



College of Information Studies

University of Maryland Hornbake Library Building College Park, MD 20742-4345

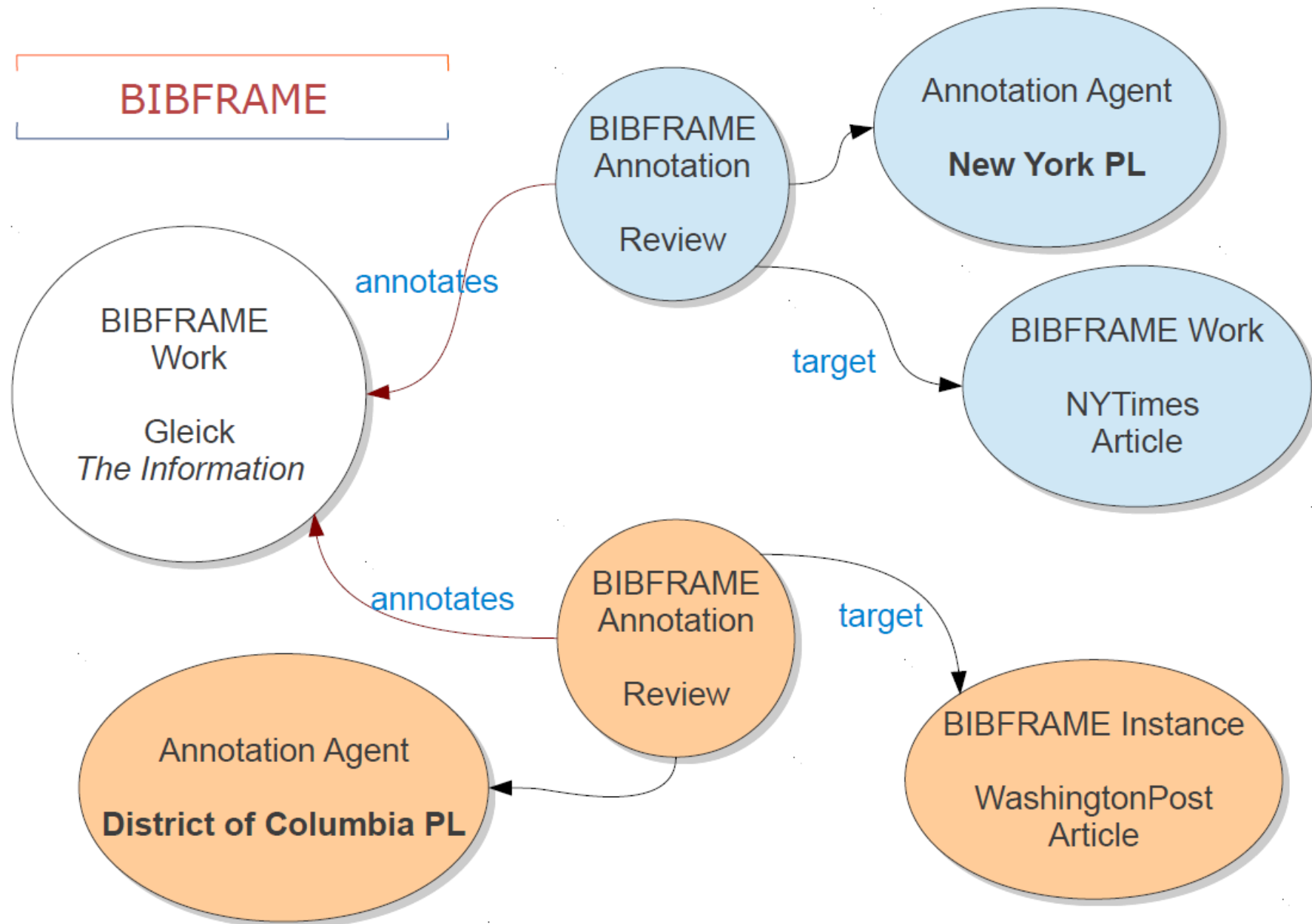
Discovery and Delivery

Week 7

LBSC 671

Creating Information Infrastructures

BIBFRAME



Tonight

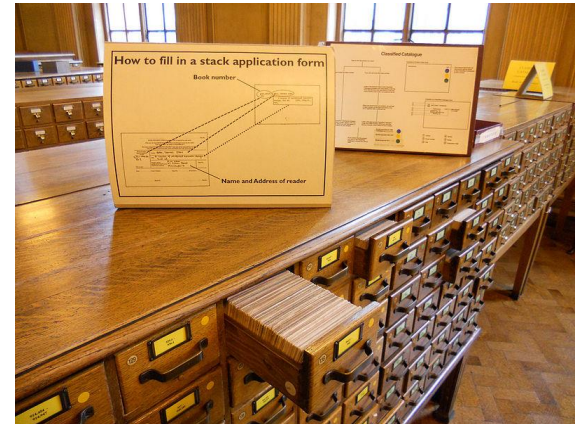
- Access points
- Discovery
- Delivery
- Midterm exam review

Authority Control

- Unify references to the same entity (synonyms)
 - Samuel Clemens, Mark Twain
- Distinguish references to different entities (homonyms)
 - Michael Jordan (basketball), Michael Jordan (computers)
- Establish “access points”
 - Canonical and variant forms, to better support “find it” tasks

Access Points

- Originally designed for card catalogs
 - One card for every “authorized” access point
- Four types “dictionary” catalog access points
 - Title (uniform titles)
 - Author (name authority)
 - Subject (controlled vocabulary)
 - Series
- Other things can serve a similar purpose
 - Call number (shelf order)
 - “Keywords” (full-text search)



Functional Requirements for Authority Data (FRAD)

- Name
 - Canonical form for display to users
- Identifier
 - Canonical form for use by systems
- Controlled access points
 - Forms that can be used as a basis for access
- Rules
 - For creating access points
- Agency
 - Organization responsible for creating access points

FRBR Bibliographic User Tasks

- Find it
 - Search (“to find”)
 - Recognize (“to identify”)
 - Choose (“to select”)
- Serve it
 - Location (“to obtain”)

FRAD Authority Control User Tasks

- Searcher tasks
 - Find
 - Identify
- Authority control tasks
 - Contextualize
 - Justify



LIBRARY OF CONGRESS AUTHORITIES

[Help](#)[New Search](#)[Search History](#)[Headings List](#)[Start Over](#)

SOURCE OF HEADINGS: Library of Congress Online Catalog

INFORMATION FOR: Oard, Doug

Please note: Broader Terms are not currently available

Select a Link Below to Continue...

[Authority Record](#)

See: [Oard, Douglas W.](#)

LC control no.: no 97043761

LCCN permalink: <http://lccn.loc.gov/no97043761>

Personal name heading: Oard, Douglas W.

Variant(s): Oard, Doug

Found in: A survey of information retrieval and filtering methods, 1995: t.p. (Douglas W. Oard) p. 1 (Electrical Engineering Dept., University of Maryland, College Park, MD)

Cross-language text & speech retrieval, c1997: t.p. (Doug Oard)

Hands On

- Find the authoritative LC name for one of ...
 - <http://ischool.umd.edu/faculty-staff/jennifer-j-preece>
 - <http://www.umiacs.umd.edu/~jimmylin/>
 - <http://terpconnect.umd.edu/~pwang/>
 - http://en.wikipedia.org/wiki/Robert_S._Taylor
 - http://en.wikipedia.org/wiki/Hans_Peter_Luhn

Entity Linking



Knowledge Base

Query

Suzanne Collins




Suzanne Collins at Time 100 Gala

Born	1962 ^[1] Connecticut
Occupation	Television scriptwriter
Nationality	United States
Genres	Fantasy Science fiction Children Young adult Suspense Action

suzannecollinsbooks.com

Jackie Collins



Jackie Collins (July 200

Born	Jacqueline Jill Collins 4 October 1937 (age London, England, UK
Occupation	Novelist
Spouse	Wallace Austin (m. 1960-1964, divor Oscar Lerman (m. 1966-1992, his c
Children	Tracy Lerman (b. 19

Susan Collins



**United States Senator
from Maine**

Incumbent

Assumed office

'Hunger Games' shooting wraps in North Carolina

We're one step closer to watching Suzanne Collins' dystopian action series come to life onscreen. "The Hunger Games" has wrapped, according to [The Hollywood Reporter](#).

Principal photography on the Lionsgate adaptation was completed in North Carolina, where the movie was shot, last Saturday.

"The Hunger Games" was initially planned as a mid-budget flick for youngsters, but due to the wide-ranging popularity of the book series, it's become a nearly \$100-million production, reports the [Los Angeles Times](#). Lionsgate is targeting movie-goers of all ages, not just teens.

The trilogy has over 12 million copies in print.

Entity Linking



✧ Given

- ✧ A mention of a person's name in a document
- ✧ A “knowledge base” containing information about a set of known entities

✧ Determine

- ✧ Whether the mentioned person is in the knowledge base
- ✧ If so, where

✧ Match unstructured text to structured knowledge source

✧ Related to:

- ✧ Record linkage: Structured to structured
- ✧ Co-reference resolution: Unstructured to unstructured

Entity Linking Task

Michael Phelps

Debbie Phelps, the mother of swimming star **Michael Phelps**, who won a record eight gold medals in Beijing, is the author of a new memoir, ...

Michael Phelps

Michael Phelps is the scientist most often identified as the inventor of PET, a technique that permits the imaging of biological processes in the organ systems of living individuals. **Phelps** has ...



818k+ entries

Michael Phelps	swimmer	1985-
Michael E Phelps	biophysicist	1939-
Mike Phelps	basketball player	1961-
Edmund Phelps	economist	1933-
...		

**Identify matching entry, or determine that entity is missing from KB.
Non-trivial due to name ambiguity, name variation, & KB absence.**

Technical Approach

“According to the CDC the prevalence of H1N1 influenza in California prisons has increased ...”

Query = “CDC”

- California Dept. of Corrections
- Cedar City Regional Airport
- Cheerdance Competition
- Communicable Disease Centre
- Congress for Democratic Change
- Consumers for Dental Choice
- Control Data Corporation
- Cult of the Dead Cow
- NIL (Absence from KB)
- US Center for Disease Control
- ...

Several phases

- 1. Candidate identification (“triage”) based on target name

Technical Approach

“According to the CDC the prevalence of H1N1 influenza in California prisons has...”

Query = “CDC”

1. California Dept. of Corrections
2. US Center for Disease Control
3. Cedar City Regional Airport (IATA code)
4. Communicable Disease Centre (Singapore)
5. Congress for Democratic Change (Liberian political party)
6. Cult of the Dead Cow (Hacker organization)
7. Control Data Corporation
8. NIL (Absence from KB)
9. Consumers for Dental Choice (non-profit)
10. Cheerdance Competition (Philippine organization)

Several phases

- 1. Candidate identification (“triage”) based on target name
- 2. Candidate selection (“ranking”) exploiting document features using supervised machine learning

Technical Approach

“According to the CDC the prevalence of H1N1 influenza in California prisons has...”

Query = “CDC”

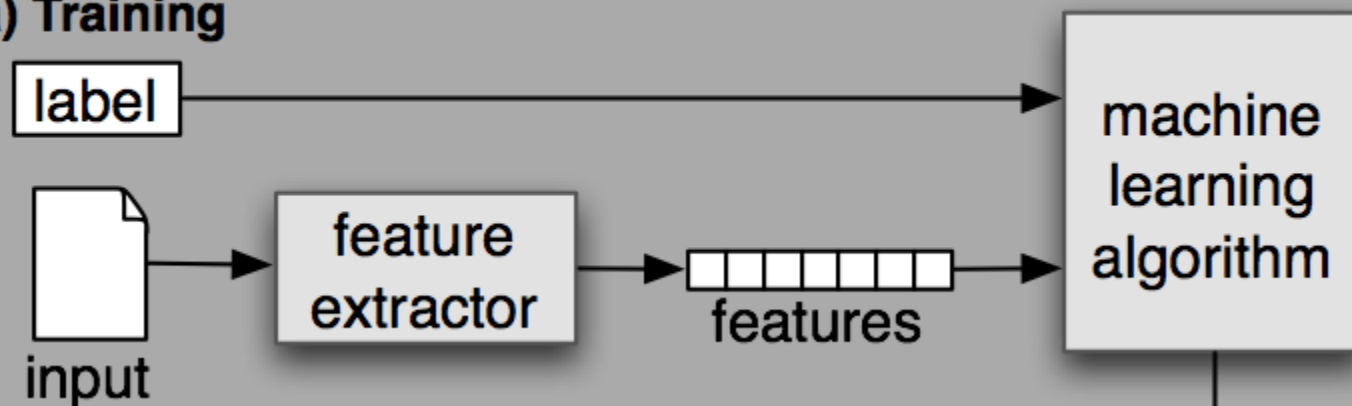
1. California Dept. of Corrections
2. US Center for Disease Control
3. Cedar City Regional Airport (IATA code)
4. Communicable Disease Centre (Singapore)
5. Congress for Democratic Change (Liberian political party)
6. Cult of the Dead Cow (Hacker organization)
7. Control Data Corporation
- 8. NIL (Absence from KB)**
9. Consumers for Dental Choice (non-profit)
10. Cheerdance Competition (Philippine organization)

Several phases

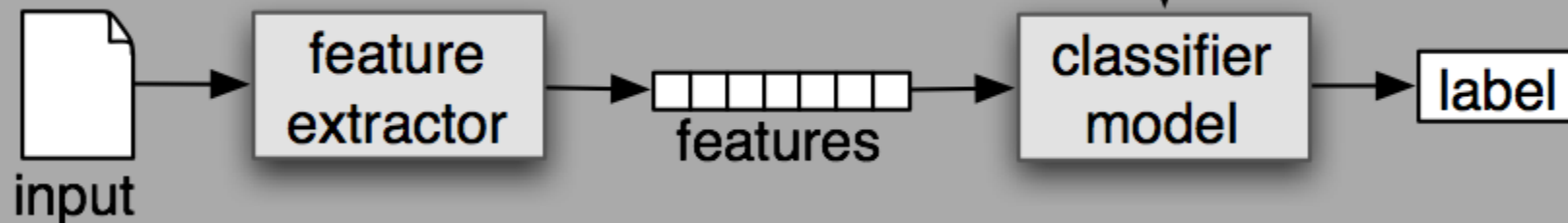
- 1. Candidate identification (“triage”) based on target name
- 2. Candidate selection (“ranking”) exploiting document features using supervised machine learning
- 3. Possibly choosing absence (NIL)

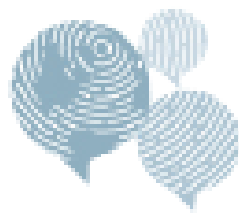
Supervised Machine Learning

(a) Training



(b) Prediction

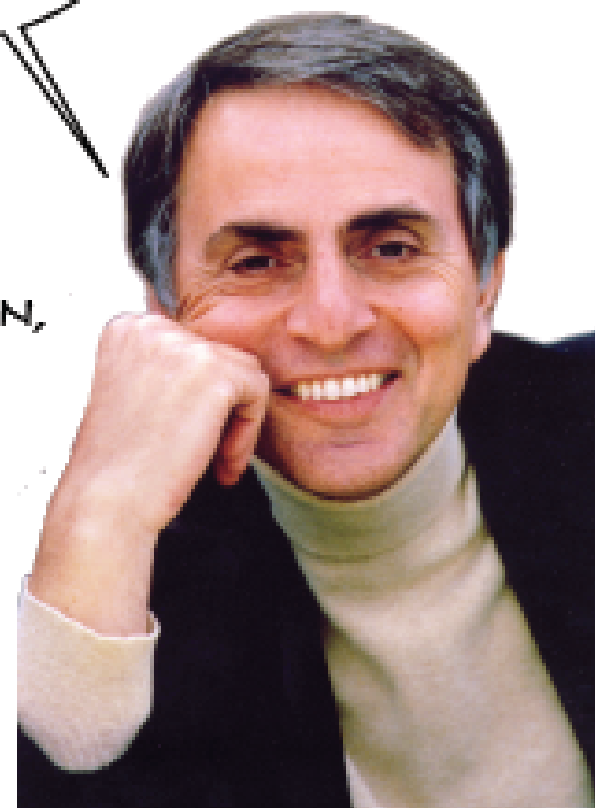




FEATURES

- ☐ NAME-MATCHING
 - ☐ ACRONYMS, ALIASES, STRING SIMILARITY
- ☐ DOCUMENT FEATURES
 - ☐ TF/IDF-WEIGHTED COMPARISONS, OCCURRENCE OF KB FACTS IN QUERY TEXT
- ☐ ENTITY TYPE, NAMED ENTITY CO-OCCURRENCES
 - ☐ TYPE (I.E., IS THIS A PERSON, ORGANIZATION, LOCATION?)
 - ☐ DO OTHER ENTITIES CO-OCCUR IN QUERY DOCUMENT AND KB RECORD?
- ☐ ABSENCE (NIL INDICATIONS)
 - ☐ DOES ANY CANDIDATE LOOK LIKE A VIABLE MATCH?

BILLIONS AND
BILLIONS OF
FEATURES



Cross-Language Entity Linking

زيارة شارون جاءت قبل شهرين فقط من قرار الحكومة البريطانية النهائي حول صفقة صواريخ تسعى إسرائيل من ورائها إلى بيع خمسة آلاف من صواريخ "سبايك" إسرائيلية الصنع المضادة للدبابات وأكدت المصادر أن الصفقة كانت أحد الموضوعات الرئيسية على مائدة المفاوضات بين شارون ونظيره البريطاني **توني بلير**. كانت وزارة الدفاع البريطانية قد أبرمت صفقة مبدئية مع الحكومة الإسرائيلية في عام 2001 اشترت بمقتضاها عدداً من

...main issues on the negotiating table between Sharon and **Blair**. It is noteworthy that the British Defense Ministry had clinched a ...

الصفقة الحاسمة التي سوف تفضي إليها إلى 200 مليون جنيه إسترليني.

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact Wikipedia

Toolbox
Print/export

Tony Blair
From Wikipedia, the free encyclopedia


For other uses, see Tony Blair (disambiguation).

Anthony Charles Lynton Blair (born 6 May 1953)^[1] is a former British Labour Party politician who served as the Prime Minister of the United Kingdom from 2 May 1997 to 27 June 2007. He was the Member of Parliament (MP) for Sedgefield from 1983 to 2007 and Leader of the Labour Party from 1994 to 2007. He resigned from all of these positions in June 2007.

Tony Blair was elected Leader of the Labour Party in the leadership election of July 1994, following the sudden death of his predecessor, John Smith. Under his leadership, the party adopted the term "New Labour"^[2] and moved away from its traditional left wing position towards the centre ground.^{[3][4]} Blair subsequently led Labour to a landslide victory in the 1997 general election. At 43 years old, he became the youngest Prime Minister since Lord Liverpool in 1812. In the first years of the New Labour government, Blair's government implemented a number of 1997 manifesto pledges, introducing the minimum wage, Human Rights Act and Freedom of Information Act, and carrying out regional devolution, establishing the Scottish Parliament, the National Assembly for Wales, and the Northern Ireland Assembly.

Blair's role as Prime Minister was particularly visible in foreign and security policy, including in

The Right Honourable
Tony Blair



Prime Minister of the United Kingdom
In office
2 May 1997 – 27 June 2007

Monarch	Elizabeth II
Deputy	John Prescott
Preceded by	John Major
Succeeded by	Gordon Brown

0.62

0.47

Doc: ummah20031216_003

Cross-Language Entity Linking



Knowledge Base

Query

Suzanne Collins



Suzanne Collins at Time 100 Gala

Born	1962 ^[1] Connecticut
Occupation	Television scriptwriter
Nationality	United States
Genres	Fantasy Science fiction Children Young adult Suspense Action

suzannecollinsbooks.com

Jackie Collins



Jackie Collins (July 200

Born	Jacqueline Jill Collins 4 October 1937 (age London, England, UK
Occupation	Novelist
Spouse	Wallace Austin (m. 1960-1964, divor Oscar Lerman (m. 1966-1992, his c
Children	Tracy Lerman (b. 19

Susan Collins



**United States Senator
from Maine**

Incumbent

Assumed office

اطلاق النار 'الاعاب ال جوع' في
ولاية كارولينا الشمالية يلتف

نحن خطوة واحدة لمشاهدة سلسلة سوزان كولينز على شاشة التلفزيون في ولاية كارولينا الشمالية، حيث تم إطلاق النار على الغيليم، يوم السبت الماضي.

استقبل التصوير الفوتوغرافي بشكلى أساسى على التليفزيون في ولاية كارولينا الشمالية، حيث تم إطلاق النار على الغيليم، يوم السبت الماضي.

"إن دورة الألعاب الجوع" وكان من المقرر في البداية باعتباره نقض الغبار من تصرف الميزانية للشباب، ولكن نظرا لشعبية واسعة النطاق من سلسلة الكتاب، أصبح من إنتاج ما يقرب من 100 مليون دولار، وتقارير صحيفة لوس أنجلوس تايمز. يوزجيت تستهدف رواد السينما من جميع الأعمار، وليس فقط المراهقين.

ثلاثية وأكثر من 12 مليون نسخة في الطباعة.

One-Best Person Linking Accuracy

Language	English	Cross-Language	
Arabic	0.9480	0.9263	(97.7%)
Bulgarian	0.9818	0.9350	(95.2%)
Czech	0.9310	0.8585	(92.2%)
Danish	0.9883	0.9789	(99.0%)
German	0.9309	0.9128	(98.1%)
Greek	0.9794	0.8840	(90.3%)
Spanish	0.9087	0.8750	(96.3%)
Finnish	0.9859	0.9368	(95.0%)
French	0.9301	0.8930	(96.0%)
Croatian	0.9799	0.9497	(96.9%)
Italian	0.9842	0.9009	(91.5%)
Macedonian	0.9778	0.8950	(91.5%)
Dutch	0.9841	0.9751	(99.1%)
Portuguese	0.9865	0.9269	(94.0%)
Romanian	0.9761	0.9738	(99.8%)
Albanian	0.9717	0.9191	(94.6%)
Serbian	0.9762	0.8370	(85.7%)
Swedish	0.9866	0.9710	(98.4%)
Turkish	0.9801	0.9642	(98.4%)
Urdu	0.9725	0.8763	(90.1%)
Average	0.9680	0.9195	(95.0%)

Classification

- Classification
 - A system for organizing knowledge
- Notation
 - Expressing the classification in a systematic way

Library of Congress Subject Headings

- Controlled vocabulary for subject access points
 - Most commonly applied to books and serials
- Used when a subject describes $\geq 20\%$ of the work
- Choose the most specific appropriate headings
 - But if more than 3 subtopics, choose a broader heading

LCSH Subdivisions

- Topical

Archaeology – Methodology

- Form

Archaeology – Fiction

- Chronological

Archaeology – History – 18th century

- Geographic

Archaeology – Egypt

Hands On

- Find the LCSH for one of:
 - <http://www.mayoclinic.com/health/heart-attack/DS00094>
 - <http://en.wikipedia.org/wiki/AS-204>
 - <http://www.apollotheater.org/>
 - <http://www.flickr.com/photos/usnationalarchives/4153755504/>
 - http://en.wikipedia.org/wiki/Operation_Entebbe

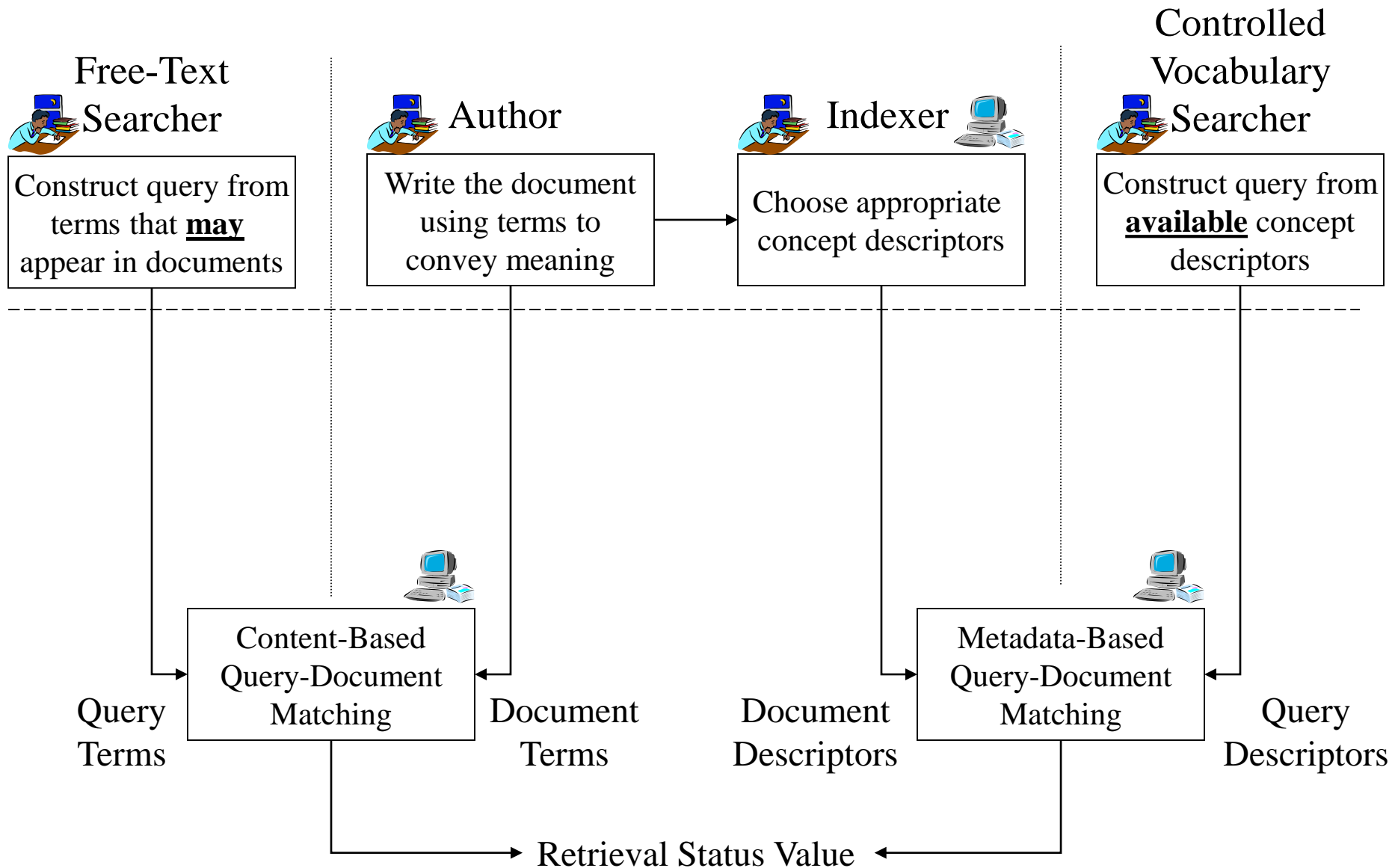
Tonight

- Access points

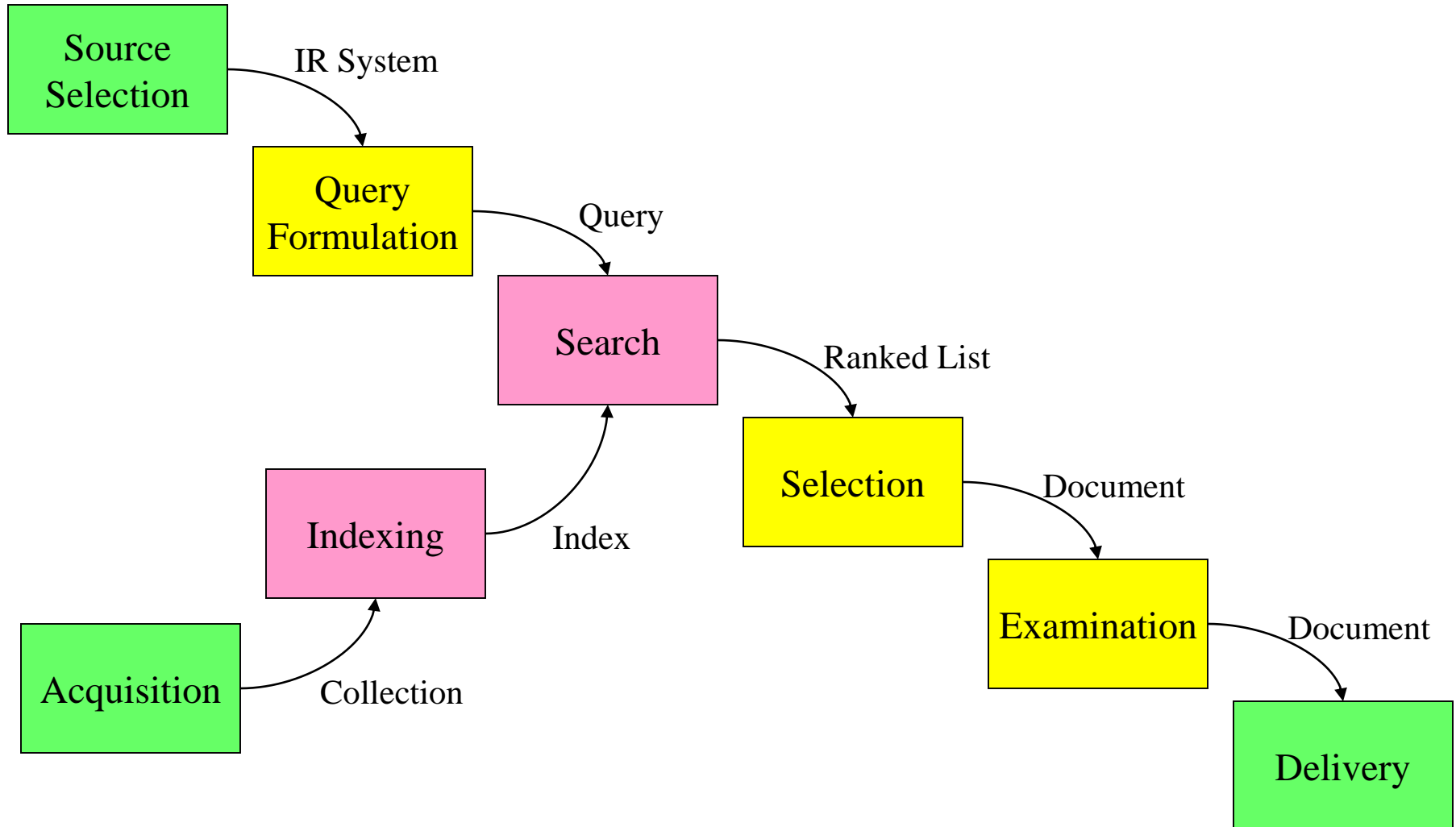
➤ Discovery

- Delivery
- Midterm exam review

Two Ways of Searching



Supporting the Search Process



Online Public Access Catalog (OPAC)

- Known-item search
 - Author, Title
- Topic search
 - Title, subject headings
- Result display
 - Sort by publication date, “relevance,” ...
- Navigation
 - Broader/narrower headings, other editions, ...
- Delivery
 - Call number or (digital content) direct delivery

Tonight

- Access points
- Discovery
- Delivery
- Midterm exam review

Delivery (“Serve It”)

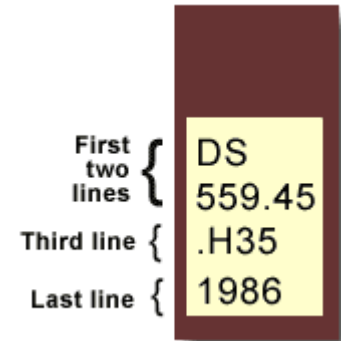
- Assigning a shelf order
- Moving physical materials
- Controlling access to digital materials

Library of Congress Classification

Book title: Uncensored War: The Media and Vietnam

Author: Daniel C. Hallin

Call Number: DS559.46 .H35 1986



The first two lines describe the subject of the book.

DS559.45 = Vietnamese Conflict

D	History
DS1-937	History of Asia
DS520-560.72	Southeast Asia
DS556-559.93	Vietnam. Annam
DS557-559.9	Vietnamese Conflict

The third line often represents the author's last name.

H = Hallin

<i>After other initial consonants</i>							
<i>for the second letter:</i>	a	e	i	o	r	u	y
<i>use number:</i>	3	4	5	6	7	8	9

<i>For expansion</i>							
<i>for the letter:</i>	a-d	e-h	i-l	m-o	p-s	t-v	w-z
<i>use number:</i>	3	4	5	6	7	8	9

The last line represents the date of publication.

The World Is Flat (in LCC)

HM846 .F74 2005

H Social sciences

HM Sociology

HM831 Social change – Causes

HM846 Technological Innovations. Technology.

.F74 Cutter number for Friedman, Thomas

The World Is Flat (in Dewey)

303.4833

300 Social science

300 Social sciences, sociology, & anthropology

303 Social processes

303.4 Social change

303.48 Causes of change

303.483 Development of science and technology

303.4833 Communication (Information technology)

Inter-Library Loan



- Users search “union catalog” to find books
- Remote library “ships” it to local library
 - Often by scanning it, where practical
 - Someone pays for this (local library or user)
- Local library manages circulation
 - Limited access period
 - Some “return” mechanism

Optimum statistical operations with celestial fix data for interplanetary navigation

Author: [David Randolph Scott](#); [Charles Joseph Januska](#); [Richard Errol Willes](#)

Publisher: 1962.

Dissertation: Thesis (M.S.)--Massachusetts Institute of Technology, Dept. of Aeronautics and Astronautics, 1962.


Edition/Format:  Thesis/dissertation : Thesis/dissertation : Manuscript  Archival Material : English

Database: WorldCat


Find a copy in the library

This item isn't held in your library system, but it is held by other WorldCat Libraries.
You may request this item from a WorldCat library through interlibrary loan.

WorldCat

 Find it in libraries globally




Request Item through Interlibrary Loan

 **Worldwide libraries own this item**

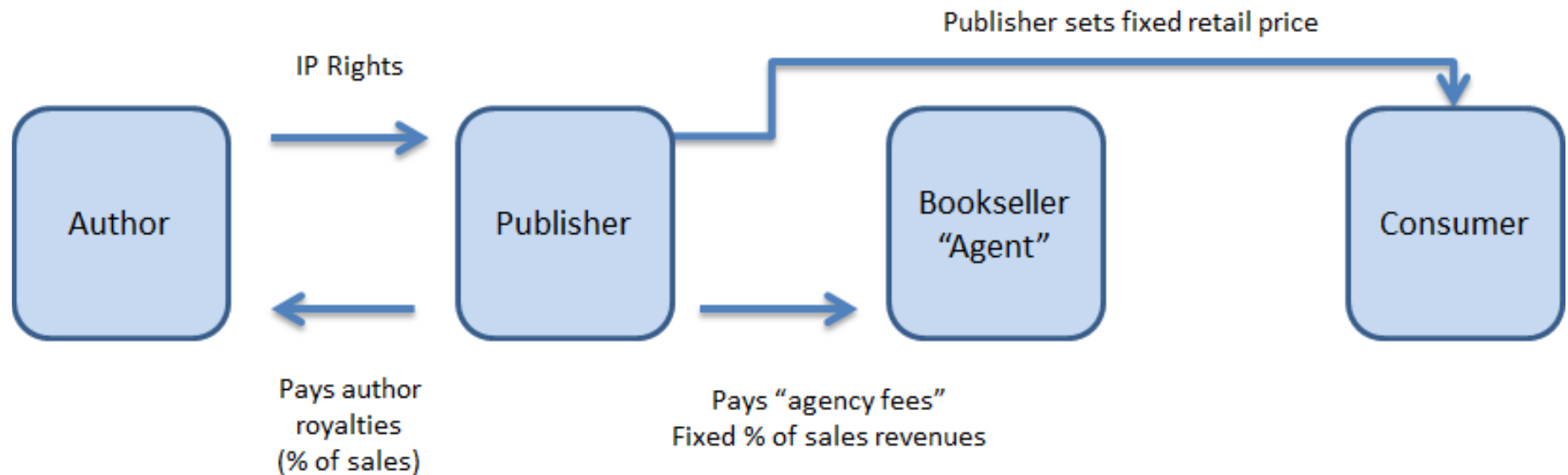
Enter your location:

Displaying libraries 1-1 out of 1

Show libraries holding [just this edition](#)

Library	Held formats	Distance	
1.  MIT Libraries Massachusetts Institute of Technology Libraries Cambridge, MA 02139 United States	  Book	382 miles <input type="button" value="MAP IT"/>	Library info Ask a librarian Add to favorites

E-Book Distribution



Copyright

- Balances two public interests
 - Incentivizing production of new information
 - Through owner's interest in monetizing assets
 - Fostering use of information
 - First sale doctrine
 - Fair use doctrine

First Sale Doctrine

- Owner may transfer access of the owned copy
 - But may not make a copy then transfer the copy
 - This is what permits “lending libraries”
 - Exception: no commercial lending of audio recordings
- Licensing can apply more restrictive rules
 - Establishes a conditional right of access
 - This is what permits limited-

Fair Use Doctrine

- Balance two desirable characteristics
 - Financial incentives to produce content
 - Desirable uses of existing information
- Safe harbor agreement
 - Book chapter, magazine article, picture, ...
- Developed in an era of physical documents
 - Perfect copies/instant delivery alter the balance

Recent Copyright Laws

- Copyright Term Extension Act (CTEA)
 - Ruled constitutional (Jan 2003, Supreme Court)
- Digital Millennium Copyright Act (DMCA)
 - Prohibits circumvention of technical measures
 - Implements WIPO treaty database protection

Digital Rights Management (DRM)

- Goal: protect intellectual property rights
 - Copyright relies on cost and quality of analog copies
- Three interlocking strategies
 - Make it difficult to produce an exact digital copy
 - Encrypt the content and then control description
 - Enforce policies to rebalance costs and benefits

Digital Rights Management

- No standards, so proliferation of one-off solutions
 - Many of which have caused unintended problems
- Unilateral implementation can result in imbalance
 - Establishing balance is a political process
- The “analog hole” is technically intractable
 - Unless interaction is needed

Midterm Exam

- Posted by 5 AM on Tuesday October 28
 - Due at **11 PM on Saturday November 2**
 - 3 Hours, same process as the quiz (email, no talking, ...)
- Comprehensive
 - Nature of information institutions
 - Have it, find it, serve it
- One question will be to create + represent a bibliographic description (w/authority control)
 - One RDA+MARC, MODS or BIBFRAME option
 - One DACS+EAD option

Before You Go!

- On a sheet of paper (no names), answer the following question:

What was the muddiest point in today's class?