Information Access in Digital Libraries

Carlo Meghini

CNR - ISTI

Aim

To sketch a conceptual framework within which to understand the many ways of accessing information in a Digital Library

Foundations:

Carlo Meghini, Fabrizio Sebastiani and Umberto Straccia. A model of multimedia information retrieval. *Journal of the ACM*, 48(5):909-970, Sept. 2001

Outline

- Basic definitions
- Information Access
- Personalization
- Distributed Digital Libraries
- Conclusions

Outline

- Basic definitions
- Information Access
- Personalization
- Distributed Digital Libraries
- Conclusions

What is Information Access

- Access: “freedom or ability to obtain or make use of something”
- In a Digital Library, Information Access is the set of tools that enable users to obtain some Resource of the Digital Library

software

programs

Information objects

Phases of information access

User Side

1. Discovery
   - Input: information need
   - Output: an object identifier
2. Request
   - Input: an object identifier
   - Output: a fruition experience

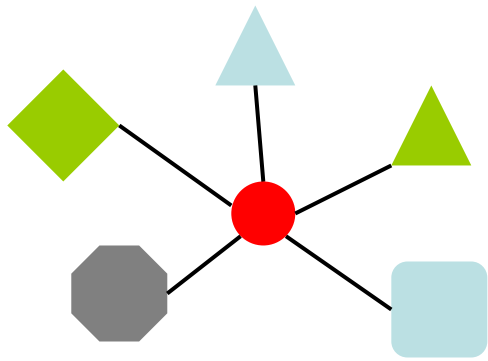
System Side

1. Query evaluation
2. Object retrieval
   a) Check permission
   b) Locate object
   c) Fetch object
   d) Render object

The basic principle of object discovery

Every information object is at the center of a very complex and rich structure.

The parts of this structure are associated to the object via relationships, which can be used as channels for discovering the object.



DELOS Reference Model

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

How?

- The user provides an example and specifies a function to match the example and the object.
- The function can do
  - Exact match
  - Best match

Outline

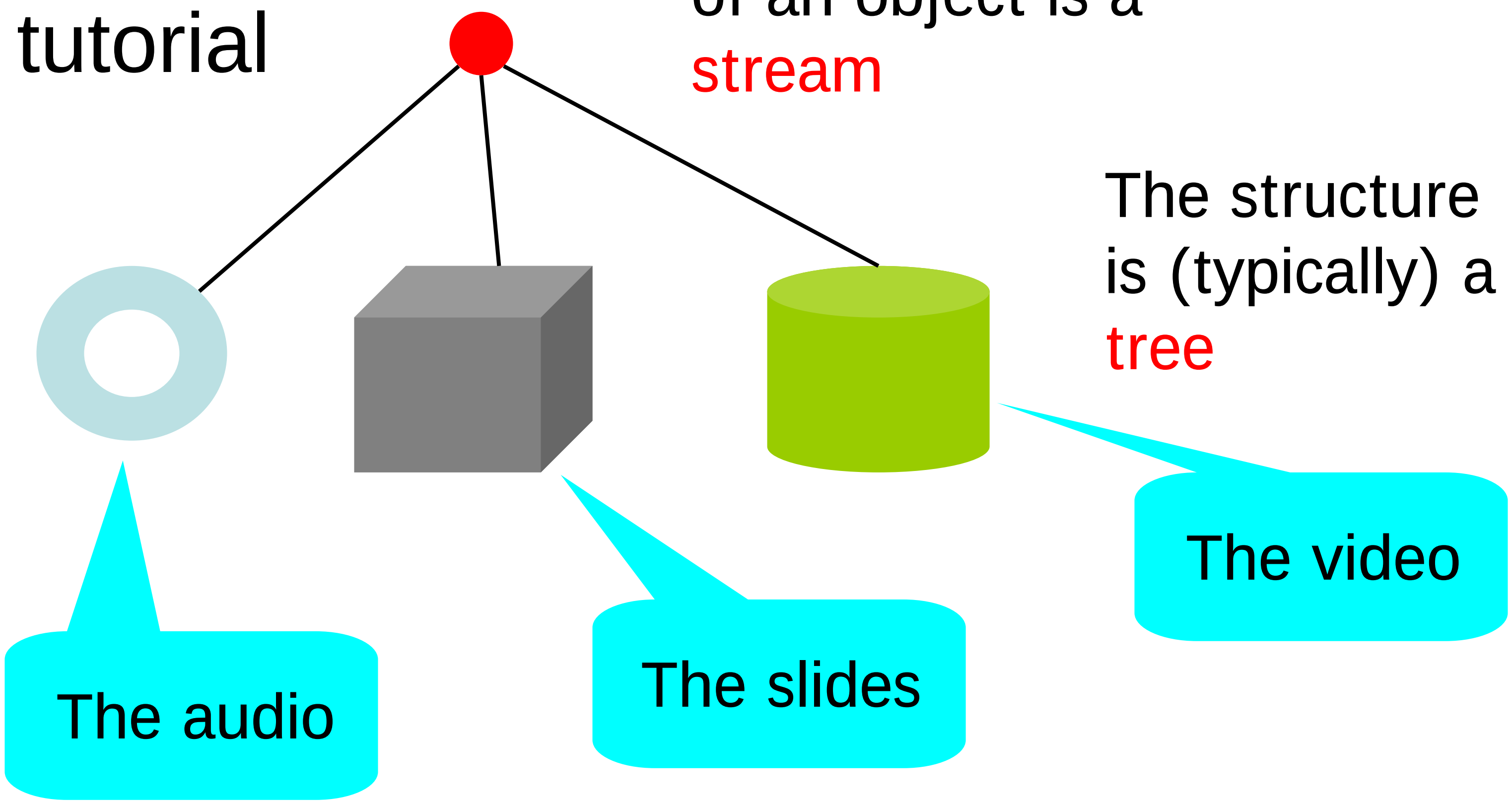
- Basic definitions
- Information Access
  - Discovery via content
- Personalization
- Distributed Digital Libraries
- Conclusions

The content channel

- A tutorial

Each primitive content component of an object is a stream

The structure is (typically) a tree



Discovery through content

- The user provides an example of a component of the object and uses a similarity function to discover the objects similar to the given one
- This is implicitly a *best match* approach

Discovery through content

- Official name: multimedia information retrieval
  - Text retrieval (early 60's) try this
  - Image retrieval (mid 80's) try this or this
  - Audio retrieval (beginning of 90's)
  - Video retrieval (mid 90's)
- Structure-based retrieval (XML)
  - To identify XML documents which have a similar tree-structure to the one given by the user
  - Not really a discovery function, because the user is expected to have already seen the object

Outline

- Basic definitions
- Information Access
  - Discovery via content
  - Discovery via associations
- Personalization
- Distributed Digital Libraries
- Conclusions

Discovery through associations

- Every object is associated with other objects for many different purposes:
  - Descriptive metadata (for discovery)
  - Keywords (for classification)
  - Annotations (for interpretation)
  - Preservation metadata
- Possibly with the aid of additional knowledge structures:
  - Taxonomies
  - Schemas
  - Ontologies

DELOS Reference Model

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

Best match

- In a best match approach, the user specifies a query in natural language
  - Vincenzo Viviani
- the system matches the query against the descriptions *seen as pieces of text*
  - *i.e.* regardless of where the query occurs as a value
- Full-text retrieval
- BRICKS simple search

Exact match

- Foundations: logic
  - The query is an open formula of a language L
  - The descriptions in the DL are seen as an interpretation of $\mathscr{L}$
  - The matching function is satisfiability: an object is returned if its description satisfies the query
- Managing additional knowledge:
  - The additional knowledge is a theory T on L
  - The descriptions in the DL are seen as part of the theory T
  - The matching function is logical implication: an object is returned if its description, conjoined with T, logically implies the query

In practice

- The techniques for managing information access vary depending on the form of descriptions and the form of the additional knowledge.
- Descriptions can be classified as:
  - Keyword-based
    - Can be managed with IR techniques
  - Record-based
    - Can be managed with database techniques
  - Net-based
    - Can be managed with database techniques if used in isolation
    - Must be managed with knowledge techniques if used in conjunction with ontologies

Keyword-based access

- Every document is assigned a set of keywords (index) from a (more or less) controlled vocabulary, including collection names
  - iconography, bibliography
  - acm:information retrieval
- The query is a Boolean combination of keywords
  - my-pictures AND NOT tintoretto
- The result consists of those objects whose indices satisfy the query

Keyword-based access

- This form of access was the first form of information retrieval, known as Boolean retrieval
- No longer used in large text collections
- Still used in accessing information through OPACs

Taxonomies

- A taxonomy is a relation between terms, capturing a specialization/generalization concept *e.g.* Yahoo Directory
  - Query: mortal
  - Object plato is indexed as: man
  - Taxonomy: mortal > man
  - plato is discovered by the query mortal

Folksonomies

- A folksonomy is a user-generated taxonomy used to categorize and retrieve web content such as Web pages, photographs and Web links, using open-ended labels called tags.
- Typically, folksonomies are Internet-based, but their use may occur in other contexts.
- Two widely cited examples of websites using folksonomic tagging are Flickr and del.icio.us.

Query expansion

- The technique to deal with taxonomies is called query expansion:
  - “mortal” is expanded into “man OR mortal” and evaluated as a Boolean query
- This technique is employed also in
  - Thesaurus-based retrieval (*e.g.* Wordnet)
    - Synonyms and specializations are included in the expanded query
  - Cross-language retrieval
    - Translations of query terms are included in the expanded query

Record-based access

- Record-based descriptions are sets of (attribute, value) pairs
  - Dublin Core metadata records
- Queries are Boolean combinations of simple conditions on attribute values
  - dc:creator CONTAINS “carlo” OR dc:date > 01.01.2004
  - [(Ey) dc:creator(x,y) AND CONTAINS(y,”carlo”)] OR
    [(Ez) dc:date(x,z) AND > (z, 01.01.2004)]
  - dc:creator, dc:date are user-defined predicates, while CONTAINS and > are predicates with fixed semantics

Record-based access

- Objects whose description satisfies the query are discovered
- This is classical database-like information access, for which we have a well consolidated technology
  - Relational DBMSs (or OO DBMSs)
  - SQL (OQL)

  which can handle up to millions of records efficiently

Net-based access

- Descriptions are bundles of objects connected by arcs, forming networks
  - Early net-based models appeared in the 70's, then termed as semantic networks or frames
  - Lack of semantics led to the formalization of these models in terms of Description Logics (mid 80's)
  - Families of Description Logics were studied for about 15 years from many point of views:
    - Logical
    - Computational
    - Pragmatical

Net-based access

- The representational principles of net-based models have been recently re-used in the context of the Semantic Web
- Result: Resource Description Framework (RDF) and its follow-ups:
  - RDF Schema
  - Ontology Web Language (OWL), in its 3 flavours:
    - OWL Light
    - OWL DL
    - OWL Full

Resource Description Framework

XML Notation for RDF

```
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
            xmlns:contact="http://www.w3.org/2000/10/swap/pim/contact#">

  <contact:Person rdf:about="http://www.w3.org/People/EM/contact#me">
    <contact:fullName>Eric Miller</contact:fullName>
    <contact:mailbox rdf:resource="mailto:em@w3.org"/>
    <contact:personalTitle>Dr.</contact:personalTitle>
  </contact:Person>

</rdf:RDF>
```

Net-based access

- A query is a logical expression referring triples, and is evaluated on the RDF graph.
- Query languages for RDF have recently been proposed:
  - SPARQL (W3C Working Draft 26 March 2007)

```
PREFIX abc: <http://mynamespace.com/exampleOntologie#>
SELECT ?capital ?country
WHERE {
  ?x abc:cityname ?capital.
  ?y abc:countryname ?country.
  ?x abc:isCapitalOf ?y.
  ?y abc:isInContinent abc:africa.
}
```

(E ?x)(E ?y) abc:cityname(?x, ?capital) AND abc:countryname(?y, ?country) AND abc:isCapitalOf(?x, ?y) AND isInContinent(?y, abc:africa)

Schemas

- The terms used in net-based descriptions are defined in schemas.
  - An RDF schema defines the terms used in RDF descriptions.
- Notions of RDF schema:
  - Classes, organized in a taxonomy
  - Properties, organized in a taxonomy

Schemas

http://www.example.org/schemas/vehicles#MotorVehicle
http://www.w3.org/2000/01/rdf-schema#subClassOf
http://www.example.org/schemas/vehicles#Truck
http://www.w3.org/2000/01/rdf-schema#subClassOf
http://www.example.org/schemas/vehicles#Van
http://www.w3.org/2000/01/rdf-schema#subClassOf
http://www.example.org/schemas/vehicles#PassengerVehicle
http://www.w3.org/2000/01/rdf-schema#subClassOf
http://www.w3.org/2000/01/rdf-schema#subClassOf
http://example.org/schemas/vehicles#MiniVan

XML Notation

```
<rdfs:Class rdf:ID="MotorVehicle"/>

<rdfs:Class rdf:ID="PassengerVehicle">
  <rdfs:subClassOf rdf:resource="#MotorVehicle"/>
</rdfs:Class>

<rdfs:Class rdf:ID="Truck">
  <rdfs:subClassOf rdf:resource="#MotorVehicle"/>
</rdfs:Class>

<rdfs:Class rdf:ID="Van">
  <rdfs:subClassOf rdf:resource="#MotorVehicle"/>
</rdfs:Class>

<rdfs:Class rdf:ID="MiniVan">
  <rdfs:subClassOf rdf:resource="#Van"/>
  <rdfs:subClassOf rdf:resource="#PassengerVehicle"/>
</rdfs:Class>
```

```
<rdfs:Class rdf:ID="Person"/>

<rdfs:Datatype rdf:about="&xsd;integer"/>

<rdf:Property rdf:ID="registeredTo">
  <rdfs:domain rdf:resource="#MotorVehicle"/>
  <rdfs:range rdf:resource="#Person"/>
</rdf:Property>

<rdf:Property rdf:ID="rearSeatLegRoom">
  <rdfs:domain rdf:resource="#PassengerVehicle"/>
  <rdfs:range rdf:resource="&xsd;integer"/>
</rdf:Property>

<rdf:Property rdf:ID="driver">
  <rdfs:domain rdf:resource="#MotorVehicle"/>
</rdf:Property>

<rdf:Property rdf:ID="primaryDriver">
  <rdfs:subPropertyOf rdf:resource="#driver"/>
</rdf:Property>
```

Ontology (from Barry Smith)

- Ontology as a branch of philosophy is the science of what is, of the kinds and structures of objects, properties, events, processes and relations in every area of reality.
- “Ontology” is often used by philosophers as a synonym of “metaphysics”
- Ontology seeks to provide a definitive and exhaustive classification of entities in all spheres of being.
  - What classes of entities are needed for a complete description and explanation of all the goings-on in the universe?
- It should be exhaustive in the sense that all types of entities should be included in the classification, including also the types of relations by which entities are tied together to form larger wholes.

Ontologies

- Ontologies focus on parts of reality
- Ontologies formalize a shared vocabulary about a domain.
- Their importance stems from the fact that they offer well thought out terminologies for different domains that can be shared and reused.

Ontologies

- Ontologies can be classified into three main categories:
  - upper
  - core
  - domain
- Upper ontologies (e.g., Cyc and WordNet) include general, domain-independent terms.
- Core -- or intermediate -- ontologies cover broad domains, such as audiovisual phenomena.
- Domain ontologies are specific to a domain, such as manufacturing, history, or soccer.

Ontologies *vs* Schemas

- Ontologies give more semantics than schemas, by specifyng constraints which may not be expressible in schemas
  - cardinality constraints on properties, e.g., that a Person has exactly one biological father.
  - that a given property (such as ex:hasAncestor) is transitive, e.g., that if A ex:hasAncestor B, and B ex:hasAncestor C, then A ex:hasAncestor C.
  - that a given property is a unique identifier (or key) for instances of a particular class.
  - that two different classes (having different URIrefs) actually represent the same class.
  - that two different instances (having different URIrefs) actually represent the same individual.
  - to describe new classes in terms of combinations (e.g., unions and intersections) of other classes
  - to say that two classes are disjoint (i.e., that no resource is an instance of both classes).
- Schemas give more implementation details than ontologies, by specifying which data types are used for implementing which ontological notions

Ontologies and information access

- An ontology (or a schema) can help the user to better understand the content of a DL
  - Browsing concepts and relationships
  - Query formulation
- Ontologies cannot in general be directly used for information access because of computational reasons.

Ontologies and information access

- Ontology:
  - PET = CAT or (BIRD and not OWL)
  - OWL is disjoint from SPARROW
  - SPARROW is-a BIRD
- Description:
  - Fido is SPARROW
- Query: PET
- Ontology + Description imply that Fido is PET
- Deriving this knowledge requires reasoning, and reasoning is computationally expensive

Semantic Interoperability

- When crossing DL boundaries, one finds different ontologies for describing the same, or similar concepts.
- This leads to the problem of Semantic Interoperability.
- Semantic interoperability is the capability of an information system to correctly interpret information coming from a different system, or to manage communicated information consistently with its intended meaning (i.e., as intended by its creators/owners).
- Semantic interoperability was recognized as a major technological challenge in AI in the early ‘90s and led to DARPA’s Knowledge Sharing initiative.
- In Databases, semantic interoperability became a major issue during the same period thanks to the web, as well as trends towards enterprise integration.

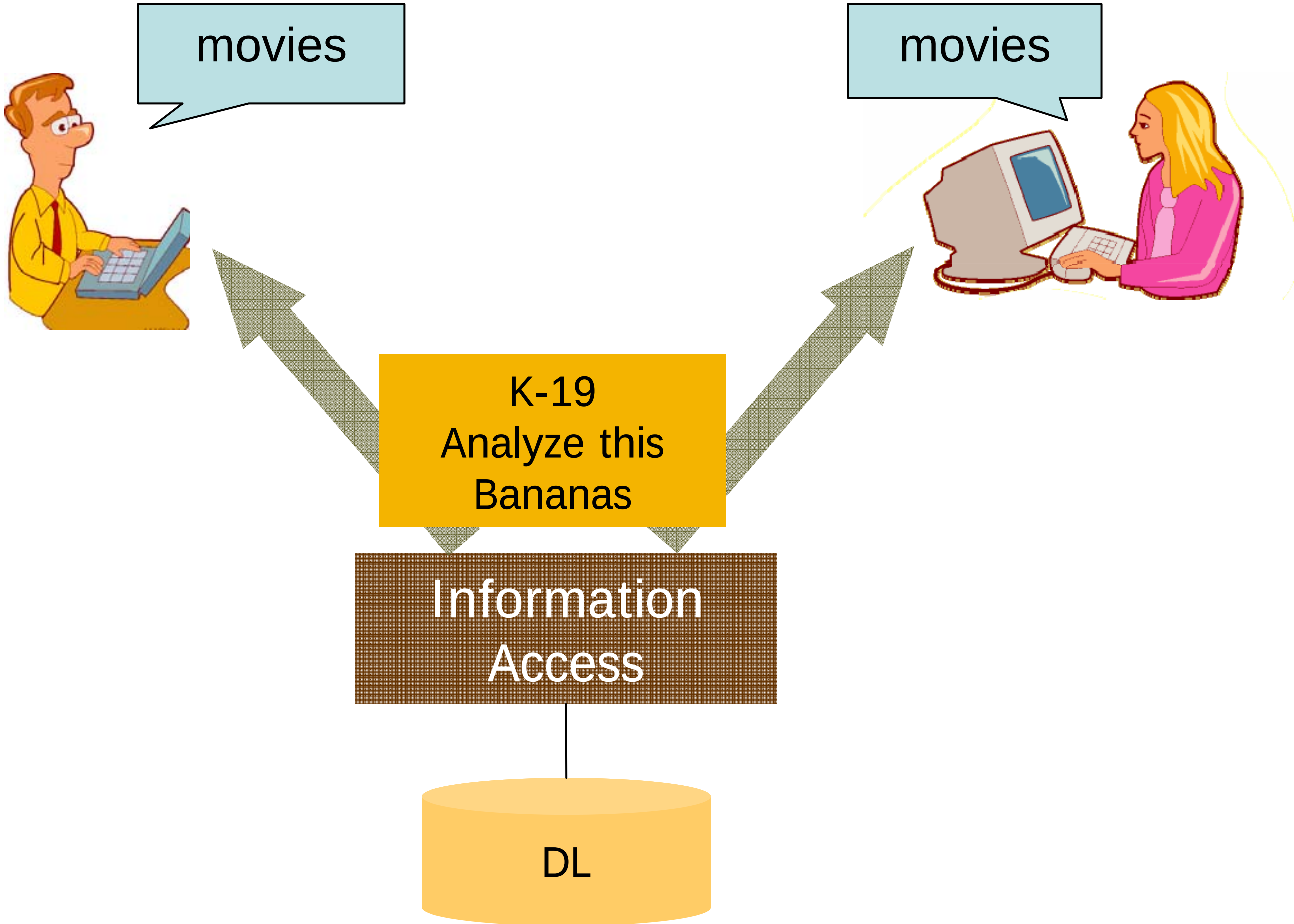
Outline

- Basic definitions
- Information Access
- Personalization
- Distributed Digital Libraries
- Conclusions

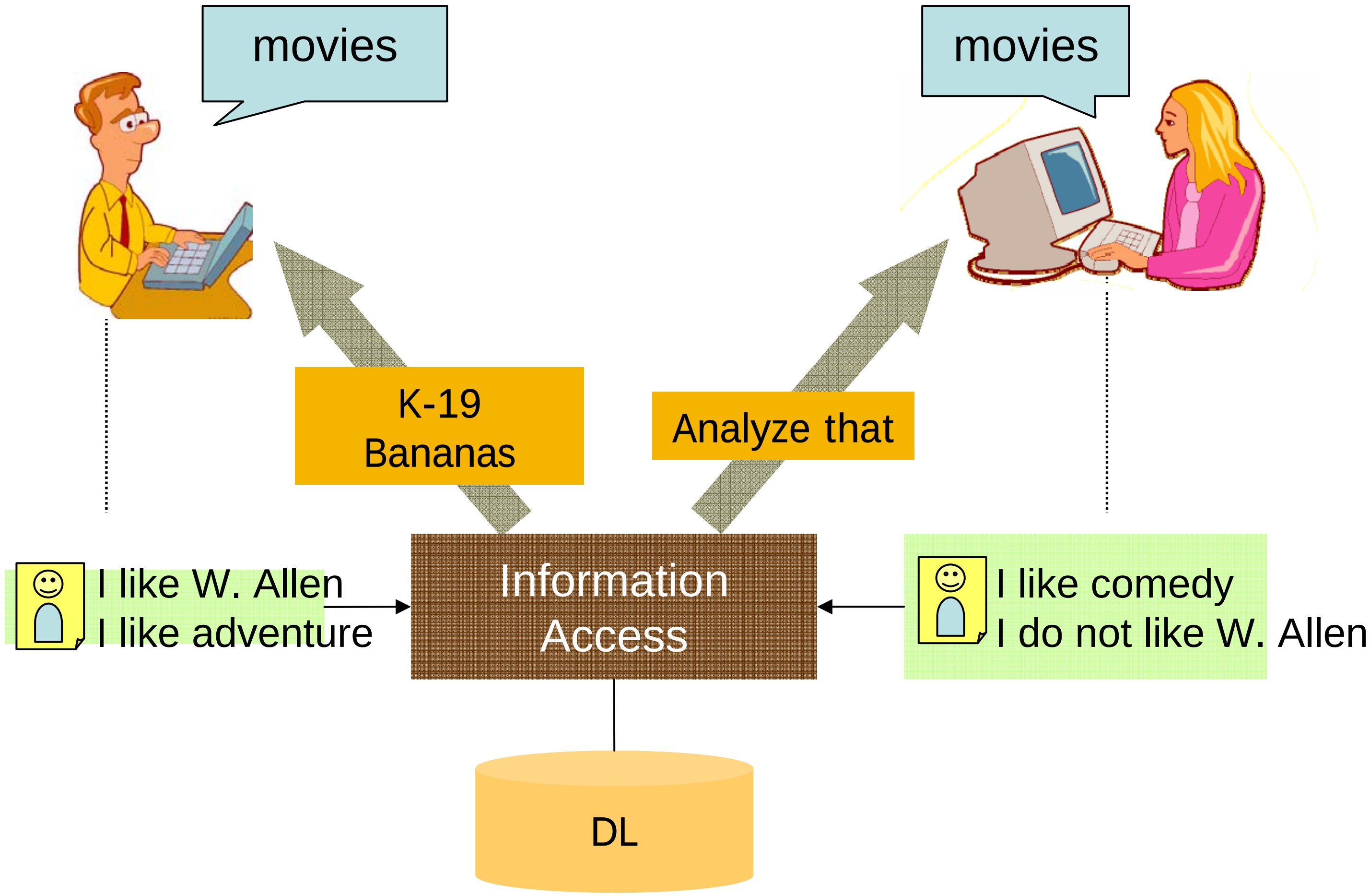
Personalization

- Personalization aims at offering to the users of a DL services which take into account their preferences.
- Every user is described by a profile
  - Identity
  - Access Control
  - Preferences

Un-personalized access



Effect of personalization



Qualitative personalization

Qualitative approach

I prefer comedies to adventures

Preferences between objects are expressed using **preference relations**

I like A better than B

Quantitative personalization

Quantitative approach

I like comedies very much

I like adventures a little

Preference for an object is expressed using **scores**

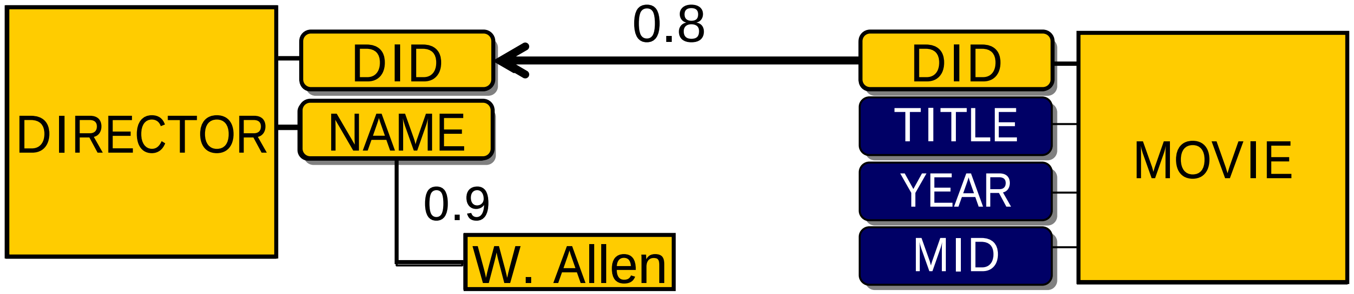
I (do not) like A that much

Quantitative preference modeling

Preference

< q, d> atomic condition q

degree of interest d $\in$ [0, 1]



< MOVIE.did = DIRECTOR.did, 0.8 >

< DIRECTOR.name = 'W.Allen', 0.9 >

Stored preferences

< MOVIE.did = DIRECTOR.did and DIRECTOR.name = 'W.Allen', 0.9 $*$ 0.8 >

Implicit preferences

Personalized information access

- Query re-writing
  - The original user query is re-written to enforce user preferences on the selected aspects
- Result re-ranking
  - The order in which query results are shown to the user is altered to make "interesting" objects higher in the rank

Outline

- Basic definitions
- Information Access
- Personalization
- Distributed Digital Libraries
- Conclusions

Distributed Digital Libraries

- Gathering descriptions for information access
- Managing query evaluation
- Coping with Syntactic and Semantic Interoperability

Gathering descriptions

- Pull mode:
  - Harvesting
    - OAI MHP
  - Crawling
    - Web search engines
- Push mode:
  - RSS

Managing query evaluation

- Query evaluation: a mediation process between the local query evaluators
- Optimization issues:
  - Parallelization
  - Index centralization
  - Asynchrony in result delivery

Syntactic Interoperability

- Service-based architectures
  - Web services

Semantic Interoperability

- Horizontal approach: ontology mapping
- Vertical approach: ontology integration
  - Merging
  - Mapping to a common ancestor (CIDOC CRM)

Outline

- Basic definitions
- Information Access
- Personalization
- Distributed Digital Libraries
- Conclusions

Conclusions

- Information access is still an open research field
  - Easy things are easy
  - More interesting things are hard!
- … and will remain so for some time
  - Knowledge is the basic good
  - Knowledge is hard to collect, represent, process, exchange, evolve, integrate
    - We basically do not know how to do it
    - Semantic Interoperability goes back to the Babel Tower
- Let's keep going!

Questions