Introduction to Web Archiving

Marc Spaniol

Saarbrücken, May 28, 2009

Agenda

- Motivation
  - Indexing vs. archiving
  - The challenge of Web archiving
  - Next generation Web archiving
- Aspects of Web archiving
  - Web archiving tools
  - Selection
  - Capturing
  - Archiving
  - Hosting
- Summary
- References

Indexing vs. Archiving

- Indexing
  - Completeness
  - Access to content
  - Scalability (speed)
  - Efficiency
  - Freshness

⇒ “Taking a Photo”

- Archiving
  - Completeness
  - Access to content
  - Scalability (coverage)
  - Authenticity
  - Coherence
  - Durability

⇒ “Shooting a Movie”

The Challenge of Web Archiving

- Digital library
  - Organized
  - Groomed content
  - Lots of metadata
  - Structured changes
  - Active preservation policies

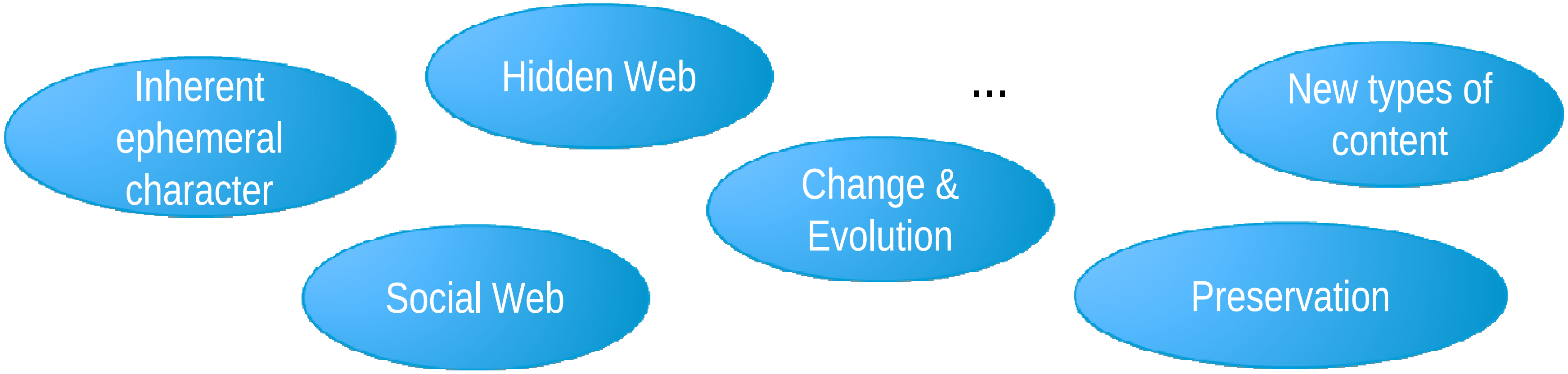
- World Wide Web
  - A disorganized free-for-all
  - Very little metadata
  - Unpredictable additions, deletions, modifications
  - No (coordinated) preservation strategy

Goals of Web Archiving

- Role of Web
  - Providing information and services for seemingly all domains
  - Reflecting all types of events, opinions, and developments within society, science, politics, environment, business, etc.
  - Giving room for the articulation for a multitude of stakeholders

⇒ Archiving this quickly changing multifaceted information space has becomes a relevant issue for cultural heritage

- Web archiving imposes various challenges:



Inherent ephemeral character

Hidden Web

…

New types of content

Change & Evolution

Social Web

Preservation

Living Web Archives

Introduction to Web Archiving

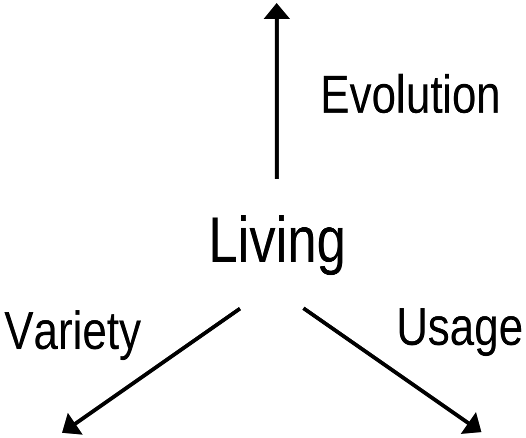
Marc Spaniol

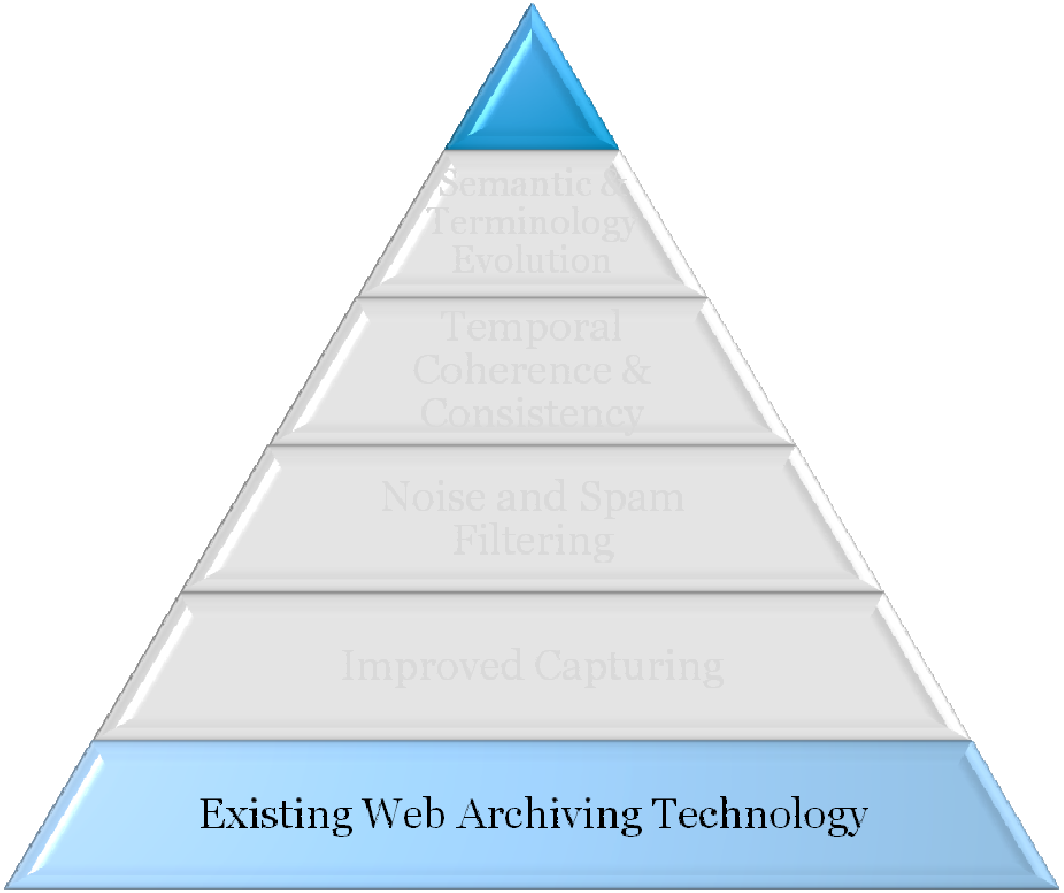
Next Generation Web Archiving

Development of Web archiving technology for

- High quality Web archives
- Long-term archive usability

⇒ From Web page storage to “Living Web Archives“



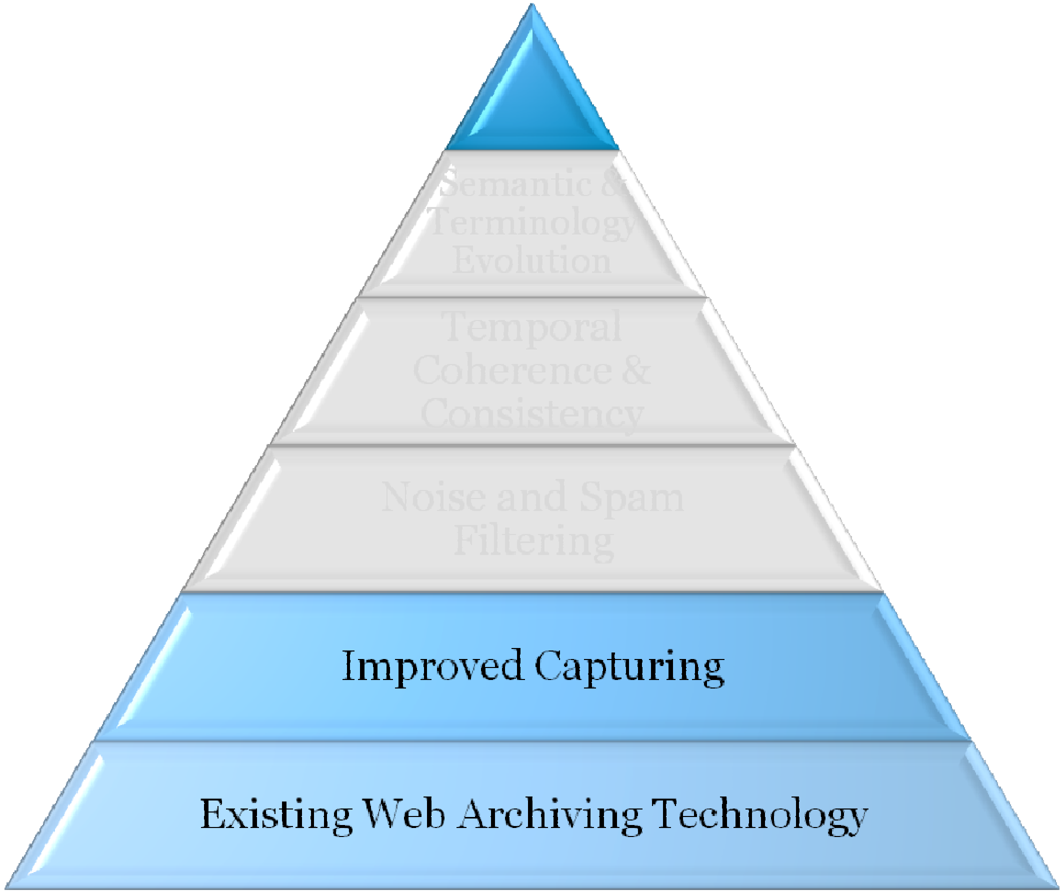


Semantic & Terminology Evolution
Temporal Coherence & Consistency
Noise and Spam Filtering
Improved Capturing
Existing Web Archiving Technology

Archive Fidelity

Next generation Web archiving methods and tools

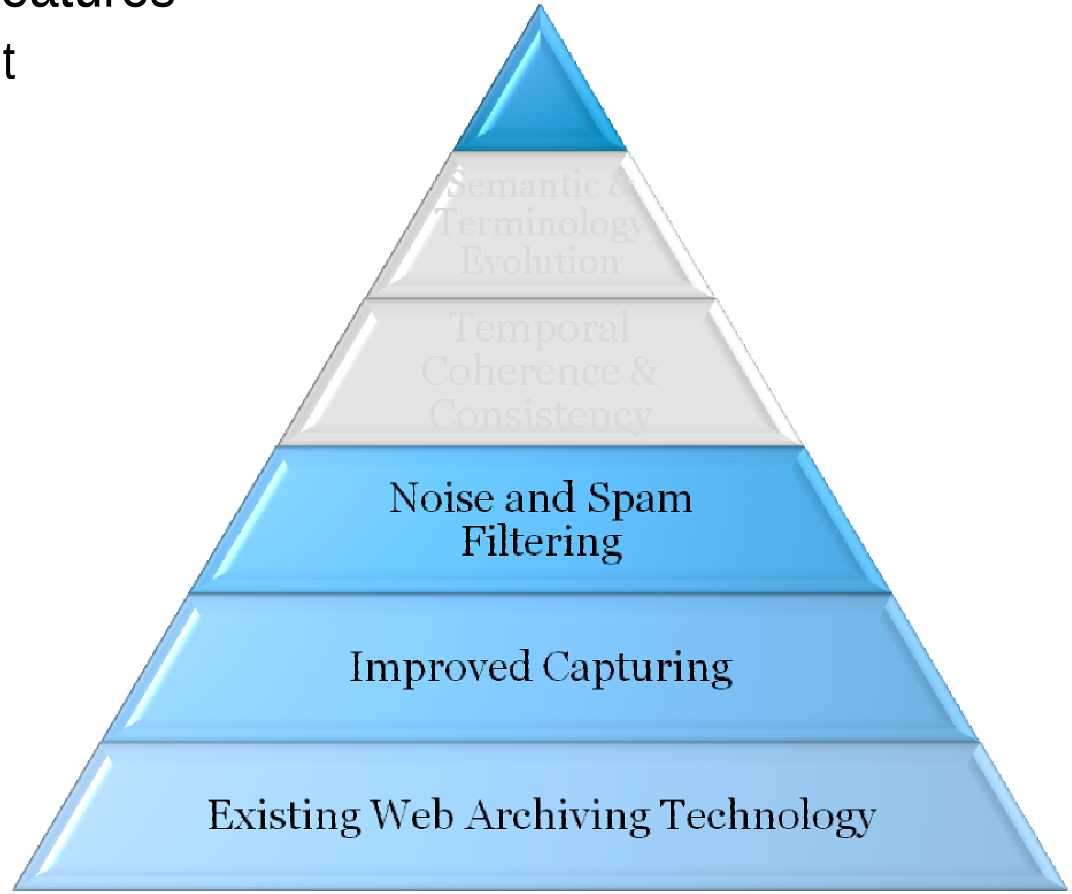
- Enhance archive fidelity and authenticity by
  - Capturing all types of content
  - Capturing of hidden Web
  - Detecting traps


Semantic & Terminology Evolution
Temporal Coherence & Consistency
Noise and Spam Filtering
Improved Capturing
Existing Web Archiving Technology

Advanced Filtering

Next generation Web archiving methods and tools:

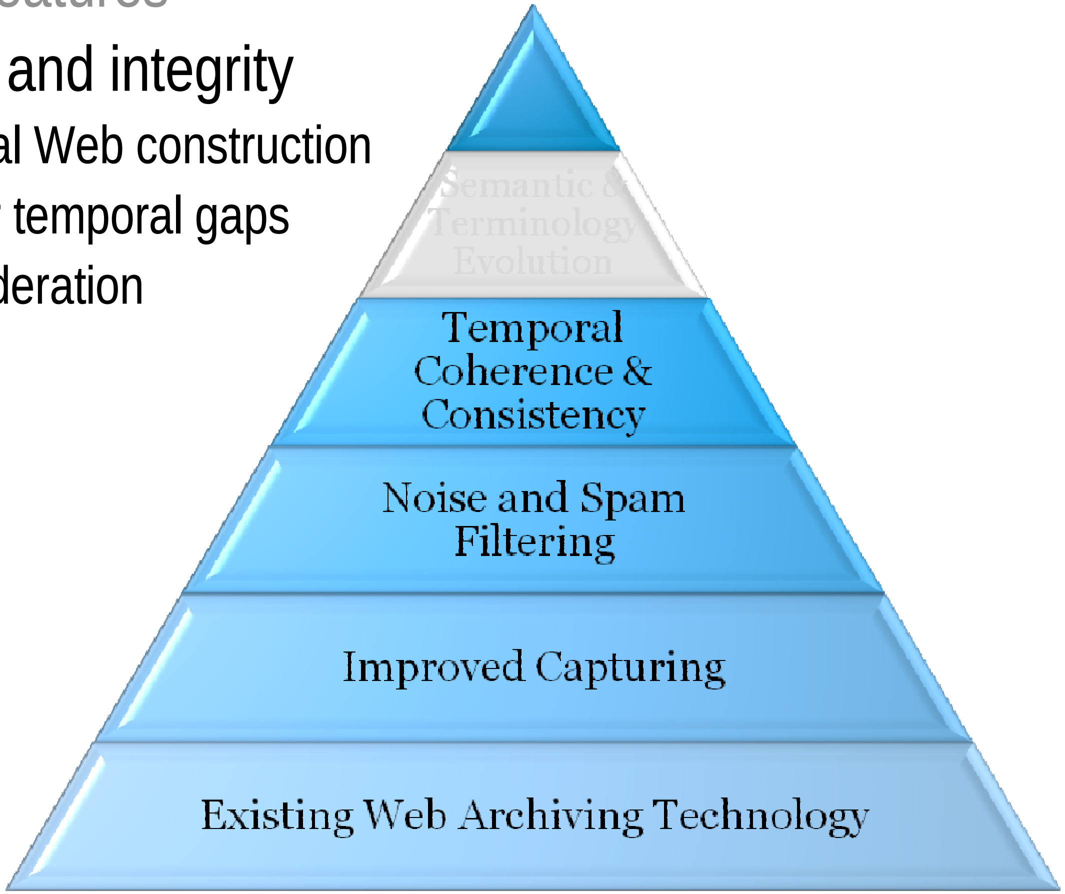
- Enhance archive fidelity and authenticity
- Provide advanced filtering features
  - Capture all types of content
  - Detect traps
  - Filtering Web spam
  - Filtering noise


Semantic & Terminology Evolution
Temporal Coherence & Consistency
Noise and Spam Filtering
Improved Capturing
Existing Web Archiving Technology

Archive Coherence

Next generation Web archiving methods and tools

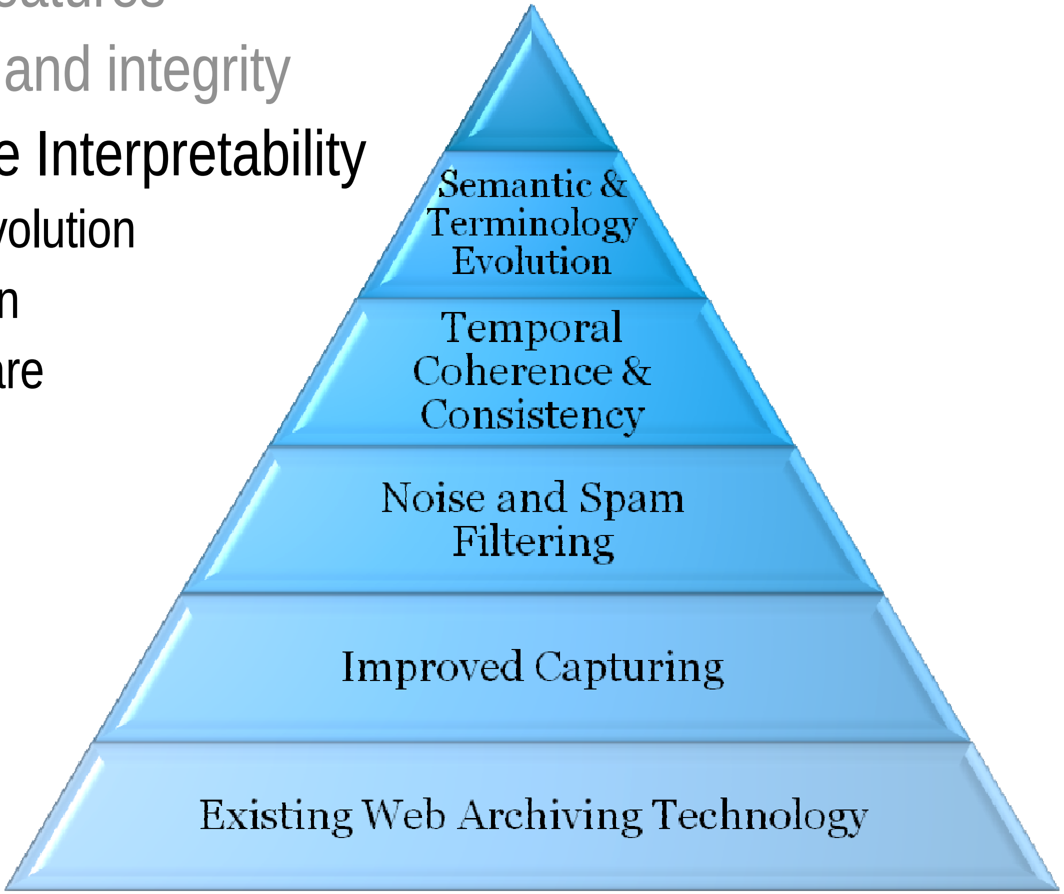
- Enhance archive fidelity and authenticity
- Provide advanced filtering features
- Improve archive coherence and integrity
  - Deal with issues of temporal Web construction
  - Identify, analyze and repair temporal gaps
  - Consistent Web archive federation


Semantic & Terminology Evolution
Temporal Coherence & Consistency
Noise and Spam Filtering
Improved Capturing
Existing Web Archiving Technology

Archive Interpretability

Next generation Web archiving methods and tools

- Enhance archive fidelity and authenticity
- Provide advanced filtering features
- Improve archive coherence and integrity
- Facilitate (long-term) archive Interpretability
  - Dealing with terminology evolution
  - Handling semantic evolution
  - Preparing for evolution aware access support



Semantic & Terminology Evolution

Temporal Coherence & Consistency

Noise and Spam Filtering

Improved Capturing

Existing Web Archiving Technology

Goals of Web Archiving Summarized

- Archiving function α applied to website W produces a capture C_W of the web site's resources and related metadata:

$$\alpha(W) \rightarrow C_W$$

- Restoration function ρ "unpacks" the capture C_W and reproduces the original site:

$$\rho(C_W) \rightarrow W$$

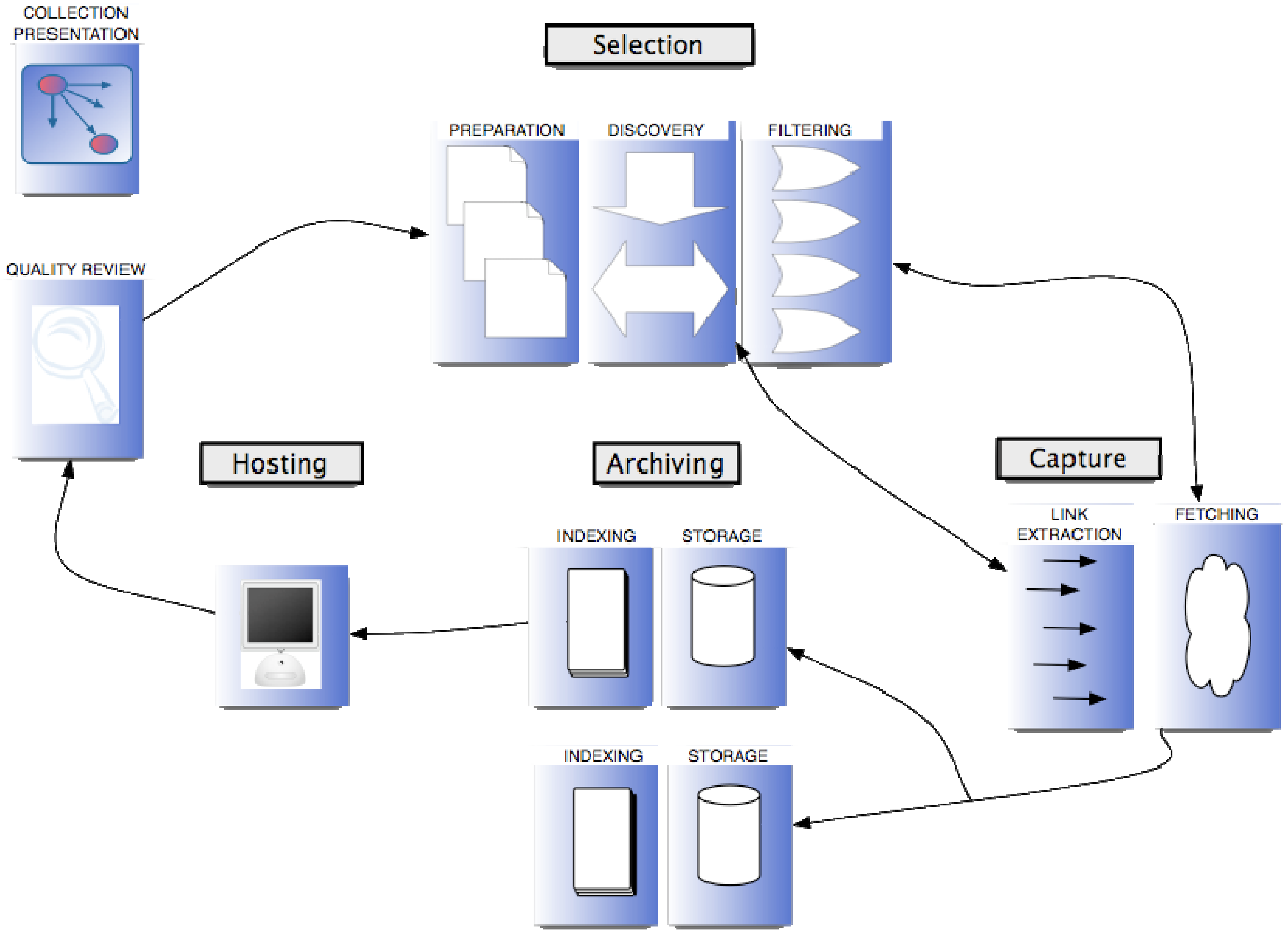
- Transformation function τ "unpacks" the capture C_W, converts the components to the modern-day equivalent, and reproduces the original site within a new environment:

$$\tau(C_W) \rightarrow W_\Delta$$

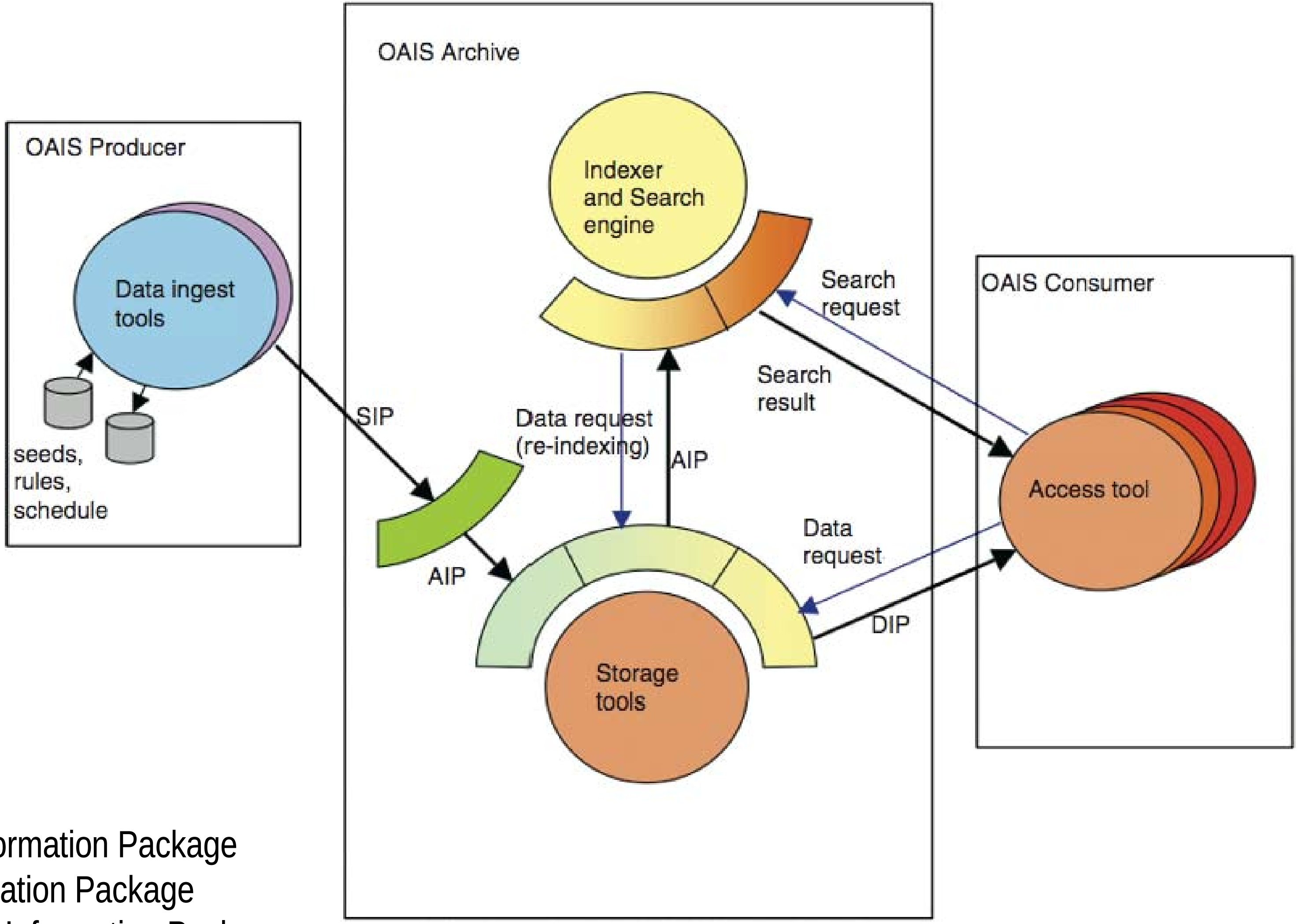
Introduction to Web Archiving

Marc Spaniol

Aspects of Web Archiving



Web Archiving Tools

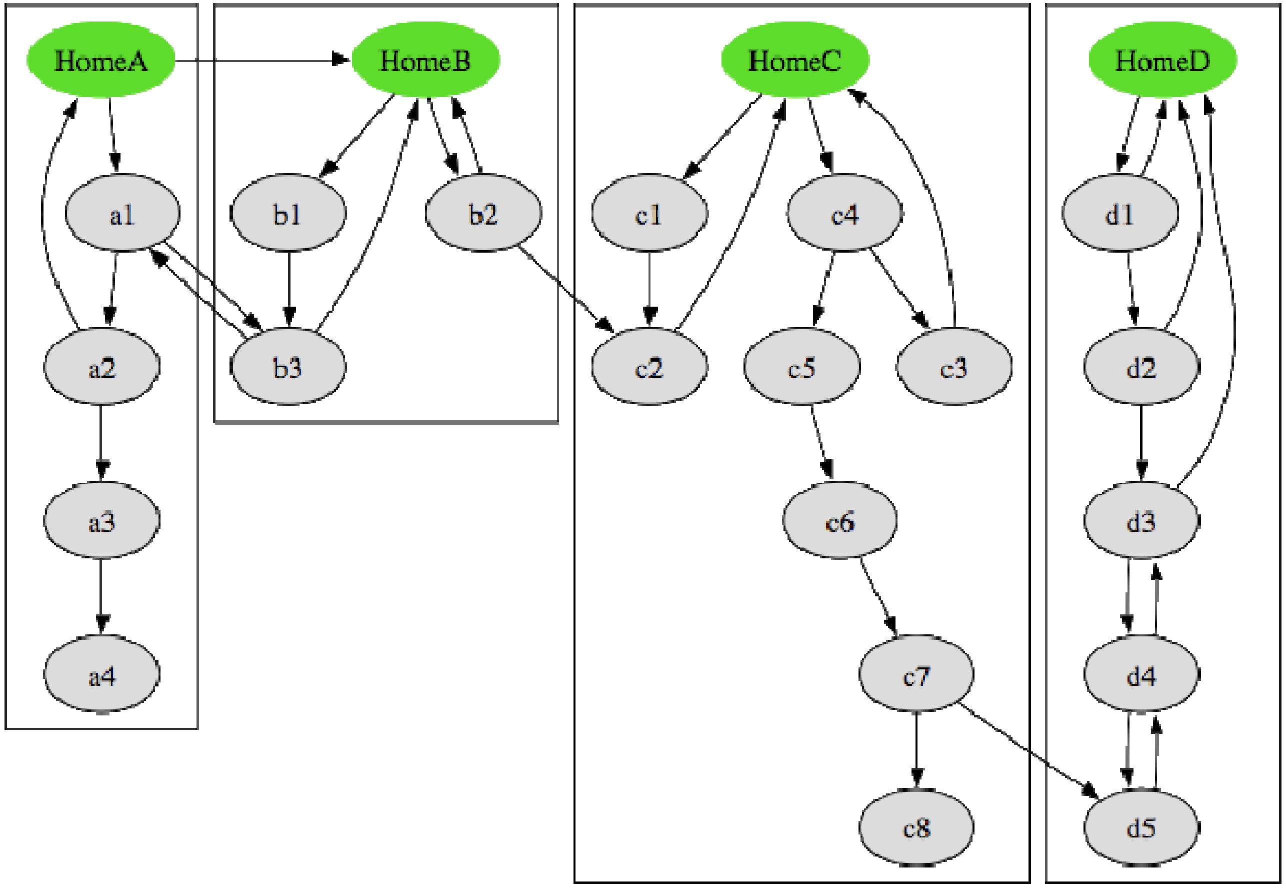


AIP: Archival Information Package
DIP: Data Information Package
SIP: Submission Information Package
OAIS: Open Archival Information System

Selection of Seed(s) and Scope

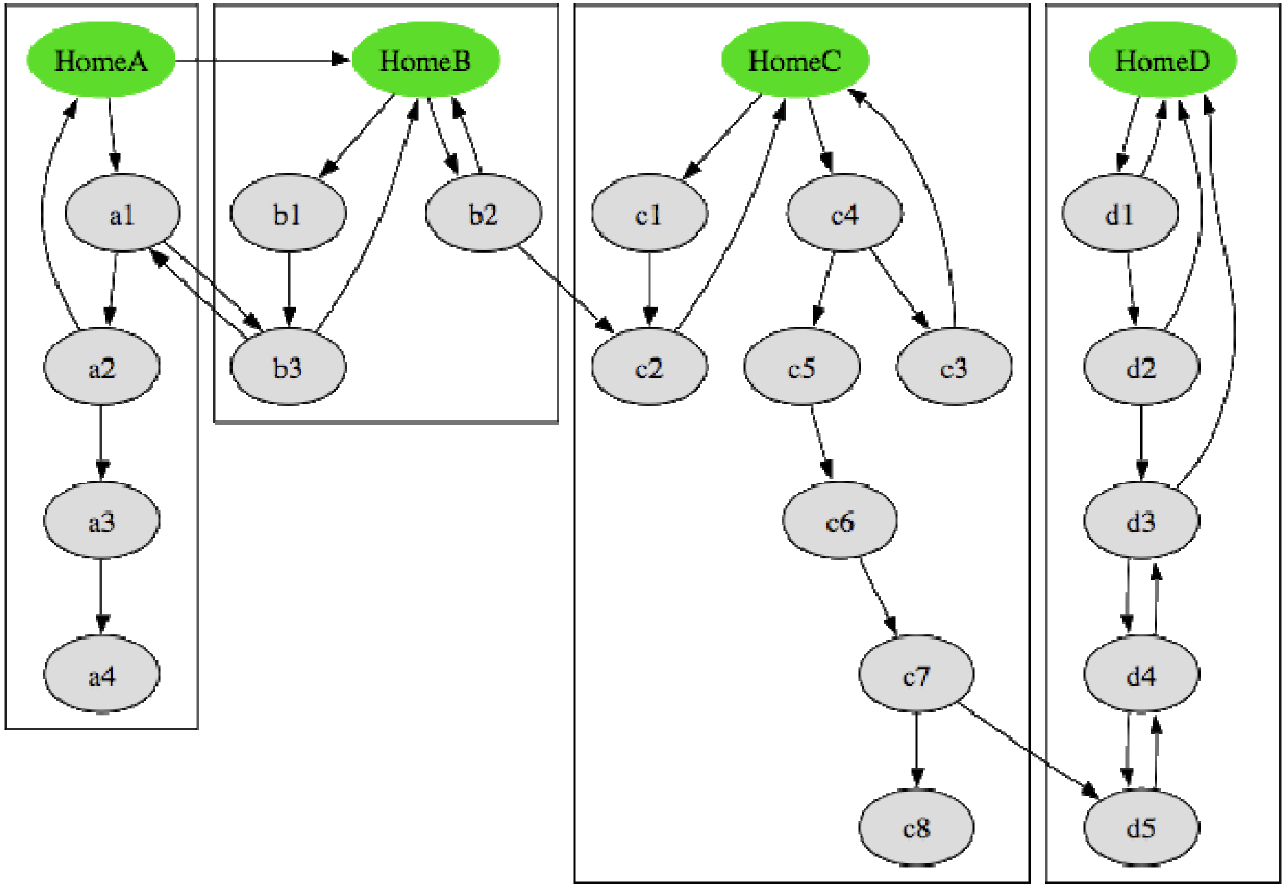
- Entry point / seed:
  Where the capturing process (crawl) starts. Top of the hypertext path that will be followed.

- Scope:
  The extent of the area that will be included in the gathering, as defined by criteria applicable to each node.



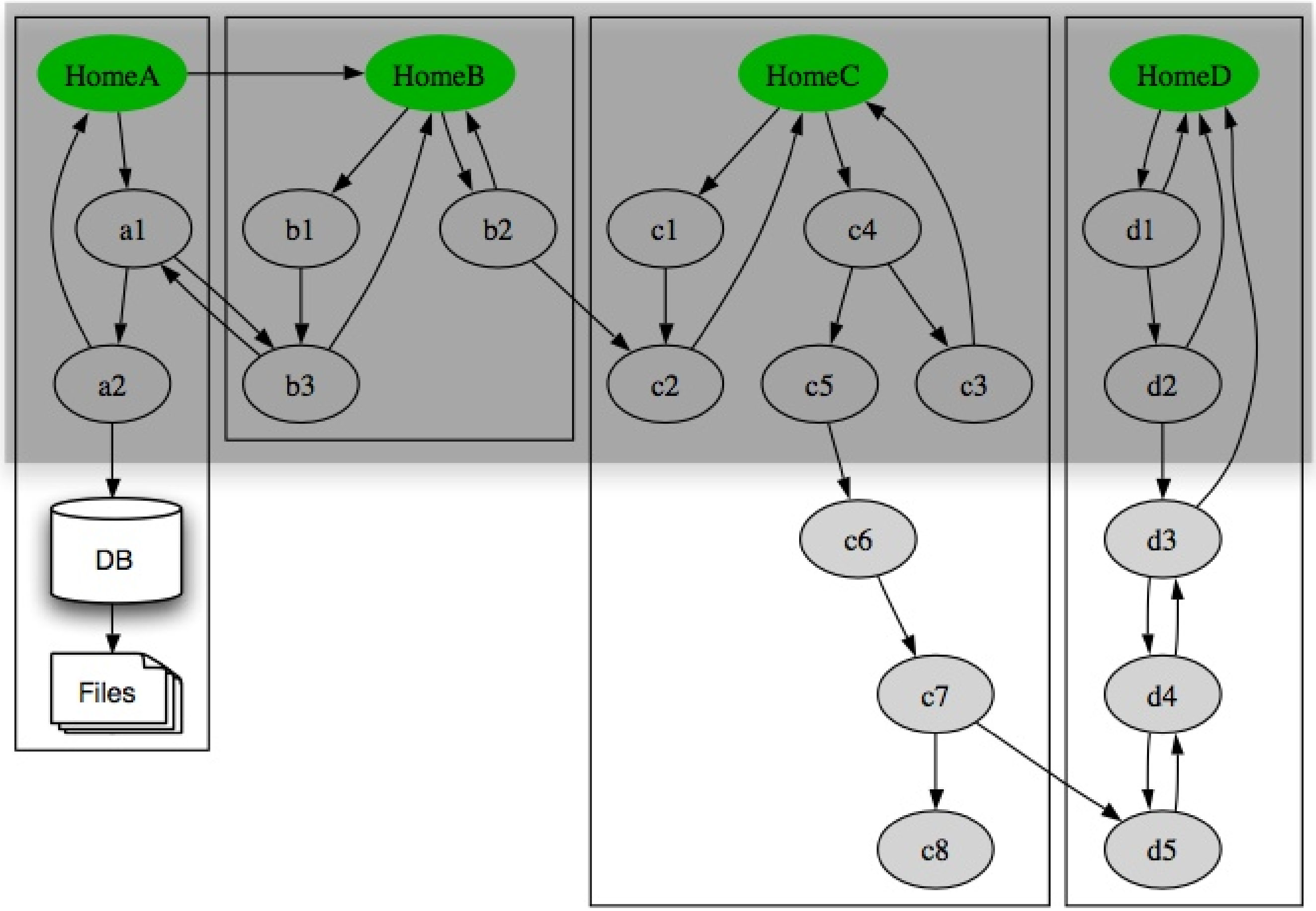
Completeness

- Vertically: Number of relevant nodes found from entry point.
- Horizontally: Number of relevant entry points found within the designated perimeter.



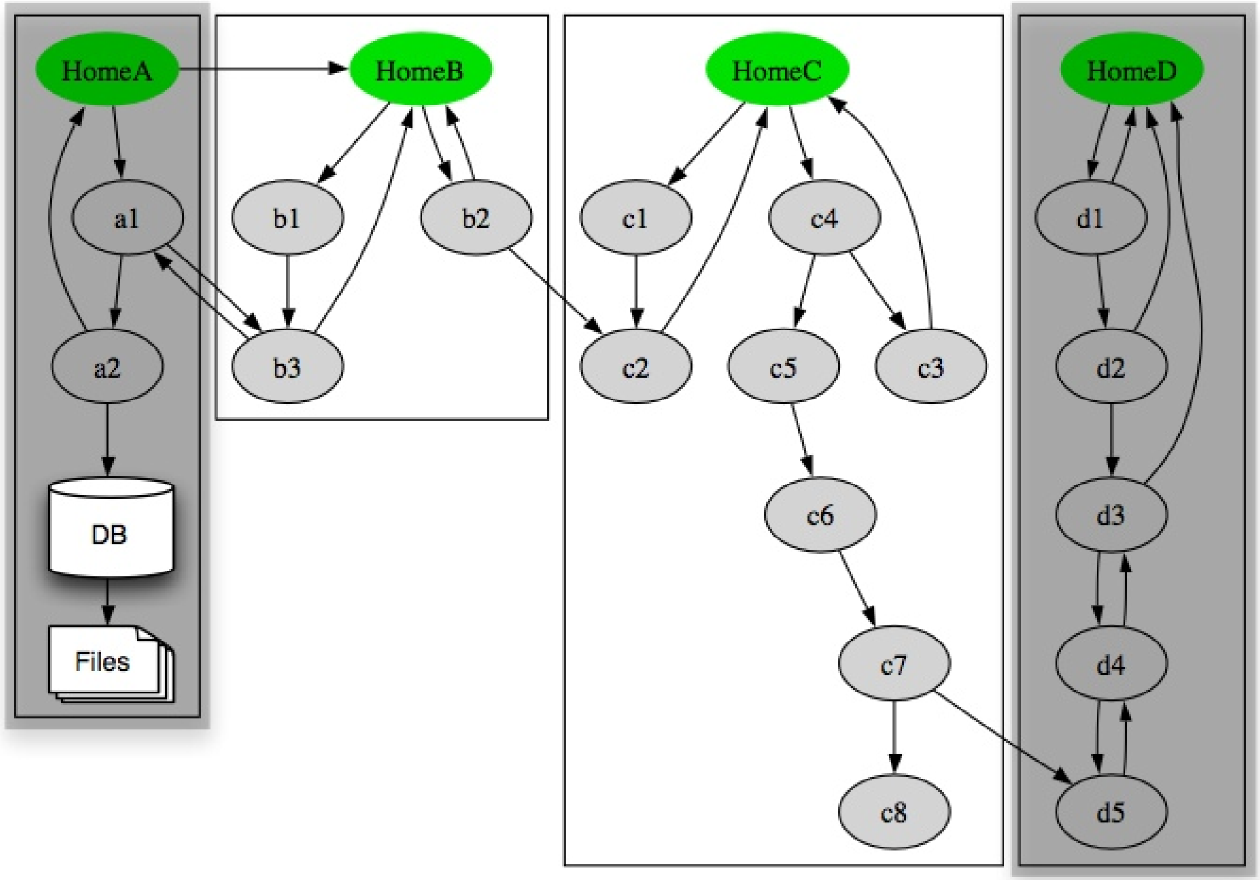
Extensive Collection

- Horizontal completeness is preferred to vertical completeness
- Holistic, domain based, or topic-centric archiving



Intensive Collection

- Vertical completeness is preferred to horizontal completeness
- Site-based archiving
- Defines the high level target of a collection
- Explicit exclusion to avoid duplicate content with other collections



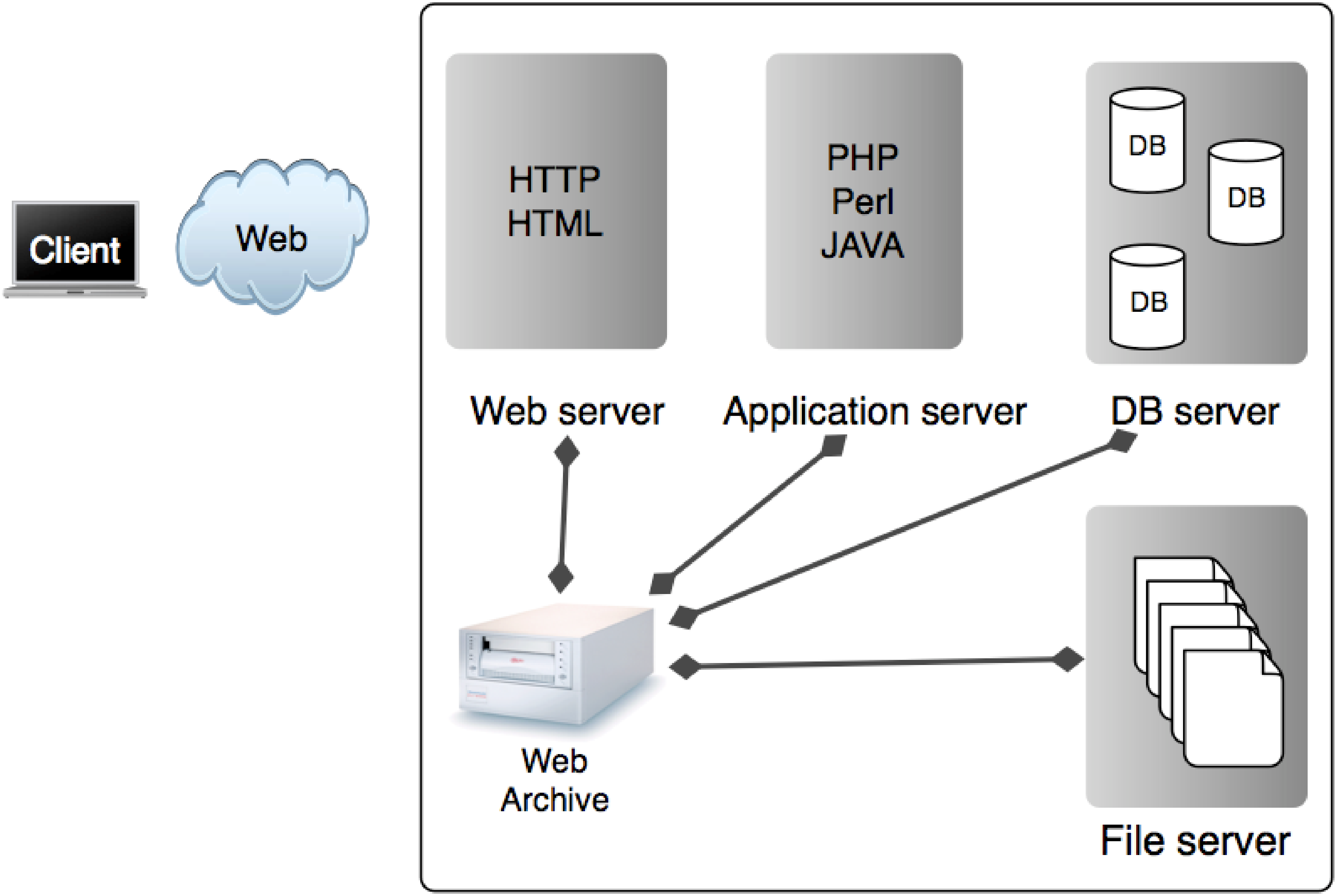
The Challenge of Web Archiving

- HTTP cannot ask for only new or modified contents
  - *Timestamps* have limited benefit
  - No list of pages that have been deleted, changed, and added
  - *Each* content must be requested, one at a time, *by name*
- There is no “SELECT *” in HTTP
  - Crawlers can only GET one resource at a time, by name
  - HTTP cannot give a crawler a list of all URLs for the site

⇒ Undiscovered or hidden resources will not be captured or refreshed

⇒ “Strategy” required

Server Side Archiving



Server Side Archiving Revisited

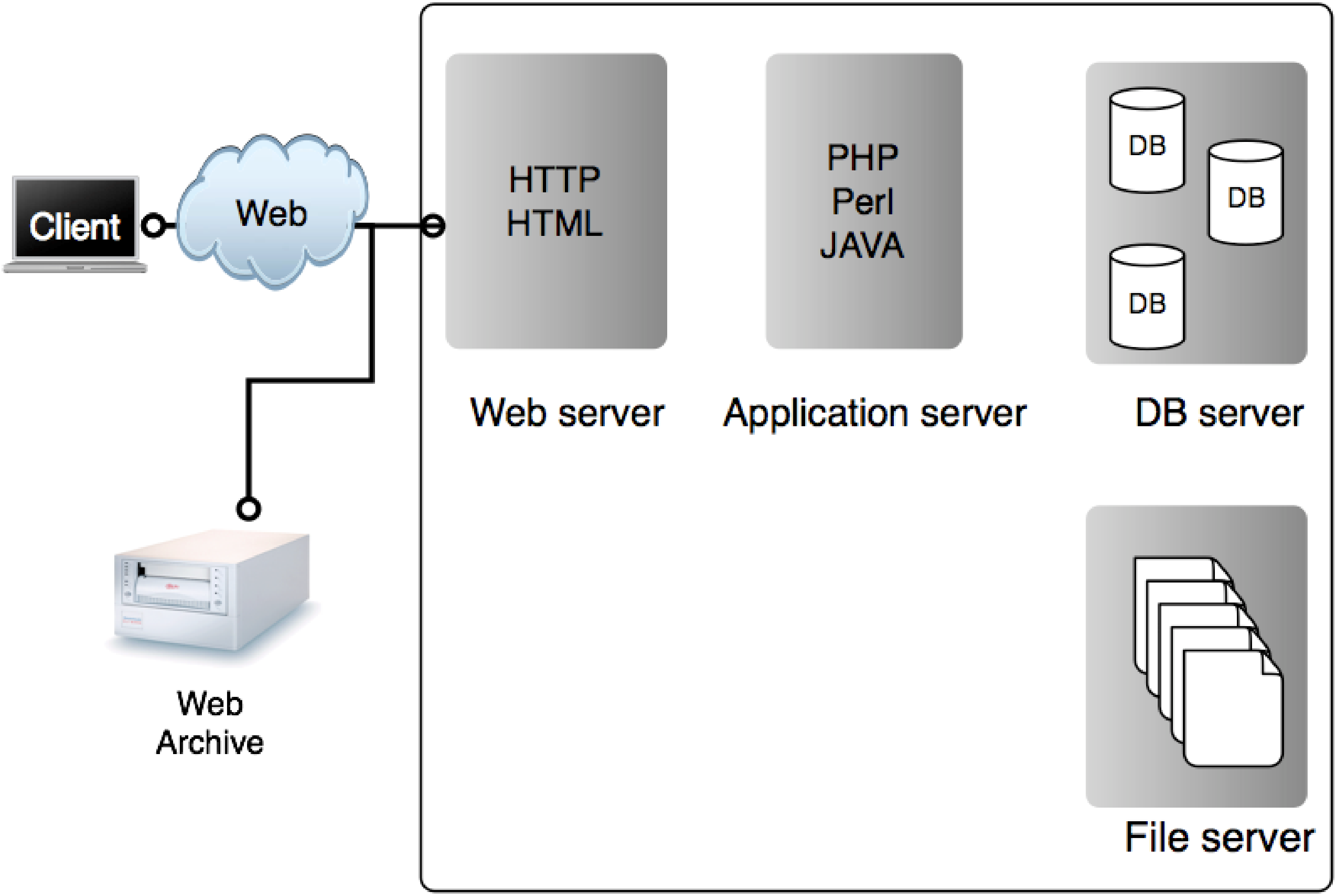
- Benefits
  - + Extremely comprehensive
  - + Changes are fully traceable (if budget permits)
  - + Instantaneous snapshots possible
  - + No network latency or limitations
  - + Deep Web compliant

- Drawbacks
  - - Change monitoring may decrease server performance
  - - Needs sophisticated set-up
  - - Requires server access

Introduction to Web Archiving

Marc Spaniol

Transaction based Archiving

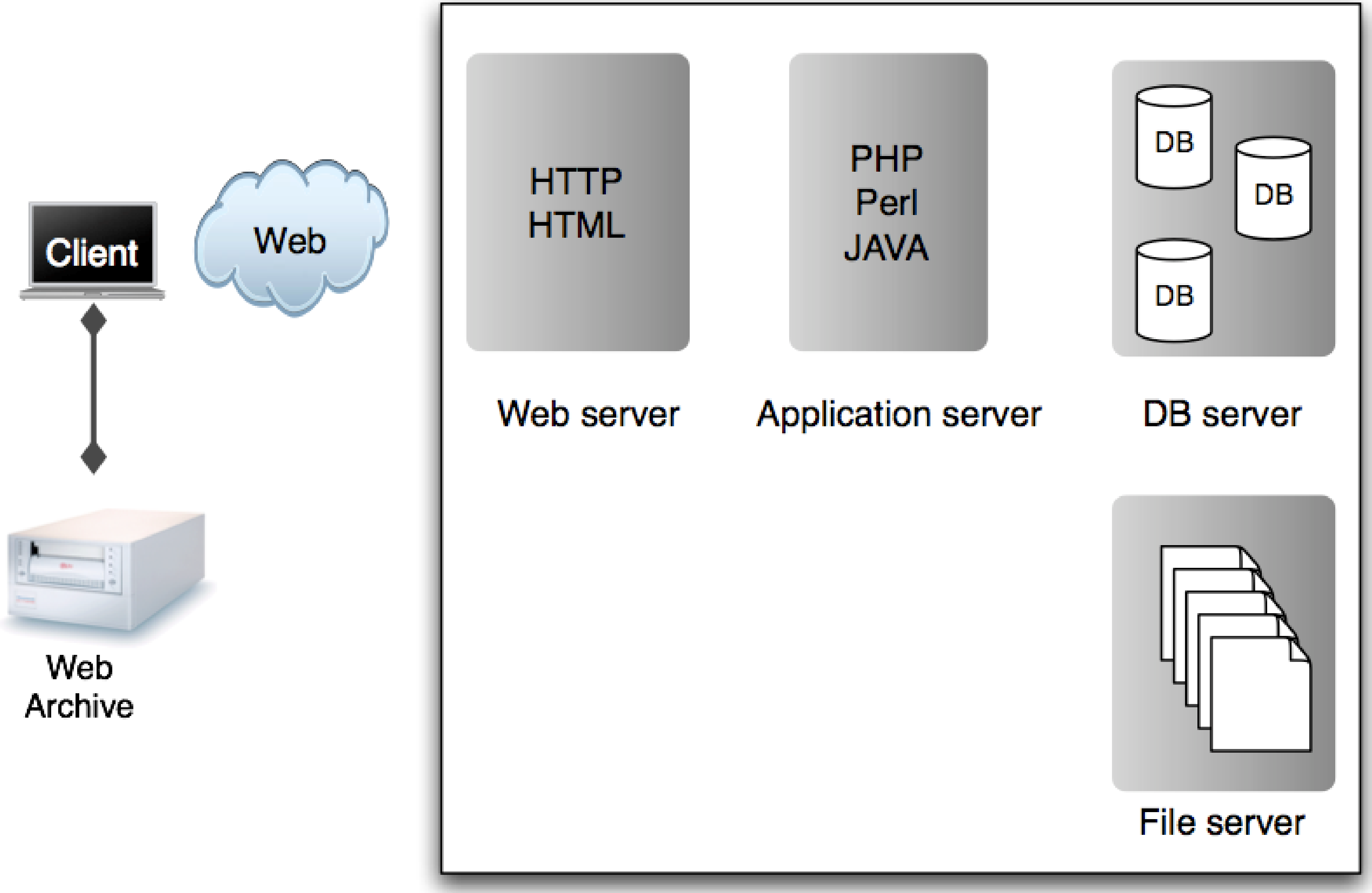


Transaction based Archiving Revisited

- Benefits
  - + Comes for “free”
  - + “Smart” coverage achieved by human interaction
  - + Simple maintenance
  - + No server collaboration/manipulation required

- Drawbacks
  - - Unsystematic
  - - Data quality is potentially poor
  - - Needs traffic monitoring
  - - Privacy issues
  - - Potential network latency or limitations
  - - Requires constant traffic

Client Side Archiving



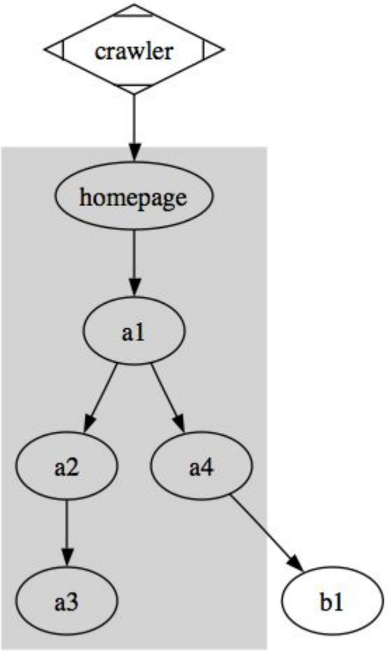
Client Side Archiving Revisited

- Benefits
  - + No server collaboration/manipulation needed
  - + Only crawler set-up required
  - + Mostly automated process (daily/weekly/monthly)

- Drawbacks
  - - Changes might get lost
  - - Good data quality requires sophisticated crawling strategies
  - - Potential network latency or limitations
  - - Computational "expensive"

<u>Next week's lecture:</u> "Data Quality in Web Archiving"

Web Capturing with Heritrix



- Internet Archive's crawler
  - Open source java implementation
  - Web-scale archiving crawler
  - Extensible
- Key components
  - Scope
  - Frontier
  - Processor chains
- Configuration options include
  - Crawl scope, e.g. via
    - SURT expression: +http://(de,mpi-inf.mpg,www,)/
    - Regular expression: ^(http|https|dns):(//)?[a-zA-Z0-9\.]*mpi-inf.mpg.de/.*
  - Lot of fine-tuning features options
    - delay-factor
    - max-delay-ms
    - min-delay-ms
    - max-retries
    - retry-delay-seconds
    - etc.

SURT
Sort-friendly URI Reordering Transform

- Transformation applied to URIs
  - Left-to-right representation matching the natural hierarchy of domain names
  - Useful when comparing or sorting URIs

- Converting URIs according to SURT
  - Make all characters lowercase
  - Change the 'https' scheme to 'http‘
  - '/' after a URI authority component only appear in the SURT form if it appeared in the plain URI form

- SURT form URIs are typically not used to specify exact URIs for fetching

- Less expressive than regular expressions → Exercises

SURT Prefix

- Used for crawl scope specification in Heritrix

- Conversion to SURT prefix:

  1. Convert the URI to its SURT form.

  2. If there are ≥ 3 slashes ('/') in the SURT form, remove everything after the last slash

     <http://(org,example,www,)/main/subsection/> ✓

     <http://(org,example,www,)/main/subsection> ⟶ <http://(org,example,www,)/main/>

     <http://(org,example,www,)/> ✓

     <http://(org,example,www,)> ✓

3. If the resulting form ends in an off-parenthesis ')', remove the off-parenthesis

   <http://(org,example,www,)> ⟶ <http://(org,example,www,>

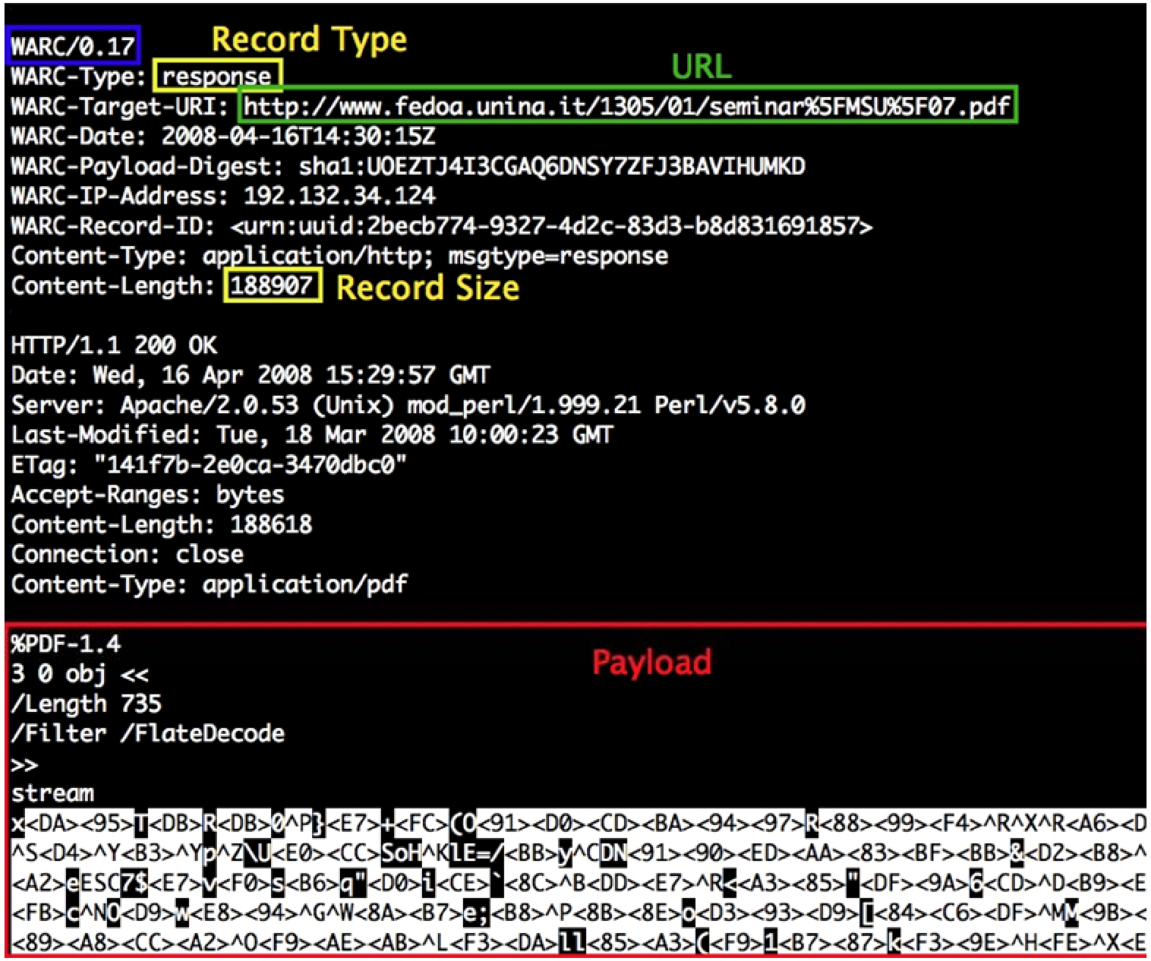
Living Web Archives

Introduction to Web Archiving

Marc Spaniol

Heritrix Output

- ARC/WARC files (Web ARChive) ~ 500 MB – 1 GB each
- “ZIP files” of content(s) and some metadata



```
WARC/0.17
WARC-Type: response
WARC-Target-URI: http://www.fedoa.unina.it/1305/01/seminar%5FMSU%5F07.pdf
WARC-Date: 2008-04-16T14:30:15Z
WARC-Payload-Digest: sha1:UOEZTJ4I3CGAQ6DNSY7ZFJ3BAVIHUMKD
WARC-IP-Address: 192.132.34.124
WARC-Record-ID: <urn:uuid:2becb774-9327-4d2c-83d3-b8d831691857>
Content-Type: application/http; msgtype=response
Content-Length: 188907

HTTP/1.1 200 OK
Date: Wed, 16 Apr 2008 15:29:57 GMT
Server: Apache/2.0.53 (Unix) mod_perl/1.999.21 Perl/v5.8.0
Last-Modified: Tue, 18 Mar 2008 10:00:23 GMT
ETag: "141f7b-2e0ca-3470dbc0"
Accept-Ranges: bytes
Content-Length: 188618
Connection: close
Content-Type: application/pdf

%PDF-1.4
3 0 obj <<
/Length 735
/Filter /FlateDecode
>>
stream
x<DA><95>T<DB>R<DB>0^P}<E7>+<FC>(O<91><D0><CD><BA><94><97>R<88><99><F4>^R^X^R<A6><D
^S<D4>^Y<B3>^Yp^Z\U<E0><CC>SoH^KLE=/<BB>y^CDN<91><90><ED><AA><83><BF><BB>&<D2><B8>^
<A2>eESC7$<E7>v<F0>s<B6>q"<D0>i<CE>`<8C>^B<DD><E7>^R<<A3><85>"<DF><9A>6<CD>^D<B9><E
<FB>c^NO<D9>W<E8><94>^G^W<8A><B7>e;<B8>^P<8B><8E>o<D3><93><D9>[<84><C6><DF>^MM<9B><
<89><A8><CC><A2>^O<F9><AE><AB>^L<F3><DA>ll<85><A3>(<F9>1<B7><87>k<F3><9E>^H<FE>^X<E
```

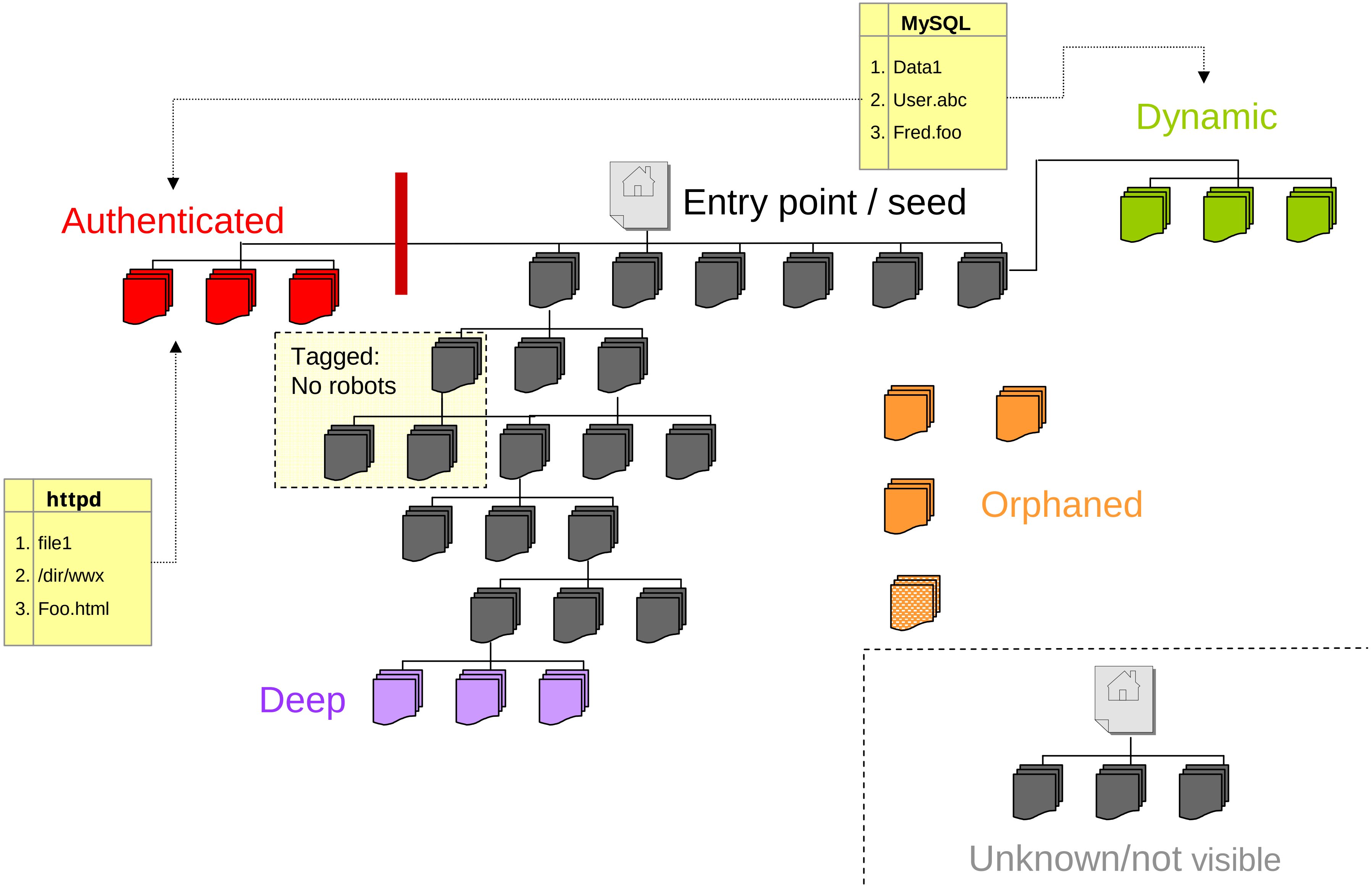
Databases and Information Systems
Prof. Dr. G. Weikum

Living Web Archives

Introduction to Web Archiving

Marc Spaniol

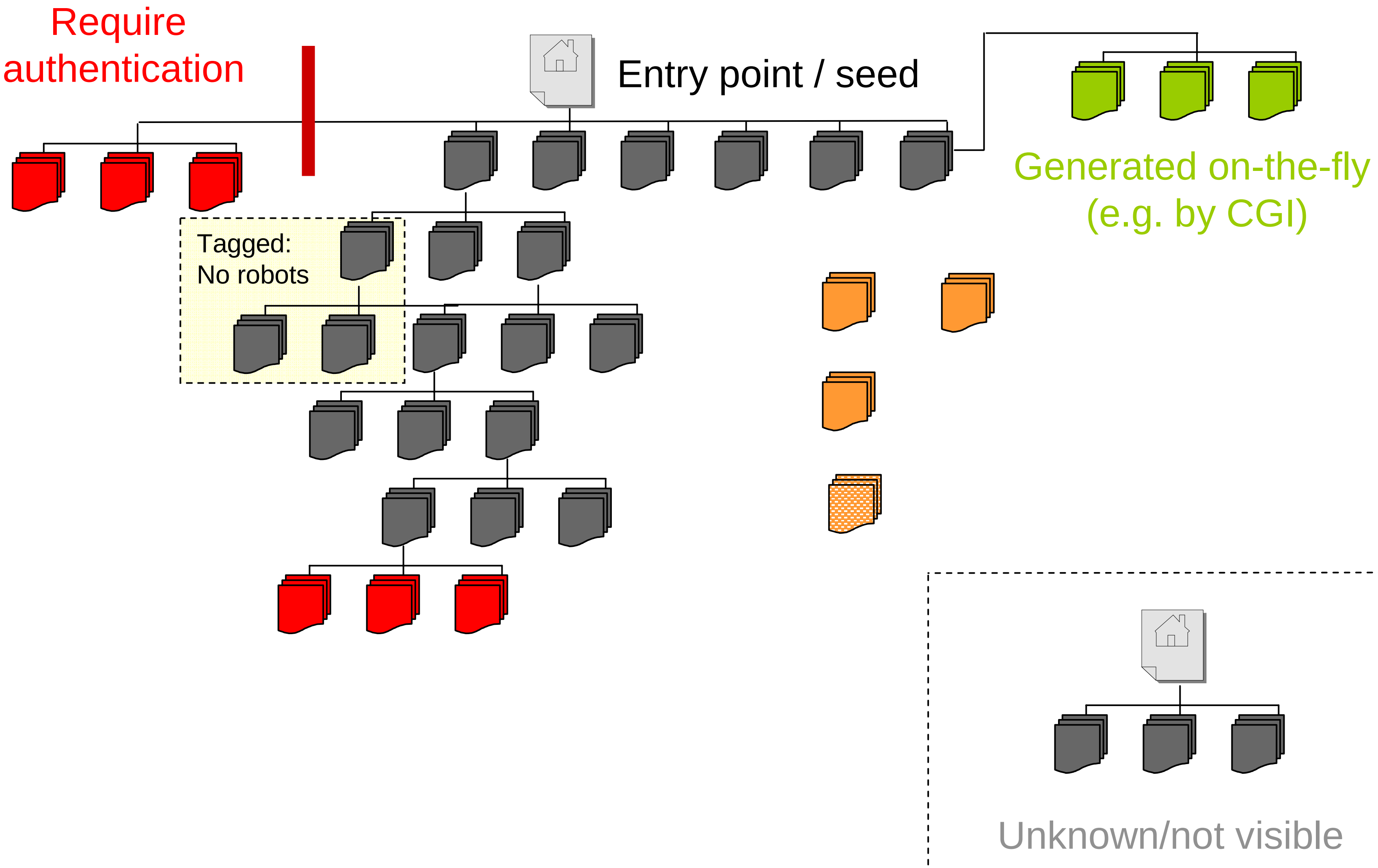
A Webmaster's Omniscient View



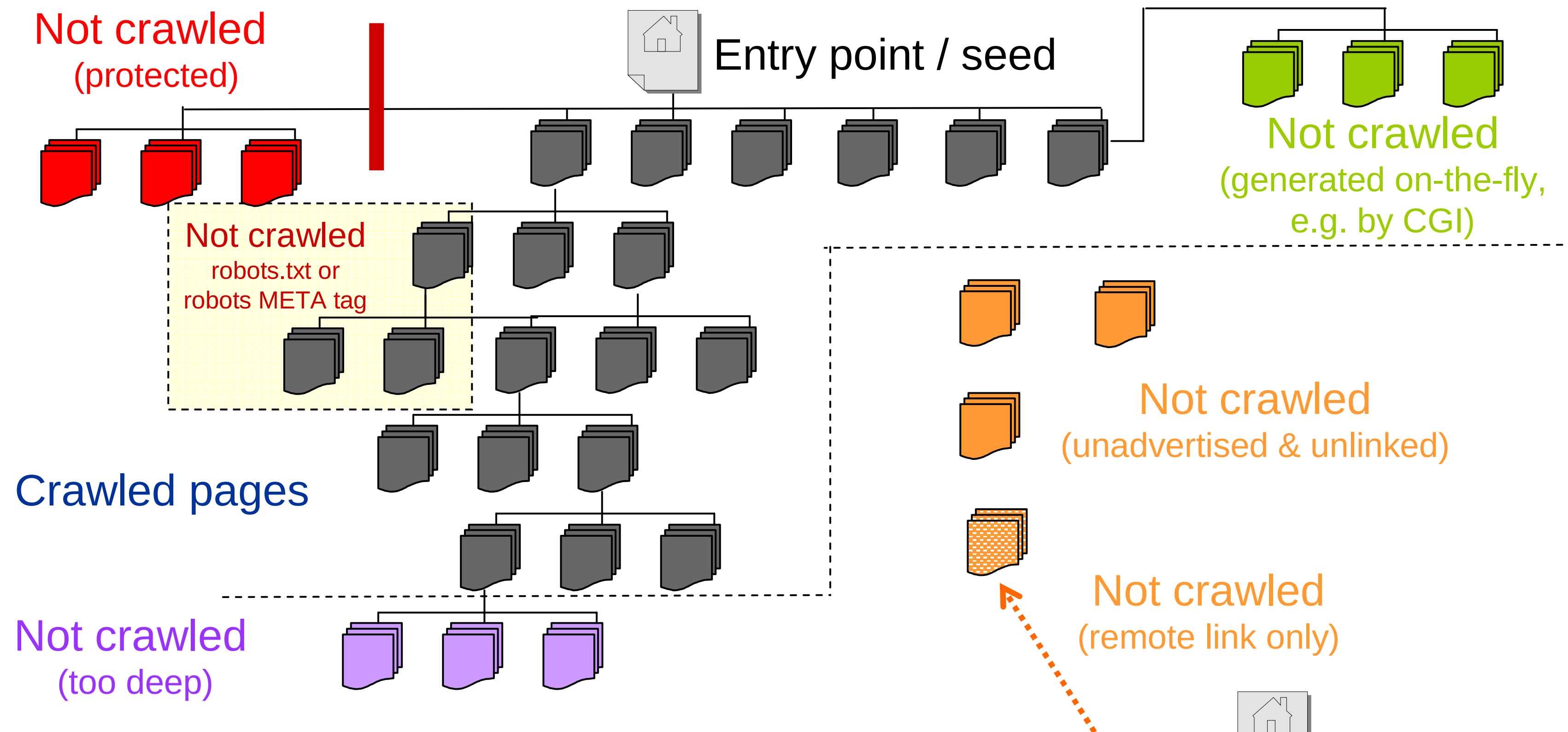
	MySQL
1.	Data1
2.	User.abc
3.	Fred.foo

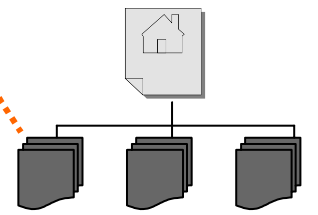
	httpd
1.	file1
2.	/dir/wwx
3.	Foo.html

Web Server’s View of a Web Site

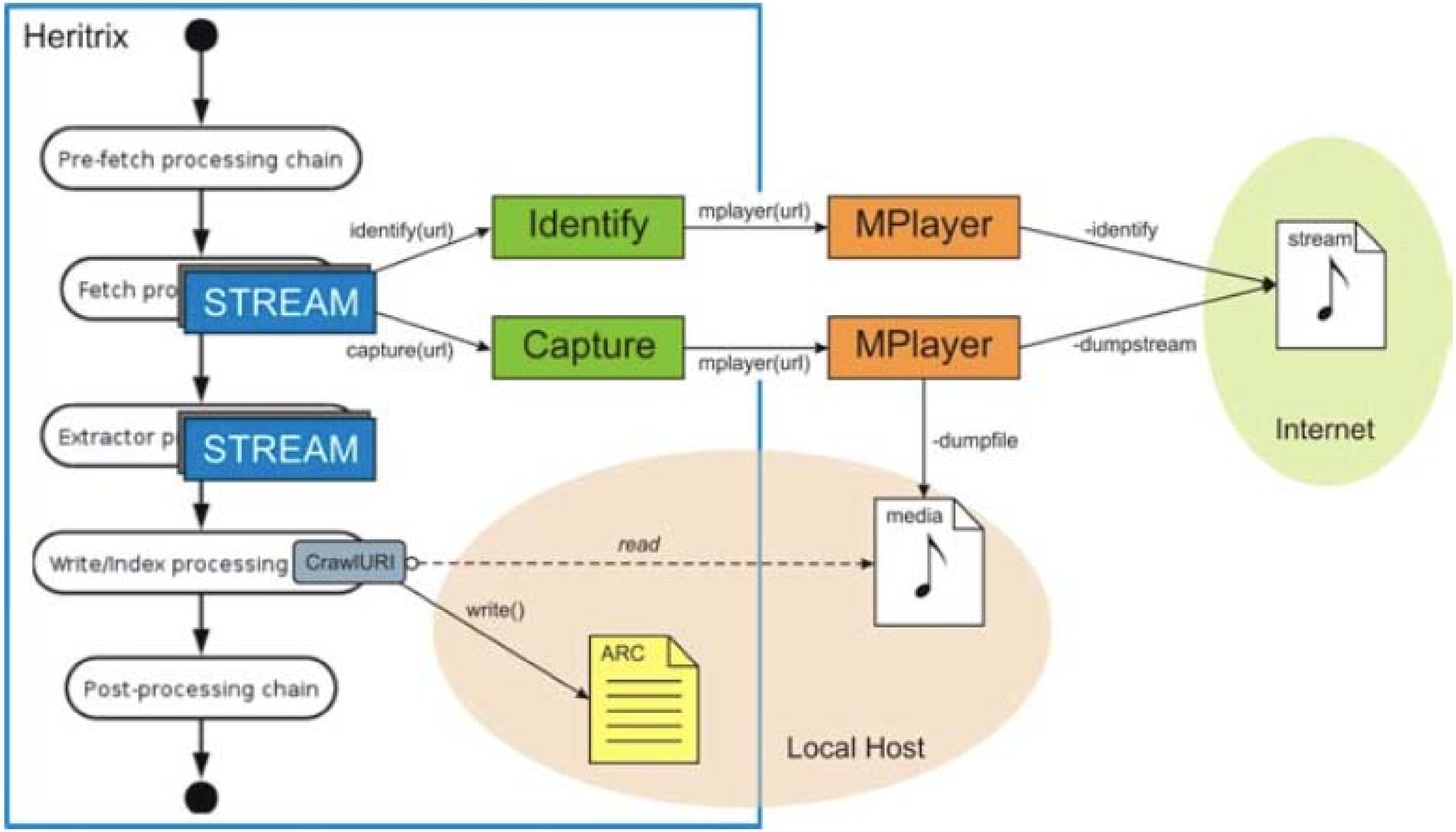


A Crawler's View of a Web Site



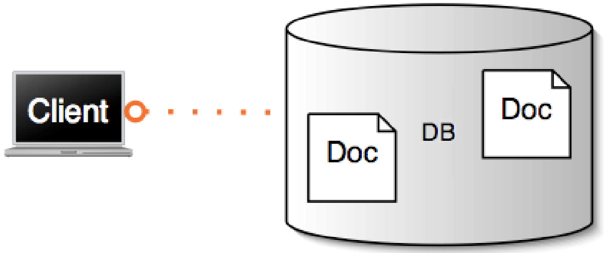


Streaming Media Capturing

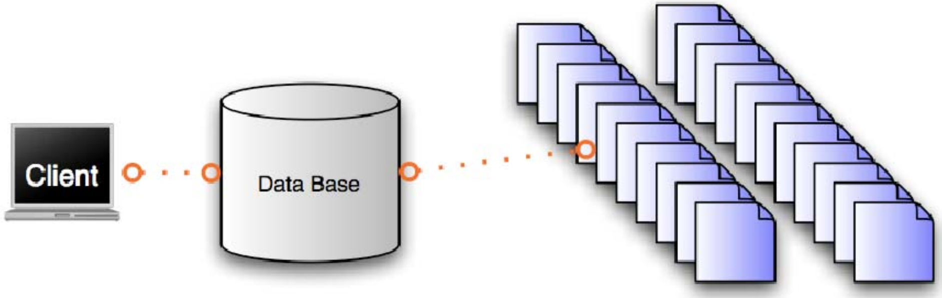


Web Information Systems

Dynamic Web sites



Hidden Web

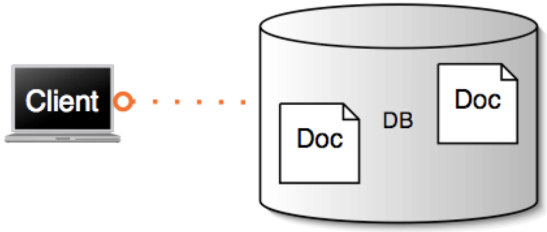


- Each interaction with a Web information system can potentially generate a unique customized response

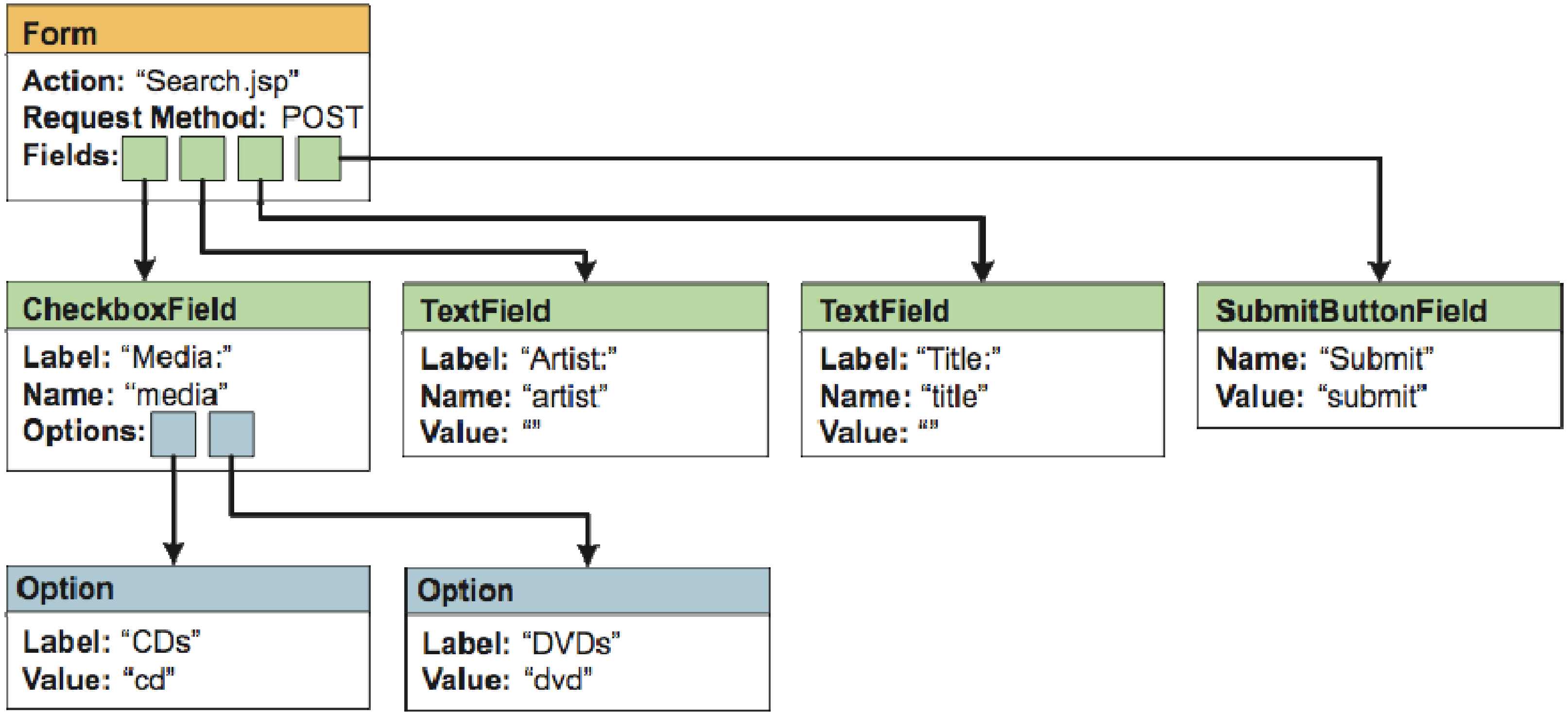
⇒ Document the context of this interaction, or pseudo-transaction

Hidden Web Archiving

- Procedure:
  1. Detect it
  2. Try to crawl it by automatic query generation
  3. Or encourage site producer to be more friendly to crawlers
- Special crawl for documentary gateways
  - Find patterns
  - Feed the fields
    - Finding forms is easy, filling them is not
    - Proximity of text near fields (beyond and on the left)
    - Tokenization and analyze
  - Reconstruct navigation or access logic



Hidden Web Archiving: HTML Form extraction

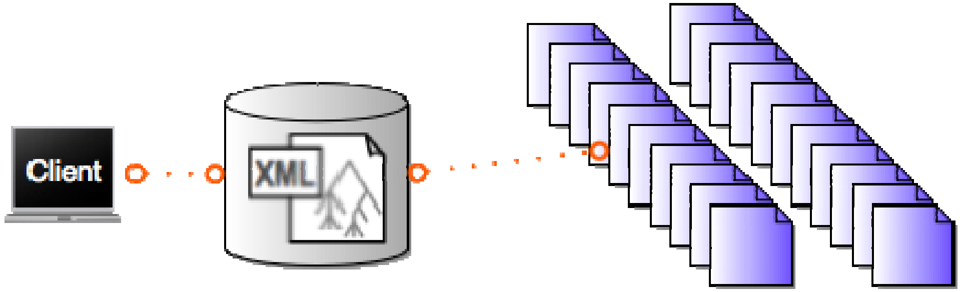


[Fontes & Soares Silva, WIDM 2004]

Crawler-Server Collaboration

- Open Archives Initiative (OAI) Protocol for Metadata Harvesting
- Provided flat list (maybe hidden for public)
- RSS feeds
- OAI server
  - Pushed by search-engines
  - Yahoo content acquisition program, google

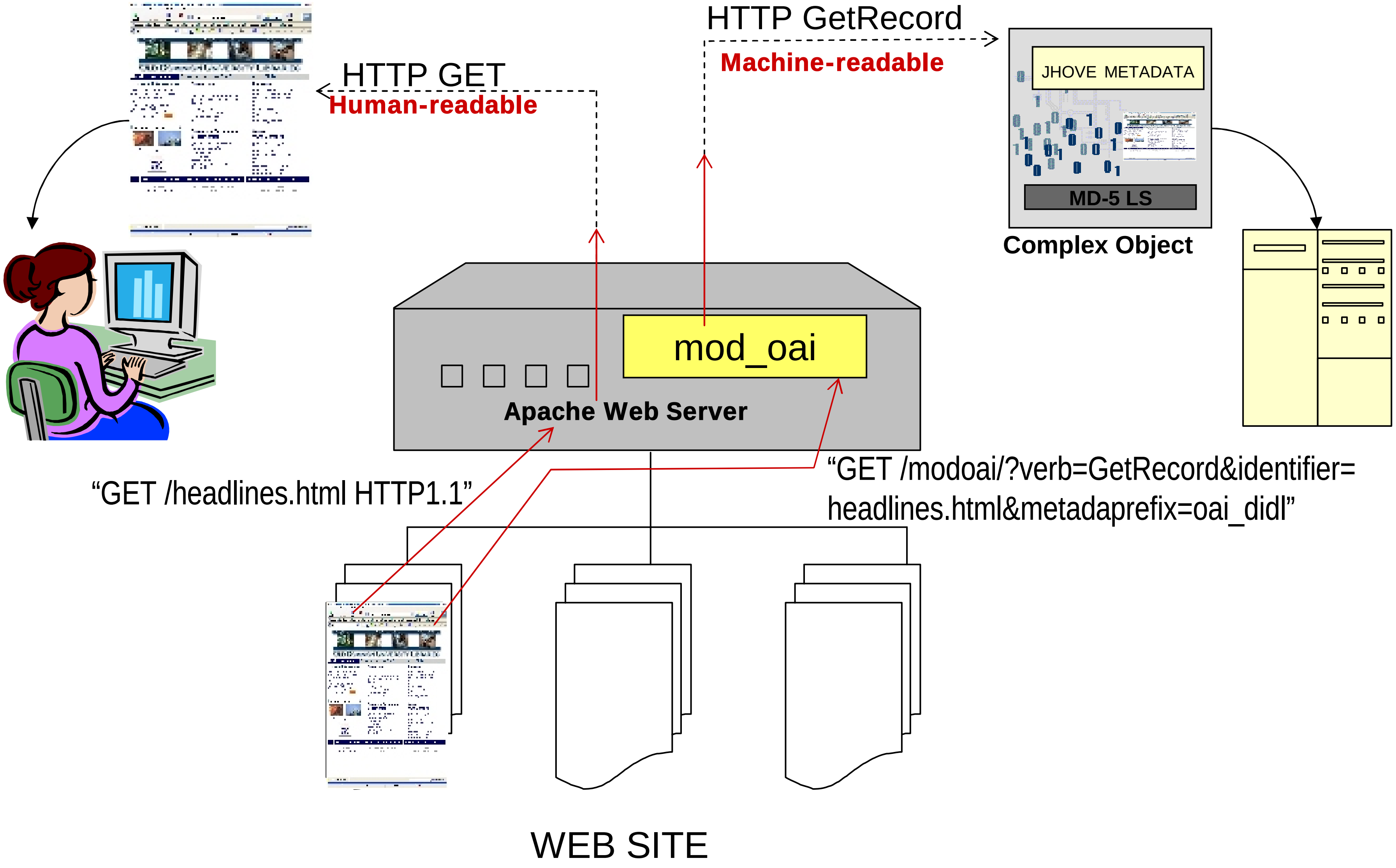
⇒ The *sitemap* standard is intended to list the resources at a site



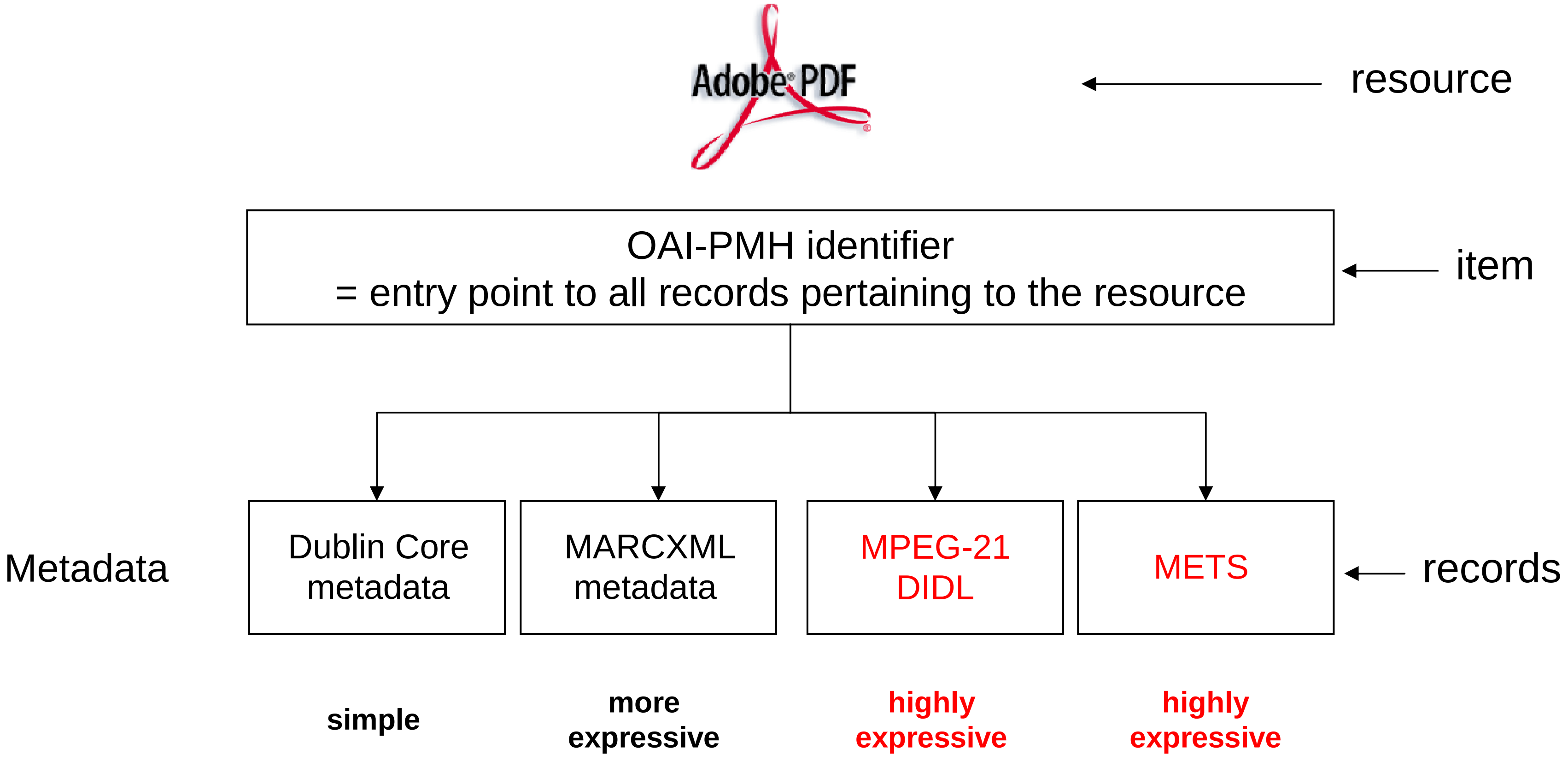
Introduction to Web Archiving

Marc Spaniol

HTTP GET vs. OAI-PMH GetRecord

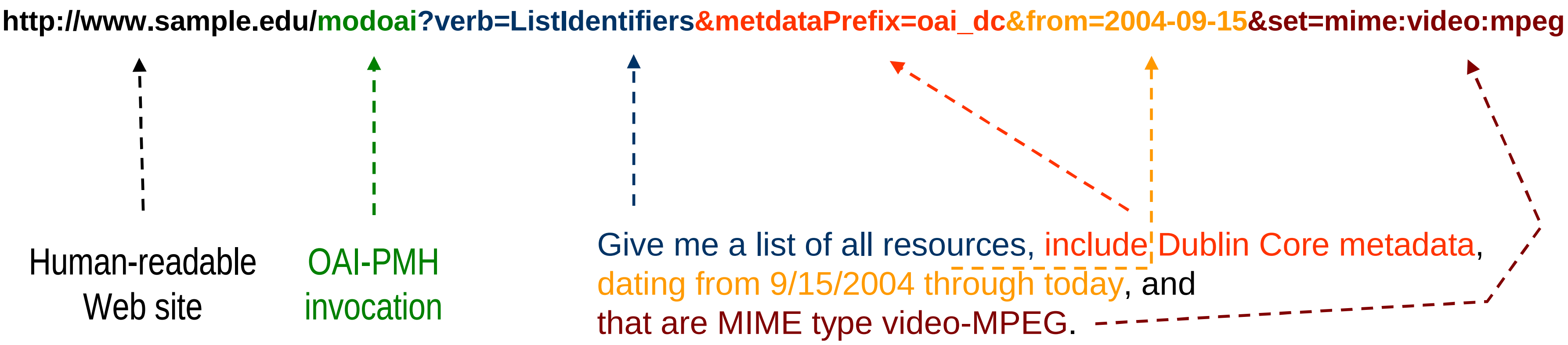


OAI-PMH data model



OAI-PMH Syntax

	Verb	Function
Repository metadata	Identify	Repository description
	ListMetadataFormats	Supported metadata formats
	ListSets	Sets defined by repository
Harvesting calls	ListIdentifiers	Unique IDs contained in repository
	ListRecords	Listing of *n* records
	GetRecord	Listing of a single record



http://www.sample.edu/modoai?verb=ListIdentifiers&metdataPrefix=oai_dc&from=2004-09-15&set=mime:video:mpeg

Human-readable Web site

OAI-PMH invocation

Give me a list of all resources, include Dublin Core metadata, dating from 9/15/2004 through today, and that are MIME type video-MPEG.

Introduction to Web Archiving

Marc Spaniol

Exemplary Application of OAI-PMH

Three from Tivoli

Takeda-Fan-Album 8/17

FANtastisch

ZU JEDEM HEIMSPIEL:
Gewinnen auf
www.takeda.de

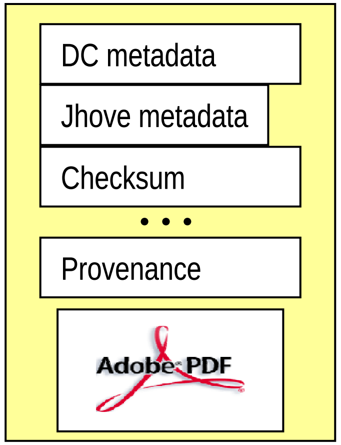
Ebenso einzigartig wie die Alemannia sind ihre Fans. Zum Beispiel Boris, Marc und Markus: Spätestens seit sie im Pokalfinale 2004 die Originalhose von George Mbwando ergattern konnten, kennt man die Clique im X-Block. Was die drei und alle anderen Fans verbindet, sind die Liebe und die Leidenschaft für Schwarz-Gelb. Auch wir von Takeda Pharma teilen diese Leidenschaft. Und stellen deshalb hier echte Alemannia-Fans vor. Weitere Fan-Porträts sowie Infos über unser Gewinnspiel finden Sie auf **www.takeda.de**.

EUREGIO PARTNER

Takeda Pharma

- Official Alemannia Aachen fan leaflet
- No. 8, Season 2005/2006
- …

http://www.takeda.de/unternehmen/pdf/fantastisch/pdf8_17.pdf
encoded as an MPEG-21 DIDL

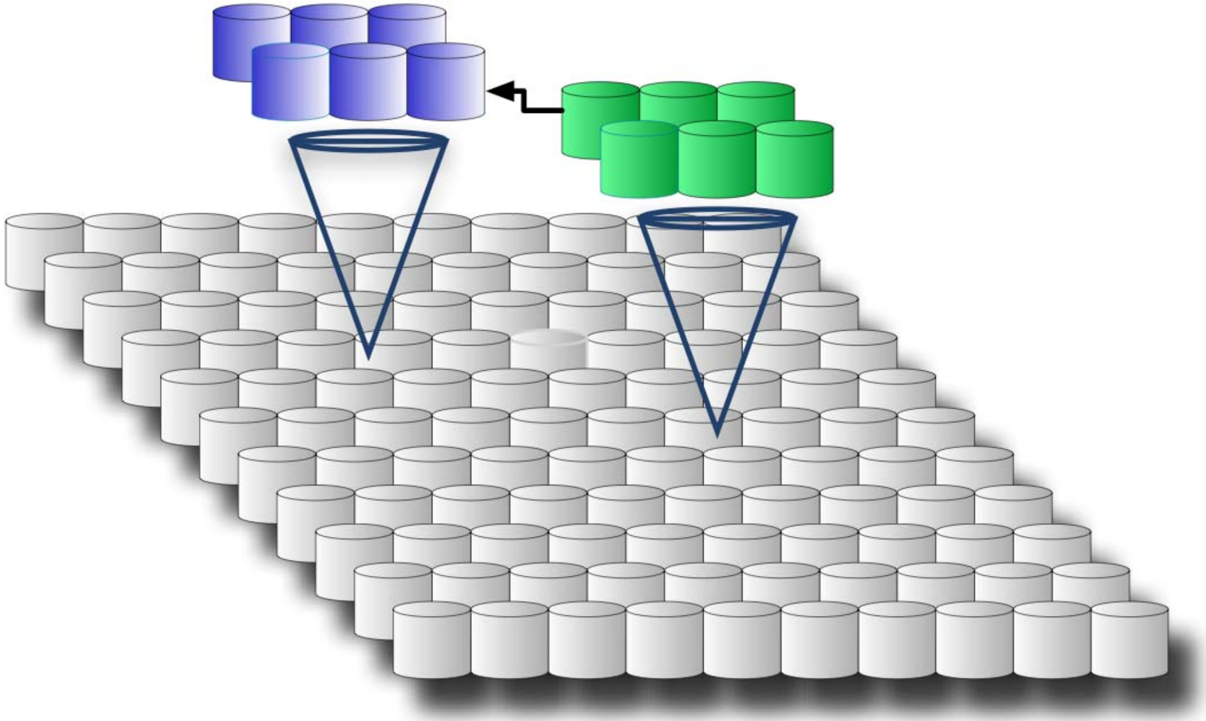


```
<didl> <metadata source="jhove">...</metadata>
    <metadata source="file">...</metadata>
    <metadata source="essence">...</metadata>
    <metadata source="grep">...</metadata> ...
    <resource mimeType="application/pdf"
    identifier="http://www.takeda.de/unternehmen/
    pdf/fantastisch/pdf8_17.pdf"
    encoding="base64">
    SADLFJSALDJF...SLDKFJASLDJ </resource>
</didl>
```

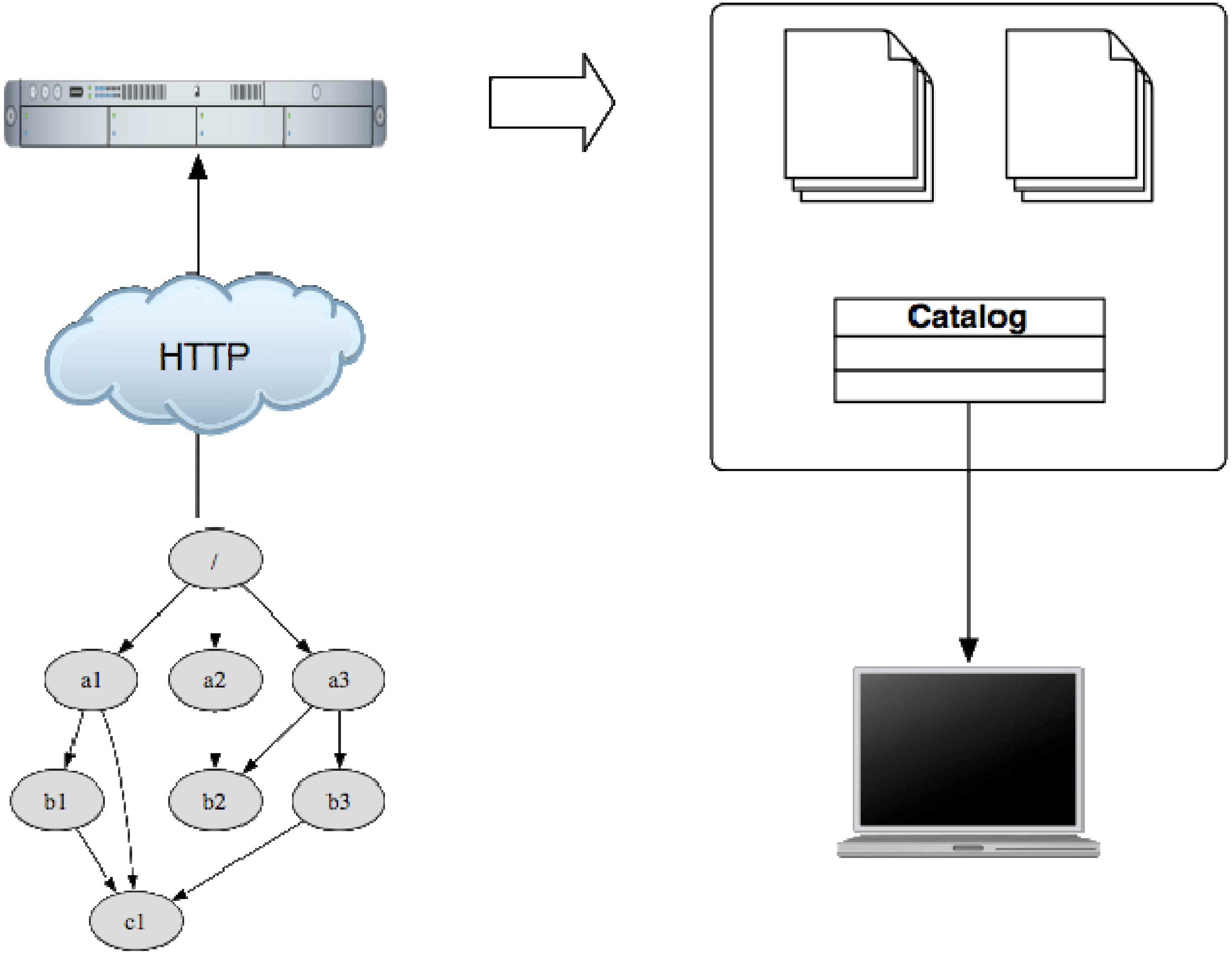
- Resource and metadata packaged together as a complex digital object represented via XML wrapper
- Uniform solution for simple & compound objects
- Unambiguous expression of locator of datastream
- Disambiguation between locators & identifiers
- OAI-PMH datestamp changes whenever the resource (datastreams & secondary information) changes

Archiving: Web Archives Grid

- Many “connected” servers
- WARC files spread among several servers
- Indexing of WARC files for access by URL and date



Hosting: Non-Web Archive

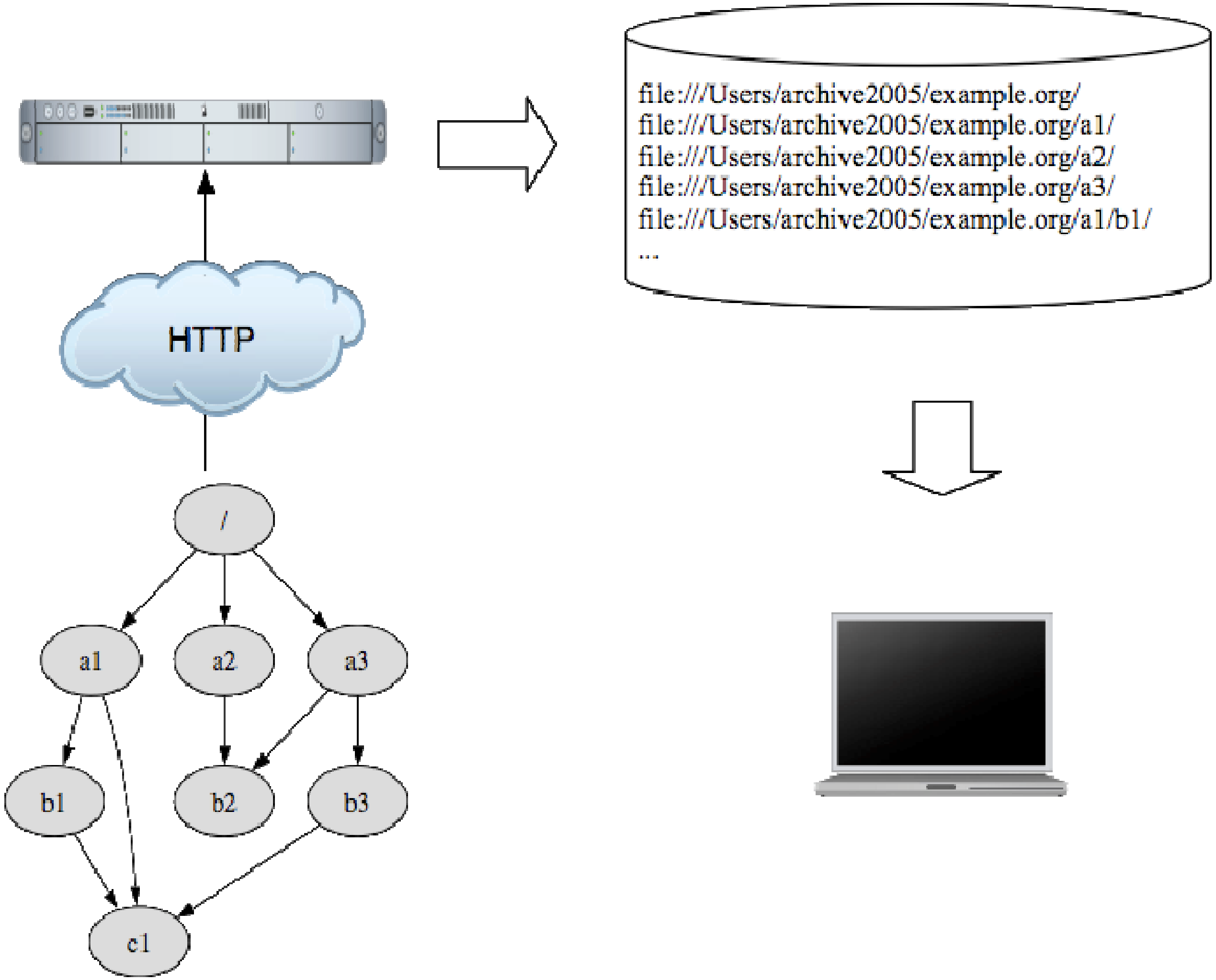


Non-Web Archive Summary

- Benefits
  - + Designed for archiving of specific (non-Web) collections
  - + Potentially fast data access

- Drawbacks
  - - Cataloging (usually) does not resemble hyperlink structure
  - - Implementation cost for cataloging logic
  - - Special search interface required

Hosting: Local File Navigation



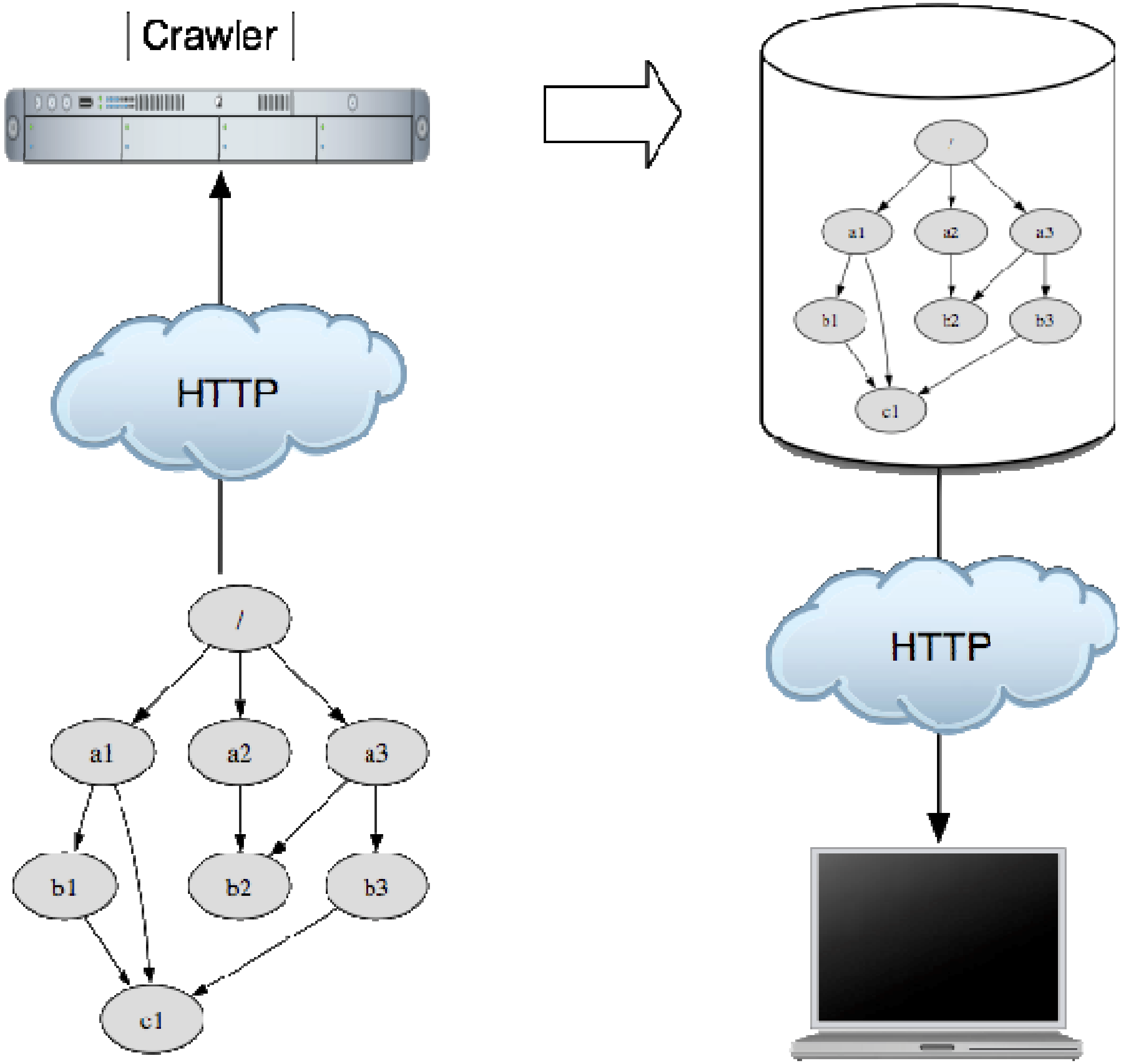
file:///Users/archive2005/example.org/
file:///Users/archive2005/example.org/a1/
file:///Users/archive2005/example.org/a2/
file:///Users/archive2005/example.org/a3/
file:///Users/archive2005/example.org/a1/b1/
...

Local File Navigation Summary

- Benefits
  - + Cheap
  - + Simple
  - + No additional infrastructure needed
  - + Fast

- Drawbacks
  - - Limited accessibility
  - - Small scale only
  - - Links are converted in relative ones
  - - Copying only

Hosting: Web-served Archive



Introduction to Web Archiving

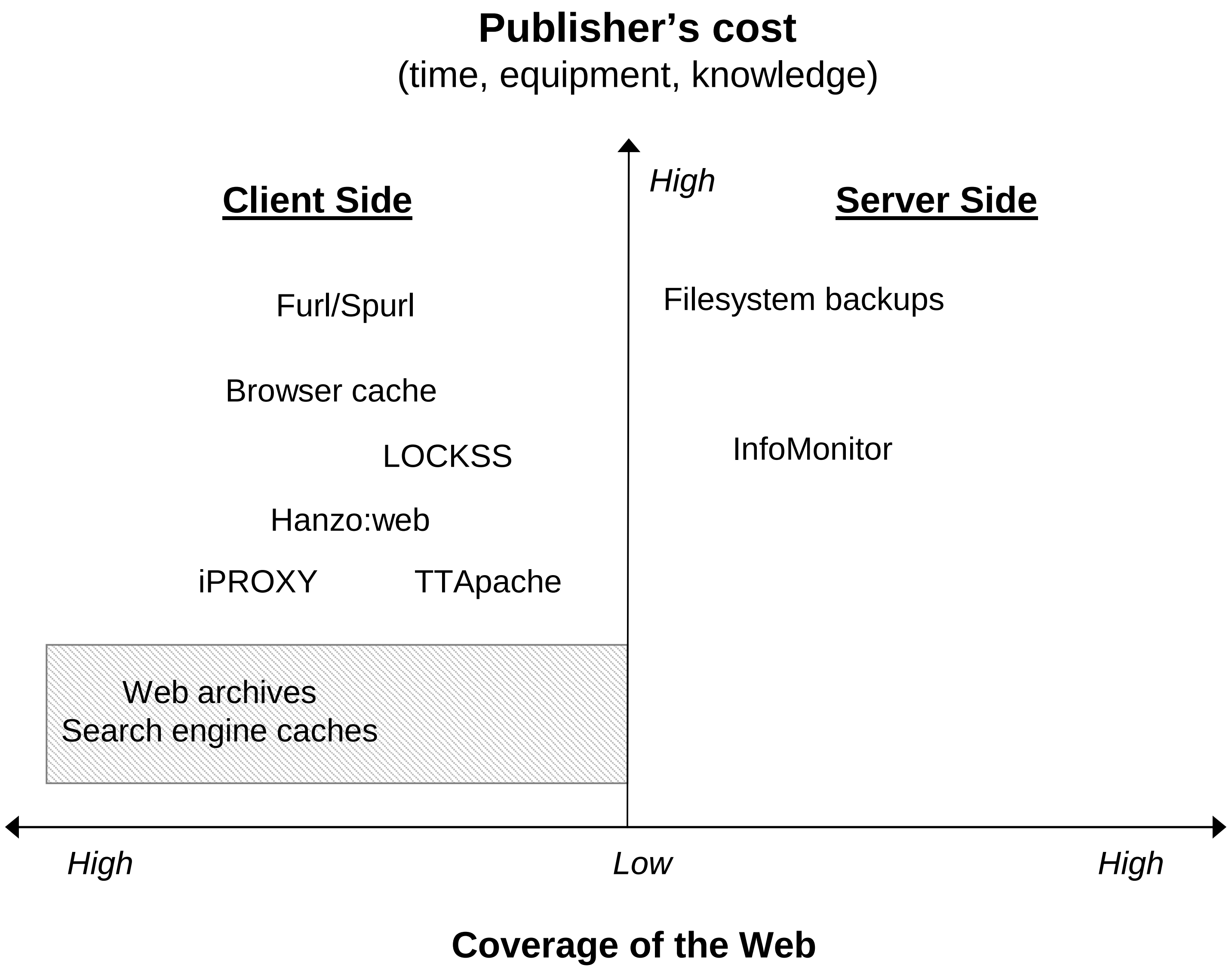
Marc Spaniol

Web-served Archive Summary

- Benefits
  - + Realistic “look&feel”
  - + Convenient navigation
  - + Time-travel also for non-technical experienced users possible

- Drawbacks
  - - Web server needed
  - - WARC/ARC file access required
  - - Indexing tool for WARC/ARC files necessary
  - - Time consuming sequential reads of WARC/ARC files

Cost of Web Archiving



Publisher's cost

(time, equipment, knowledge)

High

Client Side

Server Side

Furl/Spurl

Filesystem backups

Browser cache

LOCKSS

InfoMonitor

Hanzo:web

iPROXY

TTApache

Web archives
Search engine caches

High

Low

High

Coverage of the Web

Summary

- Web archiving is different from Web indexing
- Archiving crawlers
  - Do not aim at efficiency or freshness
  - Target at authenticity, coherence and durability
- Important aspects of Web archiving
  - Scope of archiving requires a clear definition
  - Seeds need to be carefully selected
  - Capturing of all URIs on a site and streaming media is hard
  - Preservation of hidden or dynamically generated contents is almost impossible
  - Pages may be orphaned intentionally or accidentally
  - Sitemaps rarely exist
  - WARC file processing is the bottleneck in retrieval
  - Capturing takes a long time (!!!) and contents may not fit to each other

References

[Heri09] Heritrix: “Glossary”.
http://crawler.archive.org/articles/user_manual/glossary.html
[last access: May 27, 2009]

[Masa06] J. Masanès: “Web Archiving”. Springer, New York, Inc., Secaucus, NJ, 2006.

[MKSR04] G. Mohr, M. Kimpton, M. Stack and I. Ranitovic: “Introduction to Heritrix, an archival quality Web crawler”. 4th International Web Archiving Workshop (IWAW'04), 2004.
http://crawler.archive.org/Mohr-et-al-2004.pdf
[last access: May 27, 2009]

[NCSm06] M. Nelson, F. McCown, J. Smith Thinking: “Differently About Web Page Preservation”.
http://www.loc.gov/today/cyberlc/feature_wdesc.php?rec=3896
[last access: May 27, 2009]