

# The Deep Web

## Definition (Deep Web, Hidden Web)

All the content of the Web that is not directly accessible through **hyperlinks**. In particular: HTML forms, Web services.



## Size estimate

- [Bri00] 500 times more content than on the **surface Web!** Dozens of thousands of databases.
- [HPWC07] ~ 400 000 deep Web databases.

# Sources of the Deep Web

## Examples

- *Yellow Pages* and other directories;
- Library catalogs;
- Publication databases;
- Weather services;
- Geolocalization services;
- US Census Bureau data;
- etc.

# Discovering Knowledge from the Deep Web

- Content of the deep Web hidden to classical Web search engines (they just follow links)
- But very valuable and high quality!
- Even services allowing access through the surface Web (e.g., e-commerce) have more semantics when accessed from the deep Web
- How to **benefit** from this information?
- How to do it **automatically**, in an **unsupervised** way?

# Extensional Approach



discovery

Google Scholar BETA **Advanced Scholar Search** [Advanced Search Tips](#) | [About Google Scholar](#)

Find articles with all of the words  10 results

with the exact phrase

with at least one of the words

without the words

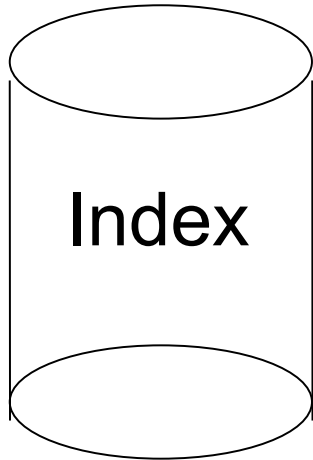
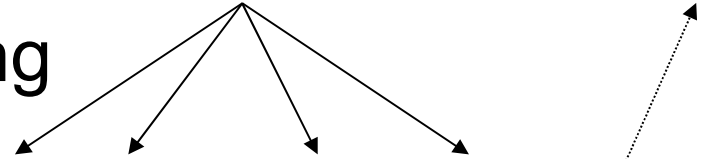
where my words occur anywhere in the article

Author Return articles written by   
e.g., "P.J. Hayes" or McCarthy

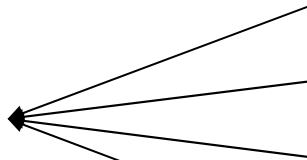
Publication Return articles published in   
e.g., J Biol Chem or Nature

Date Return articles published between  -   
e.g., 1996

siphoning



indexing



Google Scholar BETA [Web](#) [Images](#) [Video](#) [News](#) [Maps](#) [more »](#)

data  Search [Advanced Scholar Search](#) [Scholar Preferences](#) [Scholar Help](#)

**Scholar** All articles - [Recent articles](#) Results 1 - 10 of about 28,900 for monoid. (0.11 seconds)

[On finite monoids](#)  
MP Schützenberger ... 1965) On Finite monoids having Only Trivial Subgroups MP SCHÜTZENBERGER An alternative definition is given for a family of subsets of a free monoid that has ...  
[Cited by 287](#) - [Related articles](#) - [Web Search](#) - [All 3 versions](#)

[System identification: the](#)  
L Ljung - 1986 - Prentice-Hall, Inc. U  
[Cited by 7815](#) - [Related articles](#) - [Web Search](#)

[Nonlinear systems](#)  
HK Khalil, JW Grizzle - 1996 - dev.pr  
Nonlinear Systems. Second Edition.  
distinct parts: Part I covers nonlinear  
[Cited by 6462](#) - [Related articles](#) - [Web Search](#)

[Commutative, residuated l-monoids](#)  
U Höhle - Nonclassical logics and their applications to fuzzy subsets  
[Cited by 206](#) - [Related articles](#) - [Web Search](#)

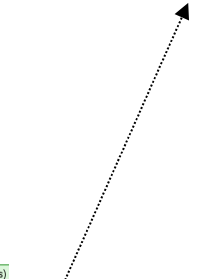
[Finite monoids and the fine structure of NC 1](#)  
DAM Barrington, D Thérien - Journal of the ACM (JACM), 1988 - portal.acm.org  
... 2. Background and Definitions A finite monoid is a finite set with an associative  
binary operation and an identity element. ... aperiodic monoid ...  
[Cited by 139](#) - [Related articles](#) - [Web Search](#) - [All 7 versions](#)

[Systems thinking system](#)  
P Checkland - 1999 - orton.ca.ie.ac.  
... Autor : Checkland, Peter, Trullo  
P. impronta : New York : Wiley 330  
[Cited by 3693](#) - [Related articles](#) - [Web Search](#)

[Rational sets in commutative monoids](#)  
S Eilenberg, MP Schützenberger - J. Algebra, 1969 - www.igm.univ-mlv.fr  
... SchÜTZENBERGER Faculté des Sciences de Paris, Paris, France Communicated by Saunders  
MacLane Received August 3, 1969 1. Rational Sets Let M be a monoid, ie a ...  
[Cited by 131](#) - [Related articles](#) - [Web Search](#) - [All 2 versions](#)

[Word problems and a homological finiteness condition for monoids](#)

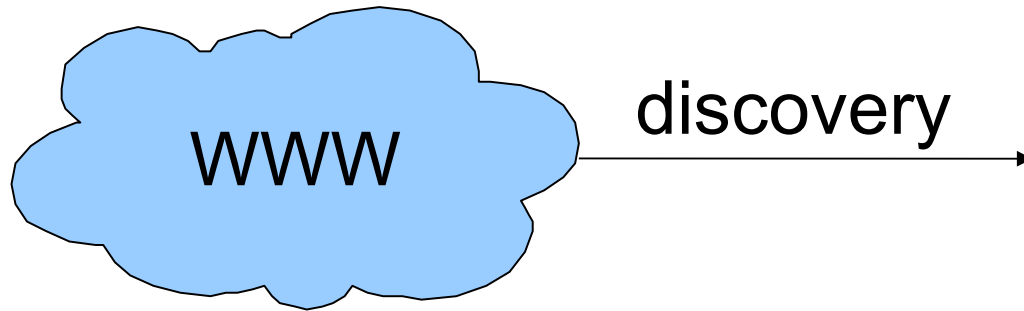
bootstrap



# Notes on the Extensional Approach

- Main issues:
  - Discovering services
  - Choosing appropriate data to submit forms
  - Use of data found in result pages to bootstrap the siphoning process
  - Ensure good coverage of the database
- Approach **favored by Google** [MHC+06], used in production
- Not always feasible (huge load on Web servers)

# Intensional Approach



Google Scholar BETA Advanced Scholar Search [Advanced Search Tips](#) | [About Google Scholar](#)

Find articles with all of the words  10 results

with the exact phrase

with at least one of the words

without the words

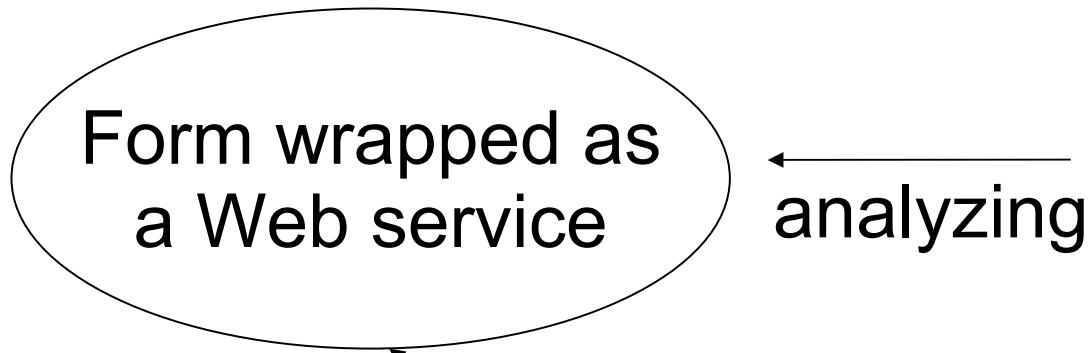
where my words occur anywhere in the article

Author Return articles written by   
e.g., "PJ Hayes" or McCarthy

Publication Return articles published in   
e.g., J Biol Chem or Nature

Date Return articles published between  -   
e.g., 1996

probing



Google Scholar BETA [Web](#) [Images](#) [Video](#) [News](#) [Maps](#) [more »](#)   [Advanced Scholar Search](#) [Scholar Preferences](#) [Scholar Help](#)

Scholar All articles - **Recent articles** Results 1 - 10 of about 91,400,000 for data [definition] (0.14 seconds)

1 Fisher R. The use of multiple measurements in taxonomic problems. [JE Psychol](#), AOO Generals, SA Genet, M Biol, BMC. ... Ann of Eugenics, 1936 - [biomedcentral.com](#)

... Cuhane A, Perriere G, Considine E, Cotter T, Higgins D. Between-group analysis of microarray data. ... [Comput Stat Data Anal](#) 2004, 46:407-425. ... [Cited by 3852](#) - [Related articles](#) - [Cached](#) - [Web Search](#)

[The protein kinase encoded by the Akt proto-oncogene is a target of the PDGF-activated...](#)

... Franke, SI Yang, TO Chan, K DATA, A Kazlauskas, DK. ... [Cell](#)(Cambridge), 1995 - [cat.inist.fr](#)

TF FRANKE, SUNG-IL YANG, TO CHAN, K DATA, A KAZLAUSKAS, DK MORRISON, DR KAPLAN, PN TSICHLIS [Cell](#)(Cambridge) 81:55, 727-736, Cell Press, 1995. [Cited by 1409](#) - [Related articles](#) - [Web Search](#) - [BI Direct](#) - [All 6 versions](#)

[RAG-2-deficient mice lack mature lymphocytes owing to inability to initiate V\(D\)J rearrangement](#)

FB Policies, DP Policy, I Subscribers, I ... - [Cell](#), 1992 - [cell.com](#)

... Both genetic and biochemical data point toward a physiological role for this complex as the elusive hairpin-opening activity in V(D)J recombination. ... [Cited by 1316](#) - [Related articles](#) - [Cached](#) - [Web Search](#) - [All 4 versions](#)

[Random data analysis and measurement procedures](#)

JS Bendat, AG Piersol - [Measurement Science and Technology](#), 2000 - [iop.org](#)

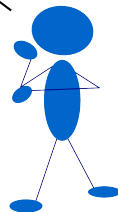
BOOK REVIEW. Random Data Analysis and Measurement Procedures. ... Chapter ten deals with data aquisition and processing, including data qualification. ... [Cited by 3944](#) - [Related articles](#) - [Web Search](#) - [SUDOC Catalogue](#) - [All 10 versions](#)

[Data mining: practical machine learning tools and techniques with Java implementations](#) - [waikato.ac.nz](#) (pdf)

IH Witten, E Frank - [ACM SIGMOD Record](#), 2002 - [portal.acm.org](#)

[Data Mining: Practical Machine Learning Tools and ...](#) Written and Frank's textbook was

query



# Notes on the Intensional Approach

- More **ambitious** [CHZ05, SMM+08]
- Main issues:
  - Discovering services
  - Understanding the structure and semantics of a form
  - Understanding the structure and semantics of result pages (wrapper induction)
  - Semantic analysis of the service as a whole
- No significant load imposed on Web servers

# Discovering deep Web forms

- Crawling the Web and selecting forms
- But **not all forms!**
  - Hotel reservation
  - Mailing list management
  - Search within a Web site
- **Heuristics:** prefer GET to POST, no password, no credit card number, more than one field, etc.
- Given domain of interest: use **focused crawling** to restrict to this domain



# Web forms

Authors	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>
Title	<input type="text"/>		Year <input type="text"/>	Page <input type="text"/>
Conference	<input type="text"/>	ID <input type="text"/>		
Journal	<input type="text"/>	Volume <input type="text"/>	Number <input type="text"/>	
<input type="button" value="Search"/>	<input type="button" value="Reset"/>	Maximum of <input type="text" value="100"/> matches		

- **Simplest case:** associate each form field with some **domain concept**
- **Assumption:** fields independent from each other (not always true!), can be queried with words that are part of a **domain instance**

# Structural analysis of a form (1/2)

- Build a **context** for each field:
  - label tag;
  - id and name attributes;
  - text immediately before the field.
- Remove **stop words, stem**
- **Match** this context with concept names or concept ontology
- Obtain in this way **candidate annotations**

# Structural analysis of a form (2/2)

For each field annotated with concept  $c$ :

- Probe the field with nonsense word to get an **error page**
- **Probe** the field with instances of concept  $c$
- Compare pages obtained by probing with the error page (e.g., clustering along the DOM tree structure of the pages), to distinguish error pages and **result pages**
- **Confirm** the annotation if enough result pages are obtained

# Bootstrapping the siphoning

- Siphoning (or probing) a deep Web database requires many relevant data to submit the form with
- **Idea:** use **most frequent words** in the content of the result pages
- Allows **bootstrapping** the siphoning with just a few words!

# Inducing wrappers from result pages

Pages resulting from a given form submission:

- share the **same structure**
- set of **records** with fields
- **unknown** presentation!

CiteSeer Find: remi gilleron Documents  
Citations

Searching for **PHRASE remi gilleron**.

Restrict to: [Header](#) [Title](#) Order by: [Expected citations](#) [Hubs](#) [Usage](#) [Date](#) Try: [Google \(CiteSeer\)](#) [Google \(Web\)](#) [Yahoo!](#) [MSN](#) [CSB](#) [DBLP](#)

7 documents found. Order: number of citations.

[PAC Learning under Helpful Distributions - Denis, Gilleron \(1997\)](#) (Correct) (10 citations)

Helpful Distributions y Francois Denis, R'emi **Gilleron** LJFL, URA 369 CNRS, Universit'e de Lille 1 59655  
1 59655 Villeneuve d'Ascq FRANCE e-mail: denis.gilleron@lifl.fr Abstract A PAC model under helpful  
on Algorithmic Learning Theory ALT'97 (Denis and **Gilleron**, 1997)Introduction It seems that many  
ftp.grappa.u

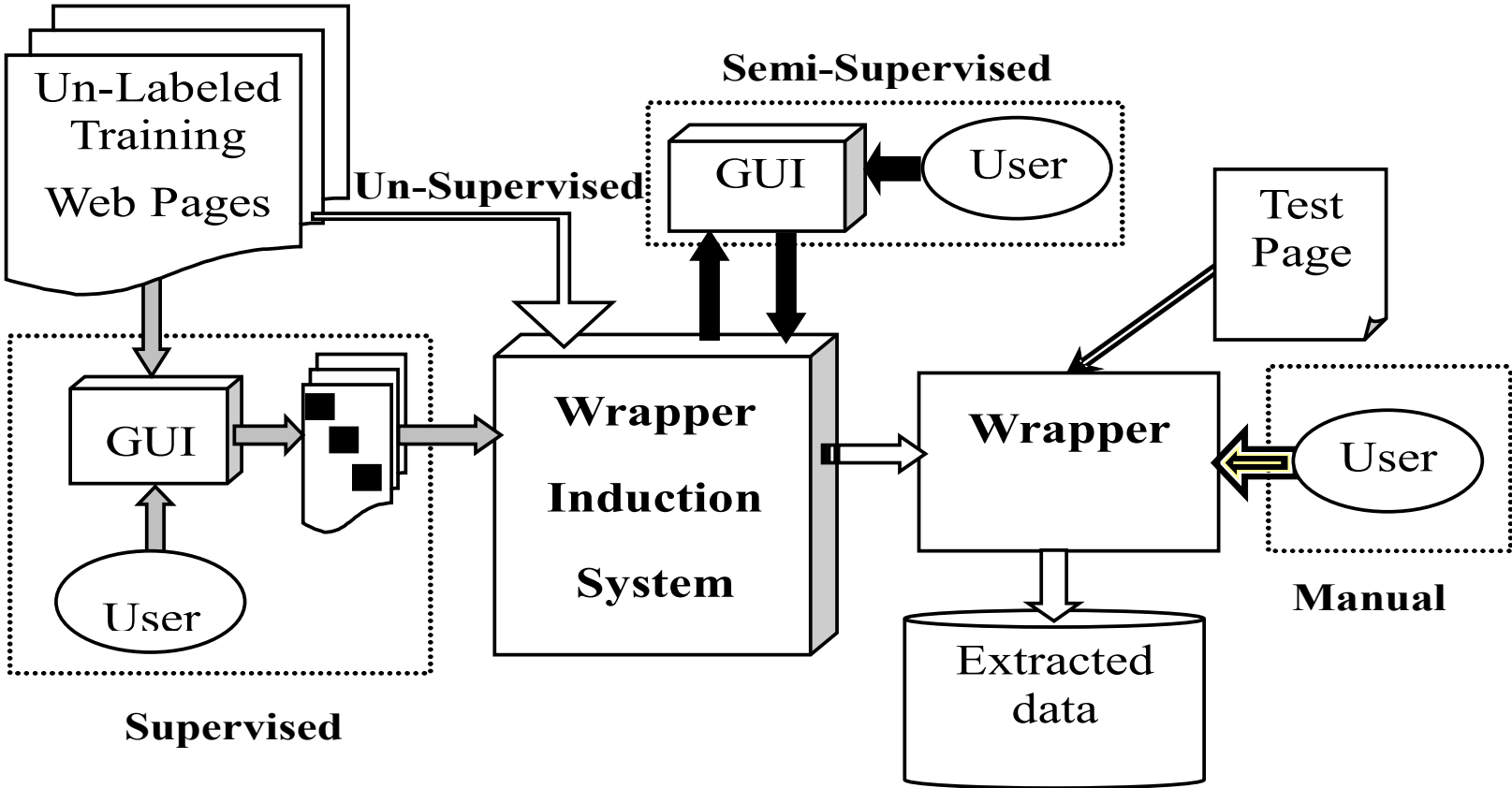
Sort by T-Meter	Sort by Title	Sort by Year
81%	<b>Grindhouse</b> Director Screenwriter Producer	2007
- N/A	<b>Death Proof</b> Director	2007
59%	<b>Hostel</b> Executive Producer	2006
- N/A	<b>Reservoir Dogs/Bad Lieutenant</b> Director	2006
- N/A	<b>Inglorious Bastards</b> Director	2006
97%	<b>Double Dare</b> Featured	2005
78%	<b>Sin City</b> Additional Directing	2005
29%	<b>The Muppets: Wizard Of Oz</b> Star	2005
0%	<b>Daltry Calhoun</b> Executive Producer	2005
85%	<b>Kill Bill Vol. 2</b> Director Screenwriter	2004
100%	<b>Z Channel: A Magnificent Obsession</b> Featured	2004
85%	<b>Kill Bill Vol. 1</b> Director Screenwriter Producer	2003

## Goal

Building **wrappers** for a given kind of result pages, in a fully automatic way.

# Information extraction systems

[CKGS06]



# Unsupervised Wrapper Induction

- Use the (repetitive) structure of the result pages to infer a **wrapper** for all pages of this type
- Possibly: use in parallel with **annotation** by recognized concept instances to learn with **both the structure and the content**



# Some perspectives

- Dealing with **complex forms** (fields allowing Boolean operators, dependencies between fields, etc.)
- **Static analysis** of JavaScript code to determine which fields of a form are required, etc.
- A lot of this is also applicable to **Web 2.0/AJAX** applications



# References

- [Bri00] BrightPlanet. **The deep Web: Surfacing hidden value.** White paper, July 2000.
- [CHZ05] K. C.-C. Chang, B. He, and Z. Zhang. **Towards large scale integration: Building a metaquerier over databases on the Web.** In *Proc. CIDR*, Asilomar, USA, Jan. 2005.
- [CKGS06] C.-H. Chang, M. Kayed, M. R. Girgis, and K. F. Shaalan. **A survey of Web information extraction systems.** *IEEE Transactions on Knowledge and Data Engineering*, 18(10):1411-1428, Oct. 2006.
- [CMM01] V. Crescenzi, G. Mecca, and P. Merialdo. **Roadrunner: Towards automatic data extraction from large Web sites.** In *Proc. VLDB*, Roma, Italy, Sep. 2001.
- [HPWC07] B. He, M. Patel, Z. Zhang, and K. C.-C. Chang. **Accessing the deep Web: A survey.** *Communications of the ACM*, 50(2):94–101 May 2007.
- [MHC+06] J. Madhavan, A. Y. Halevy, S. Cohen, X. Dong, S. R. Jeffery, D. Ko, and C. Yu. **Structured data meets the Web: A few observations.** *IEEE Data Engineering Bulletin*, 29(4):19–26, Dec. 2006.
- [SMM+08] P. Senellart, A. Mittal, D. Muschick, R. Gilleron et M. Tommasi, **Automatic Wrapper Induction from Hidden-Web Sources with Domain Knowledge.** In *Proc. WIDM*, Napa, USA, Oct. 2008.