BINGO! and Daffodil: Personalized Exploration of Digital Libraries and Web Sources

Martin Theobald
Max-Planck-Institut für Informatik

Claus-Peter Klas
Universität Duisburg-Essen

INFORMATIK

Overview

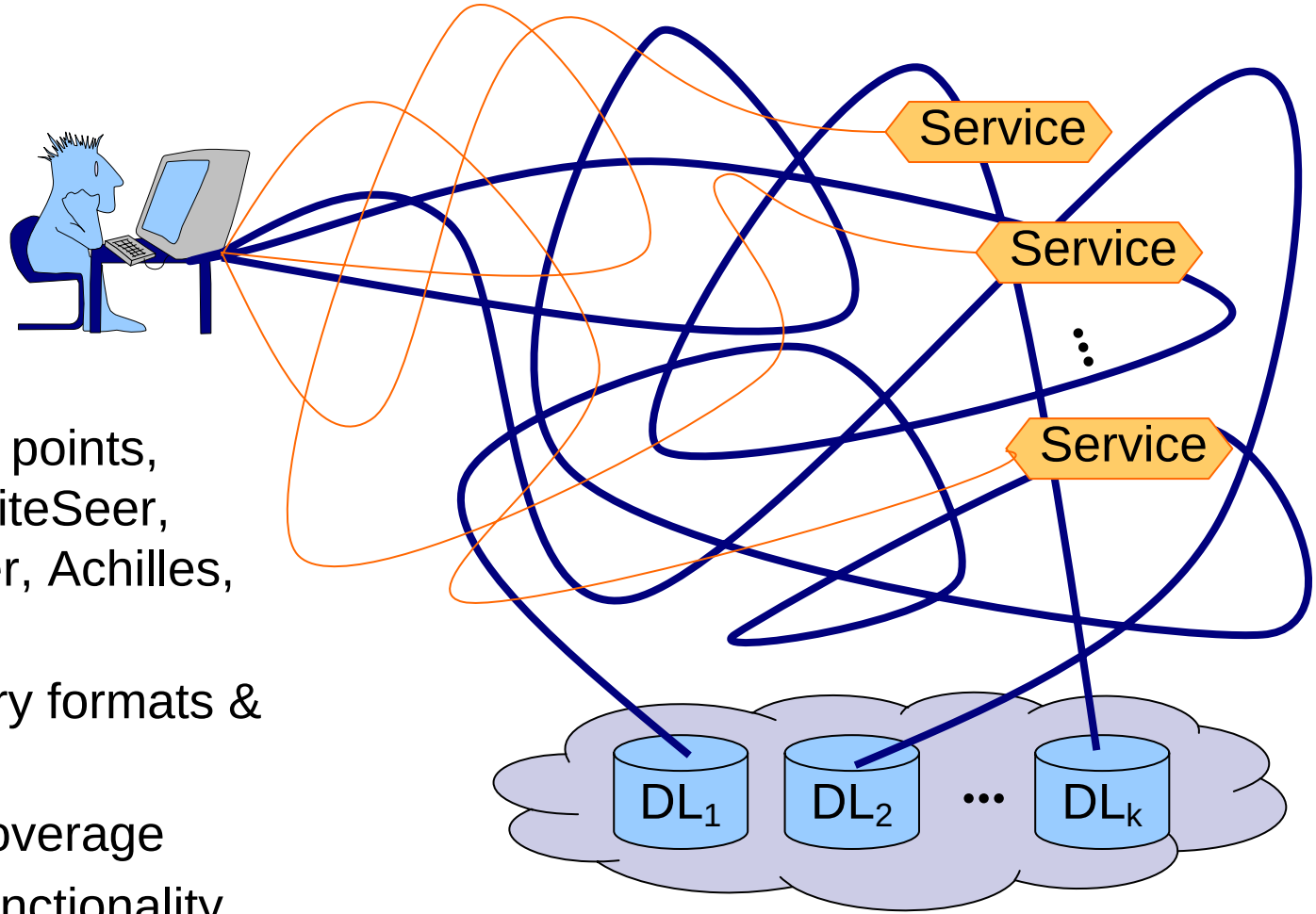
- Challenges in Web Recommendations for Digital Library Users
- DAFFODIL
- BINGO!
- Meta Data Processing
- System Integration via Web Services
- Experiments

1. Recommendations for Digital Library Users

- Current recommendations systems are based on
  - User collaboration & mining of user behaviour
  - Pre-computed document similarities (clustering, cosine measure, co-citations, etc.) limited to a single DL index

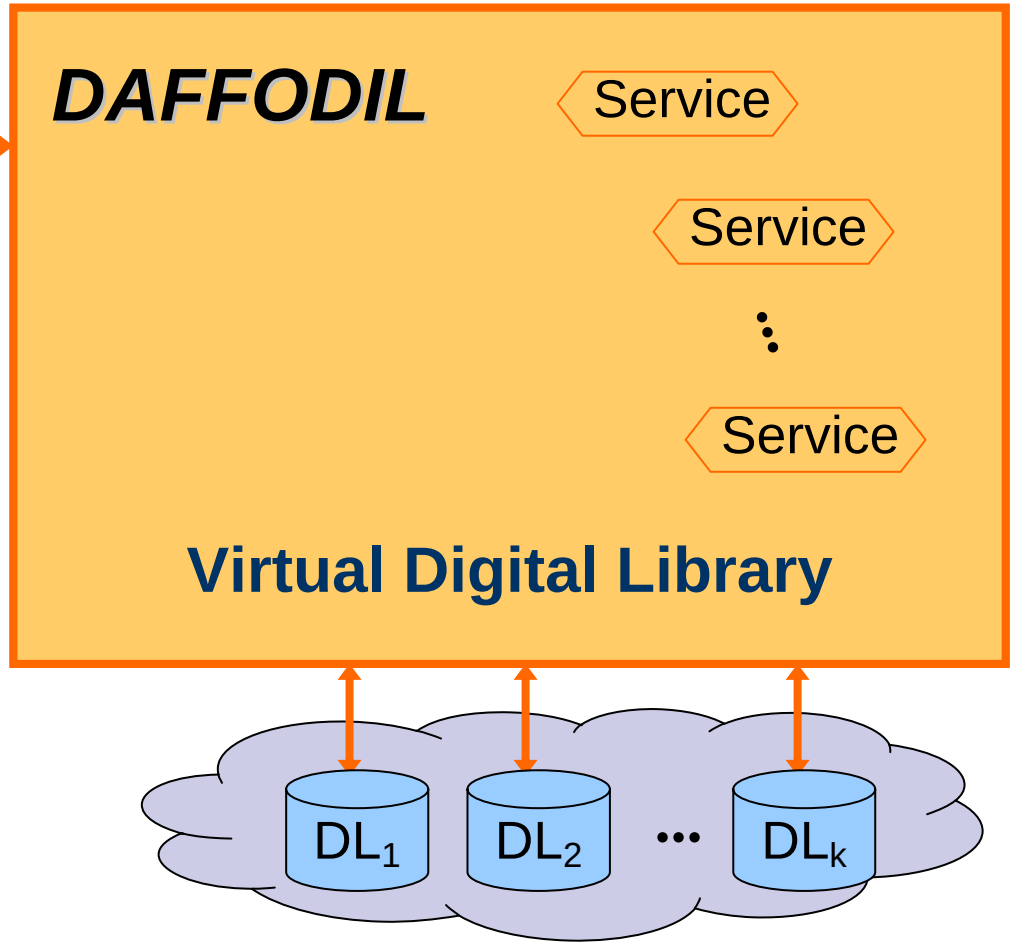
- Challenges in providing Web recommendations
  - Well organized and structured Digital Library world vs. heterogeneous and chaotic Web
  - Unified view on related publications *and* topic-specific Web documents (e.g., authors' homepages)
  - "Virtual links" from meta data through additional Deep Web search
    - Topic-specific hub pages
    - Full text sources for training
  - Filtering, user feedback, and iterative topic refinement

2. DAFFODIL: Motivation



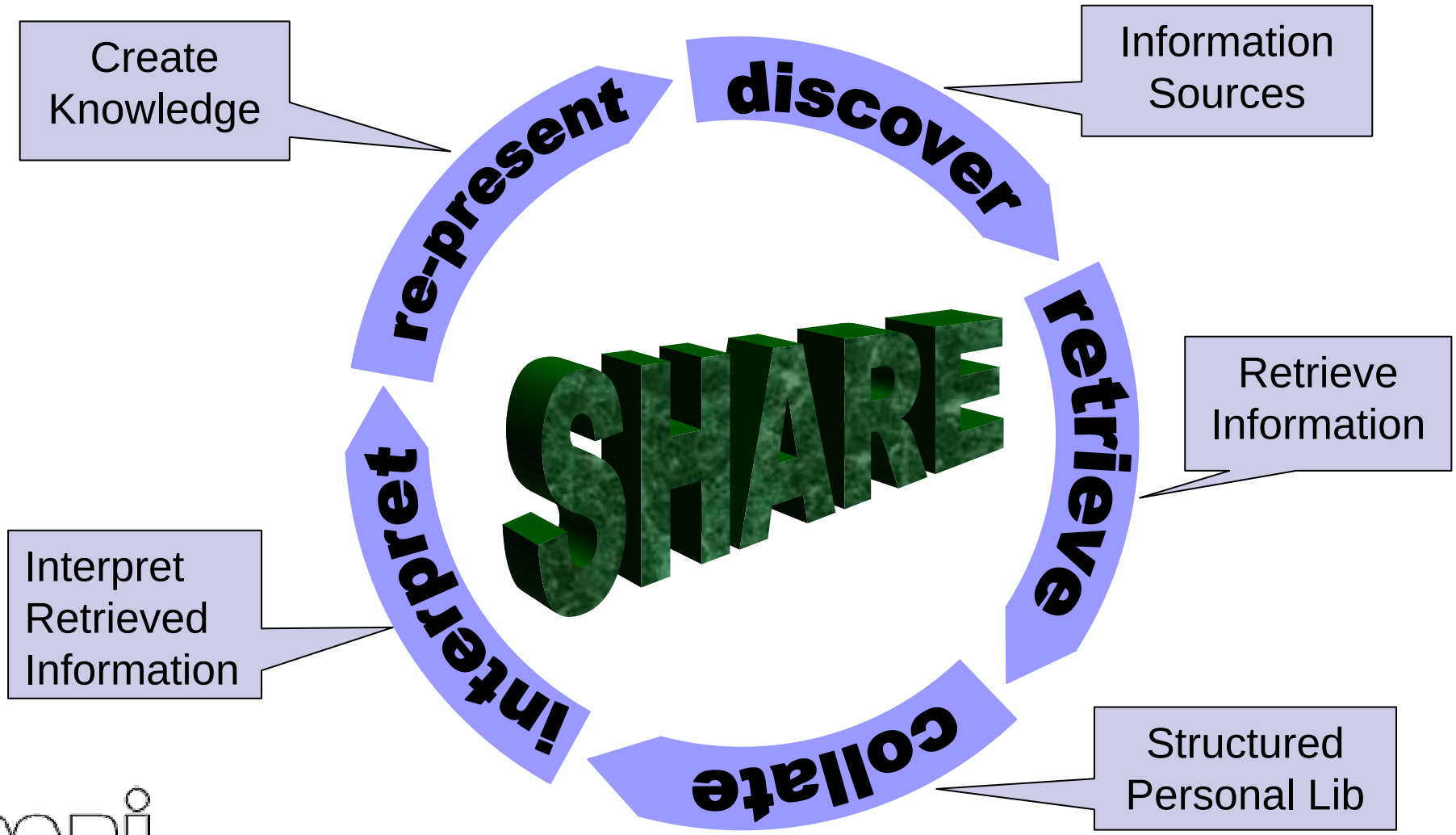
- Many access points, e.g., DBLP, CiteSeer, ACM, Springer, Achilles, HCIBIB
- Different query formats & services
- Insufficient coverage
- Insufficient functionality

2.1. Combination of Services & Information



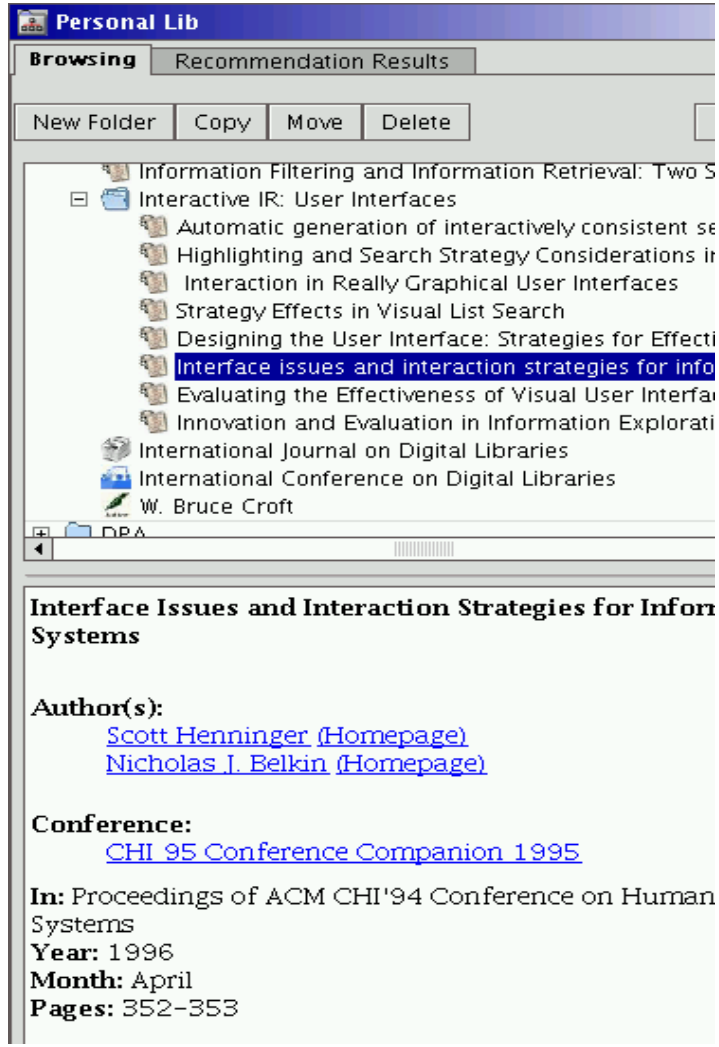
- One access point
- Integration of DLs and individual services
- Meta data extraction
- Synergy effects
  - Strategic information search
  - Information organization in personal folders

2.2. Lifecycle of Scientific Search as an Interactive and Iterative Process



2.3. Personal Library

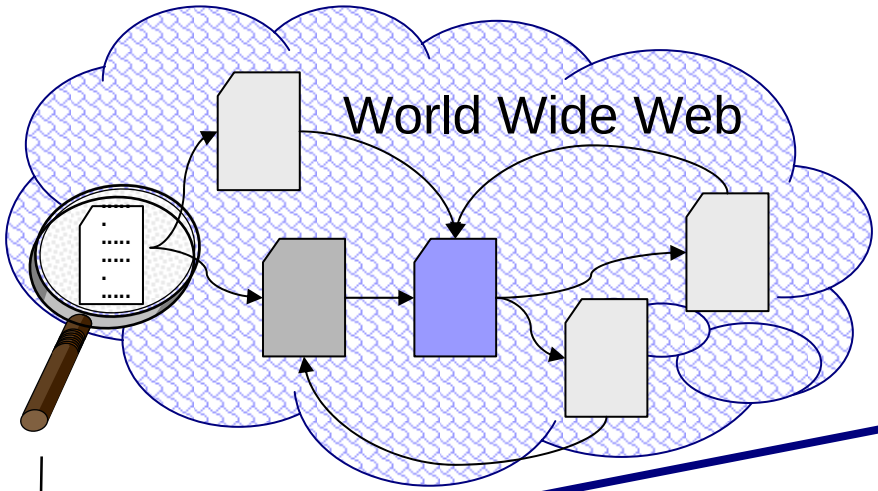
- Publications organized as structured meta data
- Digital Library Objects (DLOs)
  - Authors, title
  - Web links
  - Journals
  - Conferences
  - Keywords, abstracts
  - Queries
- Meta data export in XML or BibTex
- Basis for information exchange with BINGO!



3. BINGO!

- Focused crawler
  - Combination of traditional crawler with classifier and link analyzer for topic distillation
  - Crawls mainly within the densely connected neighbourhood graph for the topics of interest
  - Filters out irrelevant documents directly at the crawler frontier
- Information portal generation
  - Initialized by a set of training documents (e.g., *bookmarks*)
  - Extends and maintains a user-specific topic-directory (Yahoo!)
- Expert queries
  - Initialized by a single-topic keyword query
  - Exploits existing index or external search engine for retrieving initial hubs → training & seed
  - Improves recall

3.1. Focused Crawler Architecture



Training documents

Iterative (Re-)Training, Feedback

Crawler

Document analyzer

Feature selection

SVM classifier

HITS link-analyzer

URL queue

Feature vectors, TF*IDF

Topic-specific features

Taxonomy index

Hubs & Authorities

4. Meta Data Processing

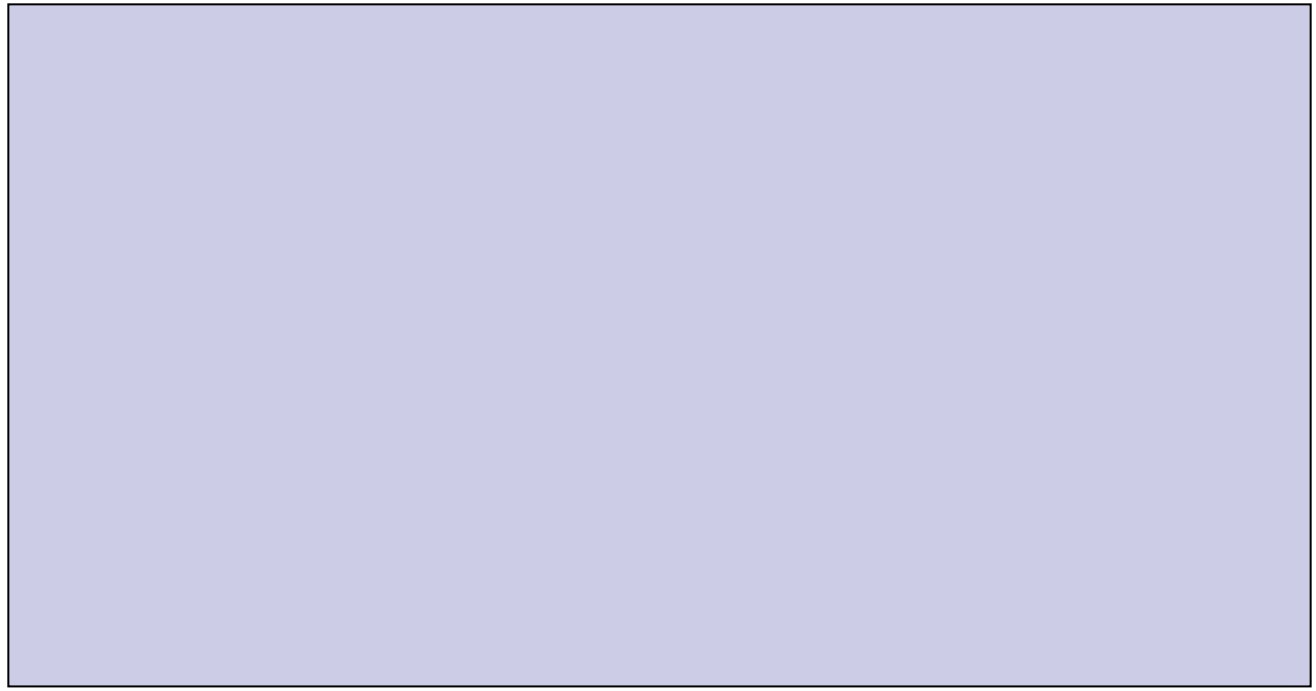
- DLOs ***do provide***
  - Comprehensive meta data describing the publication
- DLOs ***may provide***
  - Links to full text sources (e.g., PS, PDF or DOC - files)
  - Links to homepages, DL portals, or other Web references
- Focused crawler ***requires***
  - Full text documents for feature extraction and training
  - Topic-specific hubs as start pages of focused crawl

- **Solution:** Additional Deep Web queries using DLO contents

- Web Service Framework (WSF)
  - Replace form submission method of the browser (HTTP post) with new WS
    - Detection of attribute labels/names next to form fields
    - Form types determine attribute types (e.g., *String, Date,* etc.)
  - Generation of WSDL description for WS parameters
  - Automatic deployment under local Tomcat Web Server

Advanced search form of the DBLP Trier

(HTML <*form*> and <*input*> tags)



WSF

WSDL description registered at local UDDI

```
- <message name="WSF_Form0_InputMessage">
    <documentation>http://www.informatik.uni-trier.de/ley/dbbin/dblpquery.cgi</documentation>
    <part name="Authors[author]WSF_Form0Text0_Part" type="xsd:string" />
    <part name="[author]WSF_Form0Text1_Part" type="xsd:string" />
    <part name="[author]WSF_Form0Text2_Part" type="xsd:string" />
    <part name="[author]WSF_Form0Text3_Part" type="xsd:string" />
    <part name="Title[title]WSF_Form0Text4_Part" type="xsd:string" />
    <part name="Year[year]WSF_Form0Text5_Part" type="xsd:string" />
    <part name="Page[pages]WSF_Form0Text6_Part" type="xsd:string" />
    <part name="Conference[booktitle]WSF_Form0Text7_Part" type="xsd:string" />
    <part name="ID[cite]WSF_Form0Text8_Part" type="xsd:string" />
    <part name="Journal[journal]WSF_Form0Text9_Part" type="xsd:string" />
    <part name="Volume[volume]WSF_Form0Text10_Part" type="xsd:string" />
    <part name="Number[number]WSF_Form0Text11_Part" type="xsd:string" />
    <part name="Maximum of[return]WSF_Form0Select0_Part"
      type="typens:WSF_Form0Select0_Enum" />
    <part name="WSF_SubmitWSF_Form0_Part" type="typens:WSF_SubmitWSF_Form0_Enum" />
  </message>
```

4.1. Ontology-Enabled Form Matching & Portal Selection

- Ontology service
  - Incorporates WordNet, Cyc, and personalized domain ontologies
  - Mapping of attribute names to semantic concepts
  - Quantified relationships between concepts
  - Approximate matches between similar attributes in a DLO and WSDL
- Portal selection
  - Aggregation of attribute similarities between a DLO and WSDL
  - Ranking for the top matching portals to each meta data query
- Unified handling of heterogeneous portals
  - Pre-computed (static) mapping on schema level only
  - Automatic form completion and submission

- DLO extracted from Achilles

```
<metadata docid="d738000632"
  folder="/webdavroot/Uni+Saarbruecken/Workflow">
- <authors>
  - <author id="hung#patrickck">
      <firstname>Patrick C.K.</firstname>
      <lastname>Hung</lastname>
    </author>
  - <author id="karlapalem#kamalakar">
      <firstname>Kamalakar</firstname>
      <lastname>Karlapalem</lastname>
    </author>
  </authors>
  <booktitle>Australasian Information Security
    Workshop</booktitle>
  <pages>33--41</pages>
  <publisher>ACS</publisher>
  <publisheraddress>Adelaide, Australia</publisheraddress>
  <series>Conferences in Research and Practice in
    Information Technology</series>
  <title>A Secure Workflow Model</title>
  <url>Achilles</url>
  <volume>21</volume>
  <year>2003</year>
</metadata>
```

Hypernym 0.8

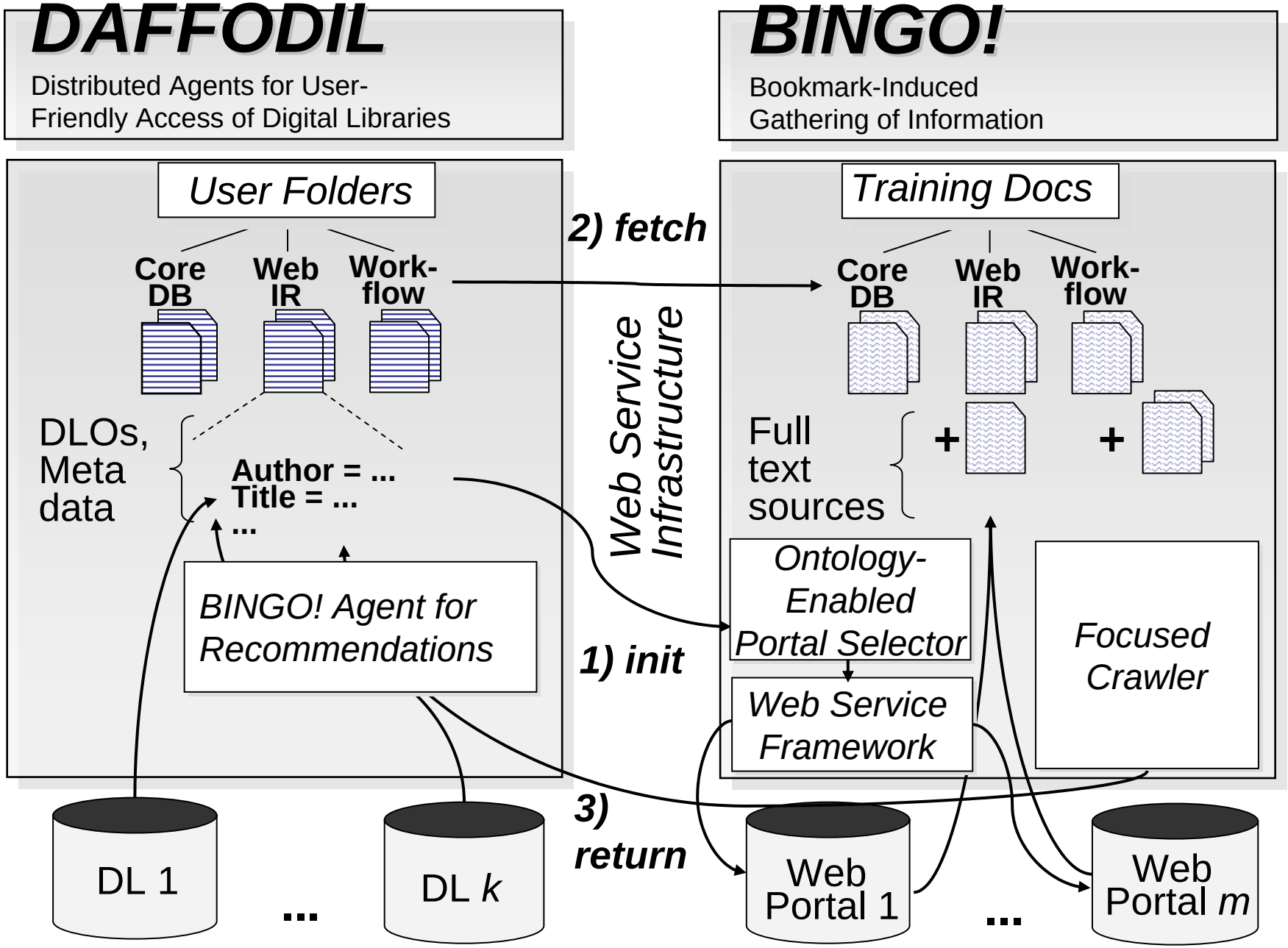
Synonym 1.0

- WSDL for DBLP Trier

```
<message name="WSF_Form0_InputMessage">
  <documentation>http://www.informatik.uni-
    trier.de/ley/dbbin/dblpquery.cgi</documentation
  <part name="Authors[author]WSF_Form0Text0_Par
    type="xsd:string" />
  <part name="[author]WSF_Form0Text1_Part"
    type="xsd:string" />
  <part name="[author]WSF_Form0Text2_Part"
    type="xsd:string" />
  <part name="[author]WSF_Form0Text3_Part"
    type="xsd:string" />
  <part name="Title[title]WSF_Form0Text4_Part"
    type="xsd:string" />
  <part name="Year[year]WSF_Form0Text5_Part"
    type="xsd:string" />
  <part name="Page[pages]WSF_Form0Text6_Part"
    type="xsd:string" />
  <part name="Conference[booktitle]WSF_Form0Text
    type="xsd:string" />
  <part name="ID[cite]WSF_Form0Text8_Part"
    type="xsd:string" />
  <part name="Journal[journal]WSF_Form0Text9_Par
    type="xsd:string" />
  <part name="Volume[volume]WSF_Form0Text10_P
    type="xsd:string" />
```

5. System Integration

- Web Service Infrastructure
  - Two standalone Tomcat Web Servers for WS routing
    - Daffodil at Universität Duisburg
    - Bingo! + WSF at MPII Saarbrücken
- Asynchronous coupling
  - Distributed multi-user requests
  - Queuing of recommendation requests at Daffodil (FIFO)
  - Integration of Bingo! into Daffodil's agent-based structure
- Simple communication protocol (via RPCs)
  - 1) Daffodil: ***Init*** request with basic crawler parameters
  - 2) Bingo!: ***Fetch*** folders along with user id & timestamp
  - 3) Bingo!: ***Return*** top recommendations


DAFFODIL
Distributed Agents for User-Friendly Access of Digital Libraries
BINGO!
Bookmark-Induced Gathering of Information
User Folders
Core DB
Web IR
Work-flow
DLOs, Meta data
Author = ...
Title = ...
...
BINGO! Agent for Recommendations
2) fetch
Web Service Infrastructure
1) init
3) return
Training Docs
Core DB
Web IR
Work-flow
Full text sources
+
+
Ontology-Enabled Portal Selector
Web Service Framework
Focused Crawler
DL 1
...
DL k
Web Portal 1
...
Web Portal m

6. Experiments: User Folder for "Workflow"

No.	Author	Title	Year
W1.	M. Gillmann, J. Weissenfels, A. Kraiss, G. Weikum	Performance Assessment and Configuration of Enterprise-Wide Workflow Management Systems	1999
W2.	M. Gillmann, G. Weikum	Benchmarking and Configuration of Workflow Management Systems	2000
W3.	M. Gillmann, J. Weissenfels, A. Kraiss, G. Weikum	Performance and Availability Assessment for the Configuration of Distributed Workflow Management Systems	2000
W4.	J. Weissenfels, M. Gillmann G. Shegalov, W. Wonner	The Mentor-Lite Prototype: A Light-Weight Workflow Management System	2000
W5.	A. Alessandra, G. Michaelis	A Light Workflow Management System Using Simple Process Models	2000
W6.	M. Gillmann, G.Shegalov, G. Weikum,	XML-enabled workflow management for e-services across heterogeneous platforms	2001
W7.	H. Zhuge, T. Cheung	A timed workflow process model	2001
W8.	P.Grefen	Transactional Workflows or Workflow Transactions ?	2002
W9.	B. Reiner, E. Siegel, J.A. Carrino	Workflow optimization: current trends and future directions	2002
W10.	P. Hung, K. Karlapalem	A Secure Workflow Model	2003
W11.	B. Kiepuszewski, A.H.M. ter Hofstede	Fundamentals of Control Flow in Workflows	2003
W12.	C. Seggewies, et al	Soarian–workflow management applied for health care	2003

- WSF: DBLP Trier, HCIBIB, Achilles, NCSTRL, Springer, Google adv.

6.1. First Iteration & Feedback

No.	URL	BINGO! Rating	Key Resource	Feed-back
1.	http://www.e-workflow.org	3.86	0	⊖
2.	http://www.waria.com	3.76	0	⊖
3.	http://www.wfmc.org/standards/docs/Glossay_German.pdf	3.32	0	
4.	http://www.wfmc.org/information/handbook2003.htm	3.05	0	
5.	http://www.wfmc.org/pr/CDROM_2001.pdf	2.95	0	⊖
6.	http://www.wfmc.org/information/info.htm	2.91	0	⊖
7.	http://www.usc.edu/dept/ATRIUM/Papers/PDI.pdf	2.89	1	⊕
8.	http://www.wfmc.org/standards/docs/TC-1011_term_glossary_v3.pdf	2.88	0	
9.	http://www.wfmc.org/pr/Workflow_Handbook_2001.pdf	2.88	0	
10.	http://www-dbs.cs.uni-sb.de/~gillmann/Publications/Demo-ICDE.pdf	2.86	1	⊕
11.	http://www.wfmc.org/information/Workflow-An_Introduction.pdf	2.75	0	
12.	http://www.wfmc.org/standards/docs/Stds_diagram.pdf	2.74	1	⊕
13.	http://www-dbs.cs.uni-sb.de/~gillmann/Publications/Demo-SIGMOD.pdf	2.68	1	⊕
14.	http://www.informatik.uni-stuttgart.de/ipvr/as/ projekte/poliflow/IPVR/adaptive_workflows.html	2.68	1	
15.	http://www.wfmc.org/information/awards.htm	2.68	0	⊖
16.	http://www.wfmc.org/standards/docs/tc003v11.pdf	2.63	0	
17.	http://www.dfki.uni-kl.de/~aabecker/Freiburg/Final/Wargitsch/Wargitsch.pdf	2.63	1	
18.	http://www.dfki.uni-kl.de/~elst/papers/kmdap_frodo_eval_submitted.pdf	2.55	1	⊕
19.	http://www.csd.uch.gr/~hy565/Papers/overview_workflow_management.pdf	2.48	1	
20.	http://www-dbs.cs.uni-sb.de/projekte/workflow_de.htm	2.37	1	
		Precision	**0.45**	

6.2. Second Iteration

No.	URL	BINGO! Rating	Key Resource
1.	http://awareness.ics.uci.edu/~rsilvafi/wonder/ISADS99/isads99.pdf	1.95	1
2.	http://www.ai.sri.com/~swim/resources/SOA-workflow.html	1.80	1
3.	http://www-dbs.cs.uni-sb.de/~gillmann/Publications/ConfigTool-EDBT.pdf	1.79	1
4.	http://awareness.ics.uci.edu/~rsilvafi/wonder/SBRC99/sbrc99.pdf	1.79	1
5.	http://www.informatik.uni-stuttgart.de/ipvr/as/projekte/apricots/atma96.pdf	1.78	1
6.	http://www.research.ibm.com/journal/sj/361/leymann.html	1.72	1
7.	http://www.cs.colorado.edu/~skip/proclets.pdf	1.70	1
8.	http://www.ifi.unizh.ch/dbtg/Projects/TRAMs/trams.html	1.64	1
9.	http://www.ifi.unizh.ch/dbtg/Projects/EVE/eve.html	1.62	1
10.	http://www.almaden.ibm.com/cs/exotica/wfmsys.pdf	1.61	1
11.	http://dis.sema.es/projects/WIDE/Documents	1.58	1
12.	http://ccs.mit.edu/klein/cscw98/paper07/	1.57	1
13.	http://ccs.mit.edu/klein/cscw98/paper08/	1.54	1
14.	http://www.computer.org/proceedings/Hiccs2/0001/00010198Babs.htm	1.51	0
15.	http://www.cs.toronto.edu/~avigal/nsfhtml/nsfhtml.html	1.47	1
16.	http://www.jeffsutherland.org/oopsla97/schulze.html	1.44	1
17.	http://osm7.cs.byu.edu/ER97/workshop4/ls.html	1.33	1
18.	http://www.informatik.uni-ulm.de/dbis/f&l/forschung/workflow/ftext-adept_e.html	1.33	1
19.	http://www.jeffsutherland.org/oopsla98/kuechler.html	1.32	1
20.	http://www.ifi.unizh.ch/groups/dbtg/Projects/SEAMAN/seaman1.html	1.22	0
		Precision	**0.9**

7. Conclusions & Outlook

- Coupling of Bingo! and Daffodil provides significantly added value for advanced users
  - For Daffodil: Context-based search over the Web
  - For Bingo!: Access to semi-structured meta data and the Deep Web

- Future work:
  - More exhaustive experimental evaluation
  - Extensive study of relevance feedback
  - DLO extraction from recommended URLs
  - Personalized crawling as a publicly available service for Daffodil users (dedicated server)

Try it out !

www.daffodil.de