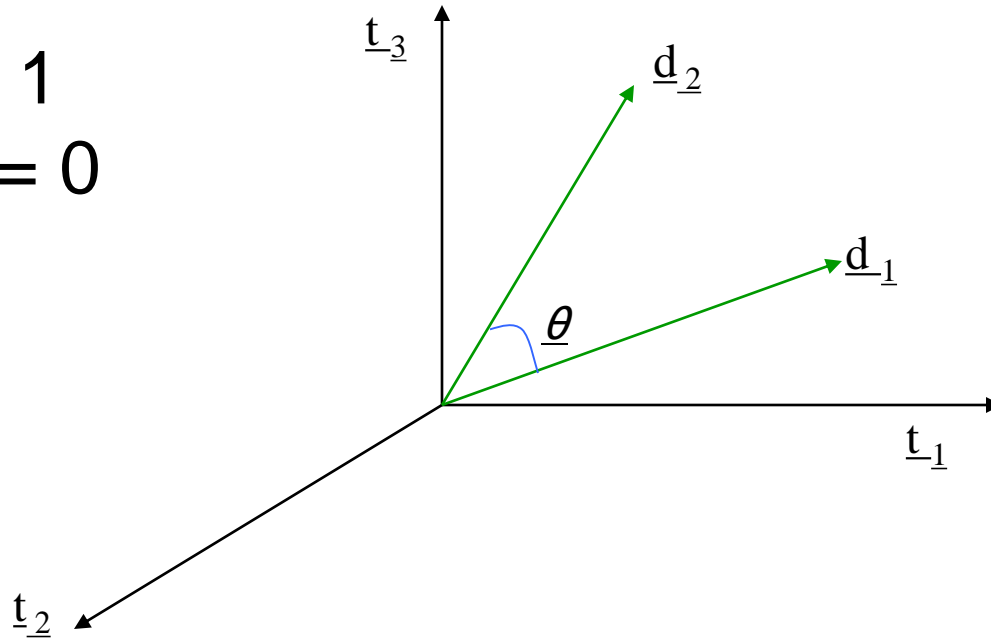
Vectors of weights

- We now have all documents represented as vectors of weights, which take care of both the **term frequency** (how many times a term appears in a document) and the **document frequency** (in how many documents a term appears)
- We can easily represent also the query as a vector of weights
- We need now a formula (an algorithm) to measure the similarity between the vector representing the query and the vector representing a document, whose value does not depend on the length of a document; we will consider this similarity as the measure of relevance

- Similarity between vectors d_1 and d_2 is *captured* by the **cosine** of the angle x between them.

$$\cos 0^\circ = 1$$

$$\cos 90^\circ = 0$$



- The cosine of the angle between two vectors can easily be computed with the formula

$$X \cdot Y = |X||Y| \cos \theta$$

$$\cos \theta = \frac{X \cdot Y}{|X||Y|}$$

- $|X|$ is the modulus (the length) of X

$$|X| = \sqrt{\sum_{i=1}^n x_i^2}$$

- The cosine is therefore

$$\cos \theta = \frac{X \cdot Y}{|X||Y|} = \frac{\sum_{i=1}^n x_i y_i}{\sqrt{\sum_{i=1}^n x_i^2} \sqrt{\sum_{i=1}^n y_i^2}}$$

- The cosine formula applied to documents

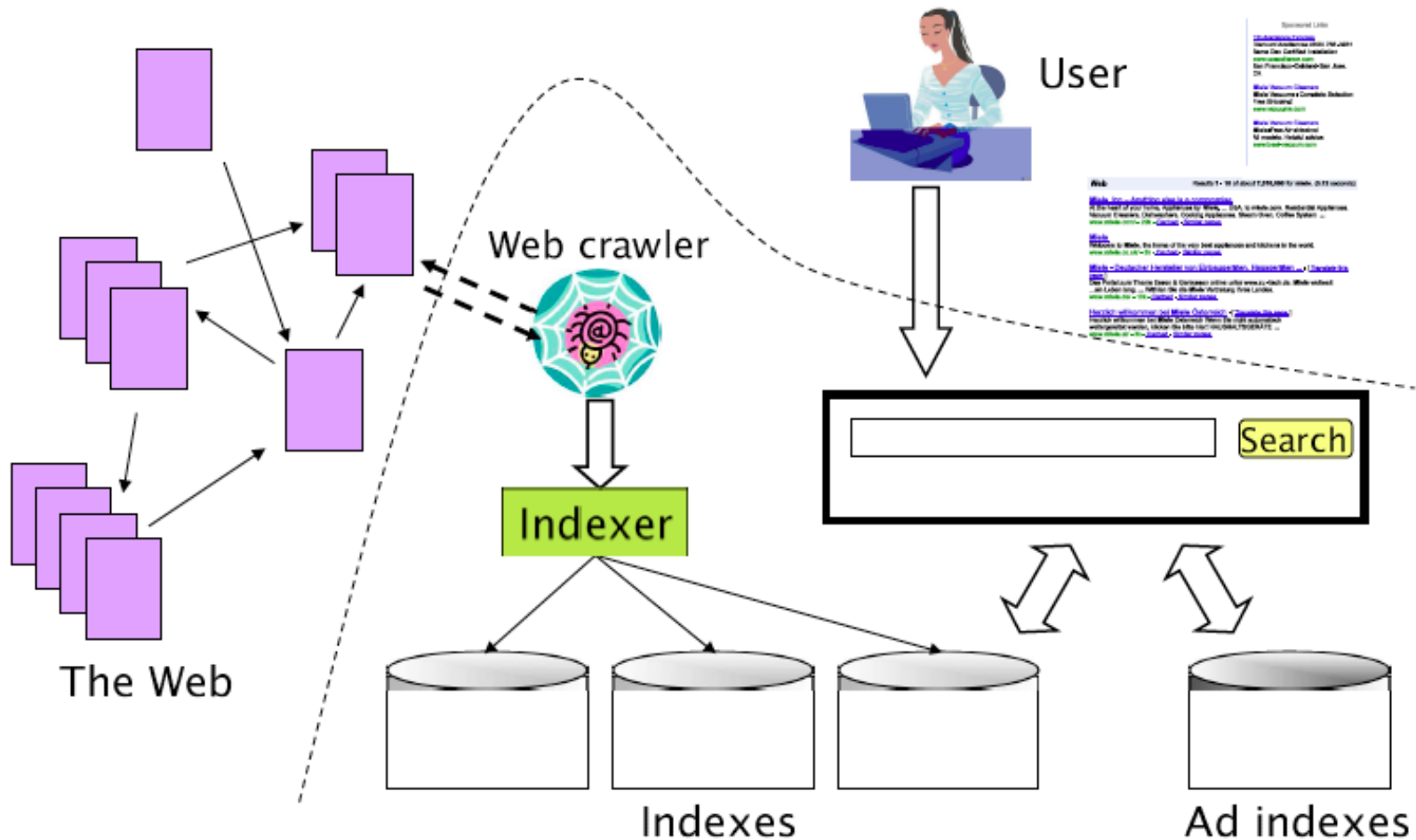
$$\text{cosine}(Q, D_d) = \frac{Q \cdot D_d}{|Q||D_d|} = \frac{1}{W_q W_d} \sum_{t=1}^n w_{q,t} \cdot w_{d,t}$$

- W_q and W_d represent the “length” of the vectors
- In the formula above the factor W_q can be ignored as it is just a multiplying factor that does not affect the ranking as it is the same for all documents

Summary of retrieval and ranking

- Build a “term-document matrix”, assigning a weight to each term in a document (instead of just a binary value as in the simple approach)
 - Usually the weight is *tf.idf*, i.e. the product of the “term frequency” (number of occurrences of the term in the document) and the “inverse of the “term document frequency” (number of documents in which the term appears)
- Consider each document as a vector in n-space (n is the number of distinct terms, i.e. the size of the lexicon)
 - The non-zero components of the vector are the weights of the terms appearing in the document
 - Normalize each vector to “unit length” (divide each component by the modulus – the “length” – of the vector)
- Consider also the query as a vector in n-space
 - The non-zero components are just the terms appearing in the query (possibly with a weight)
 - Normalize also the query vector
- Define the similarity measure between the query and a document as the cosine of the “angle” between the two vectors
 - If both vectors are normalized, the computation is just the inner product of the two vectors

Architecture of a Web Search Engine

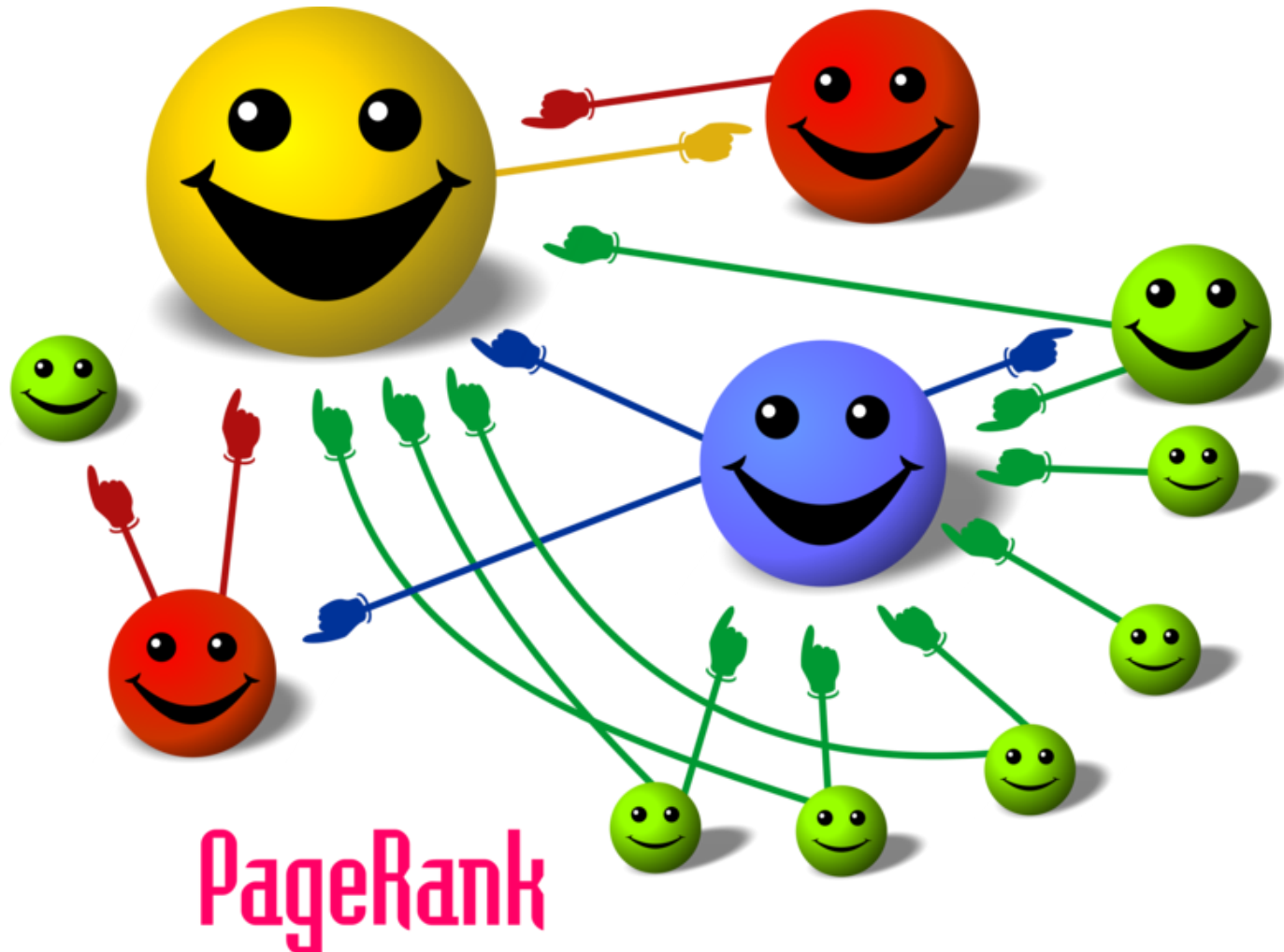


Main functions of a search engine

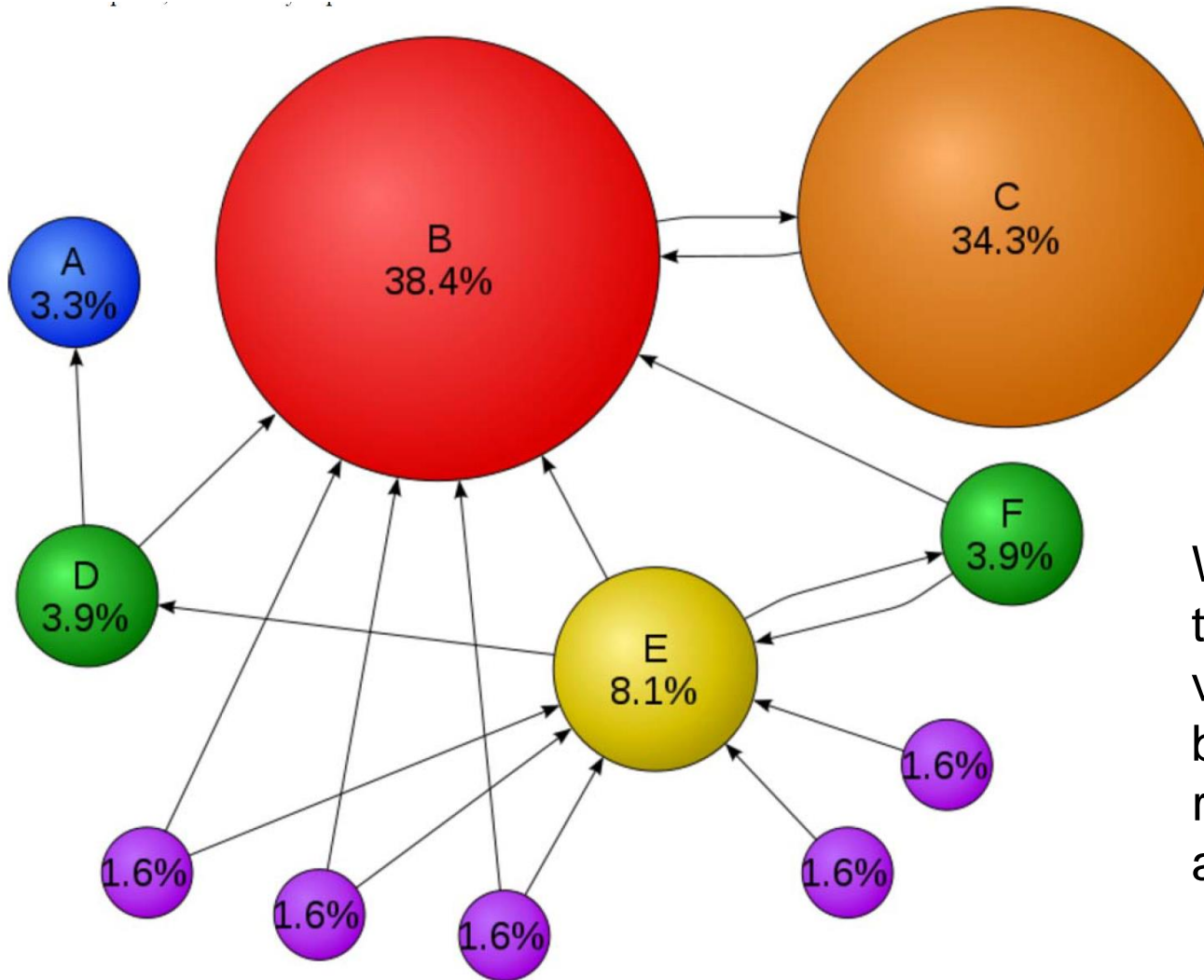
- Continuously
 - Crawling
 - Indexing (in parallel with crawling)
- At query time
 - Retrieving relevant pages
 - Ranking them based on page content
 - Adjust the ranking based on their presence in the web
 - Display the results

- In IR, searching and ranking for “relevant documents” in a collection depends only on the content of the documents (full text search/retrieval)
- In the web, however, in addition to the page content there is the information provided by the hyperlinks from one web page to another
- The idea is therefore to rank the relevance of a web page based also on its “popularity” in the web, i.e. on the “value” of the links pointing to it from other web pages
- This measure is called the PageRank

The PageRank idea



The PageRank values



We can consider the PageRank value as a number between 0 and 1, represented here as a percentage

- Search Engines Optimization (SEO)
 - increase the number of incoming links (link farms)
 - increase the PageRank of the pages pointing to it
 - divide a Web site into many pages
- Collection of query data (for statistics)
 - topics
 - time and location
 - number of clicks
- Advertising on search engines
 - high volume of visitors
 - “knowledge” of web page content
 - targeted advertising

Top Search Engines

| | NETMARKETSHARE | STATISTA | STATCOUNTER |
|------------|----------------|----------|-------------|
| GOOGLE | 72.38% | 86.86% | 92.26% |
| BING | 12.31% | 6.43% | 2.83% |
| BAIDU | 11.26% | 0.68% | 1.14% |
| YAHOO! | 1.86% | 2.84% | 1.59% |
| YANDEX | 1.16% | 0.63% | 0.5% |
| DUCKDUCKGO | 0.45% | N/A | 0.5% |

<https://www.reliablesoft.net/top-10-search-engines-in-the-world/>

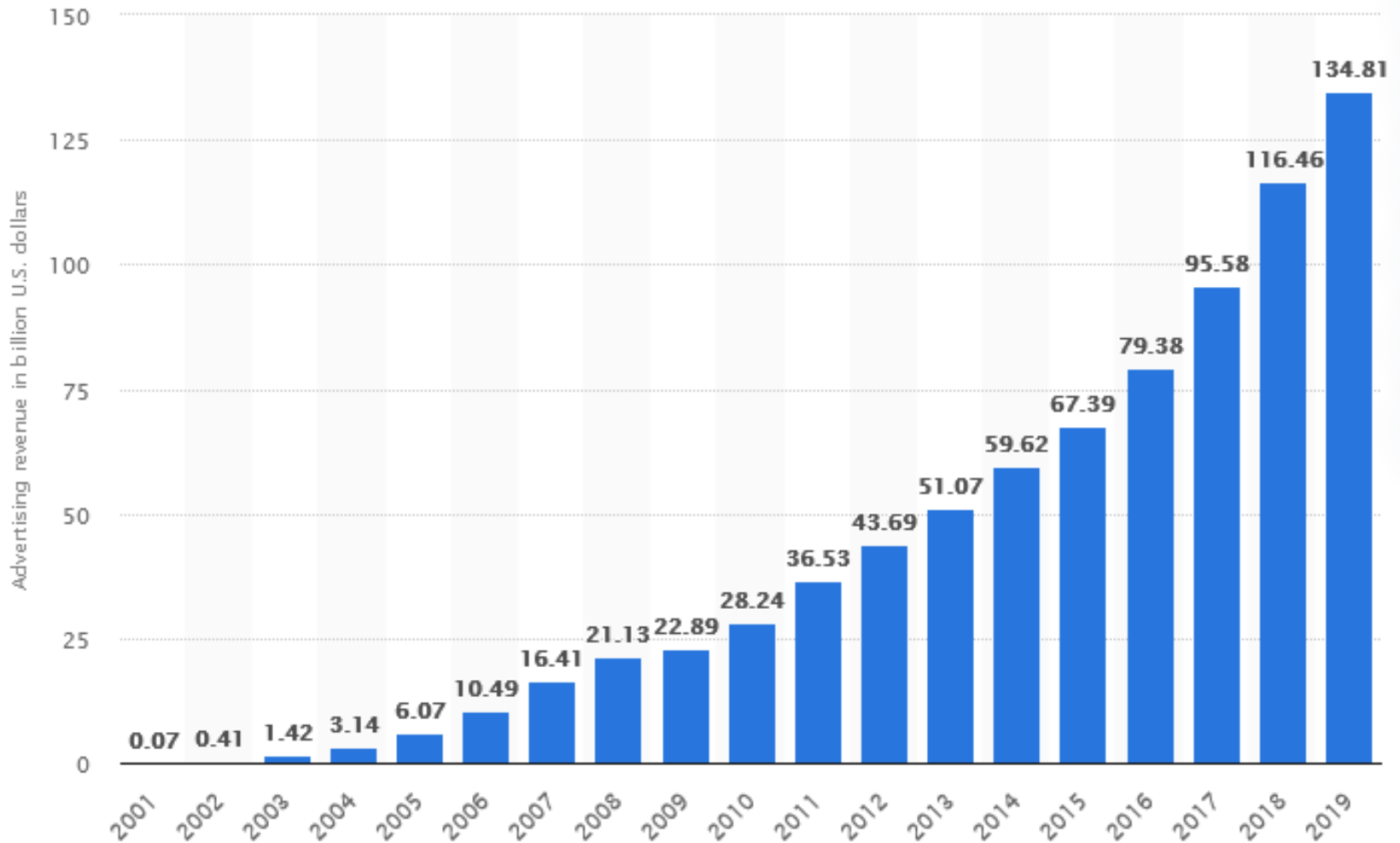
Google searches

| Year | Annual Number of Google Searches | Average Searches Per Day |
|------|---|--------------------------|
| 2016 | 3,293,250,000,000 | 9,022,000,000 |
| 2015 | 2,834,650,000,000 | 7,766,000,000 |
| 2014 | 2,095,100,000,000 | 5,740,000,000 |
| 2013 | 2,161,530,000,000 | 5,922,000,000 |
| 2012 | 1,873,910,000,000 | 5,134,000,000 |
| 2011 | 1,722,071,000,000 | 4,717,000,000 |
| 2010 | 1,324,670,000,000 | 3,627,000,000 |
| 2009 | 953,700,000,000 | 2,610,000,000 |
| 2008 | 637,200,000,000 | 1,745,000,000 |
| 2007 | 438,000,000,000 | 1,200,000,000 |
| 2000 | 22,000,000,000 | 60,000,000 |
| 1998 | 3,600,000 <i>*Googles official first year</i> | 9,800 |



- advertising is associated to “key words” (Google AdWords)
- ads are published on the result page of a query containing a keyword
- ads are paid “per click”
- ads may be published also on “partner sites” (Google AdSense)

Google advertising revenues



Mediators between information and users

- Selection
 - Definition of collections

Google

digital

- Acquisition
 - Physical objects

crawlers, spiders, bots, etc.

- Description
 - Catalogs

Dublin Core

the Web

- Access
 - Shelves

- Preservation
 - Controlled environment

a) forever

b) the next five years
whichever comes first

Conclusions

- Web Search (Google) and Digital Libraries share **similar but complementary missions**
- Celebrate the **diversity of missions**, and **concentrate on strengths** whether as web search engine or digital library
 - Web search engines: scale, universal delivery, universal services
 - Digital libraries: specialized collections, specialized services, “library” services
- Focus on **delivering value to users** through useful and relevant (web) services (“Focus on the user and all else will follow”)
- **Web search** is a service that **Digital Libraries** should exploit to ensure **universal access** to information and services