Towards Self-Validating Knowledge-Based Archives*

Bertram Ludäscher
Richard Marciano
Reagan Moore
{ludaesch,marciano,moore}@sdsc.edu

San Diego Supercomputer Center
University of California, San Diego

*sponsored by the National Archives and Records Administration and Advanced Research Projects Agency (NARA)

Digital Archives

- Problem
  - How to achieve long-term preservation of information (for the archivist: "records") and sustained access?
- Challenges and Opportunities
  - fight archives' obsolescence (*in the presence of | with*) …

  rapidly changing storage, data formats, software environment, hardware, …
- Approaches
  - "Time out" (do nothing: assume hardware, software, data formats, etc. all work 400 years from now ...)
  - Emulation (emulate hardware and software infrastructure)
  - Migration (migrate to new infrastructure)
- Factors
  - What do you need to archive? (records, data, programs, … ?)

  ==> determine usefulness and cost of emulation vs. migration

  archival of electronic records => data-centric => **migration**

Src: [Ludaescher-et-al, RIDE'01]

2

What is it That We Try to Archive??

- What constitutes a record?
  - … *beats me...*
- … but there are hierarchies of information / abstractions:
  - data ... information ... knowledge ... wisdom?
  - instance ... schema ... model ... metamodel ... metametamodel ...
  - object serialization ... data structure ... data model ... meta model …
- What is the nature of the information?
  - data …….. functions/programs
  - extensional data …….… intensional/virtual/derived data (facts/rules)
- Managing complexity using layers
  - protocol stacks (e.g. ISO/OSI, "SemanticWeb", Semantic Mediation)

  =*going up*==> abstract, correlate, aggregate, index, … the lower levels

Src: [Ludaescher-et-al, RIDE'01]

3

Archival Processes and Functions

- Data Submission/Accessioning:
  - ***loop***: information producer <==> *"archival engineer"* (ok: *archivist*)
- Ingestion:
  - a sequence of information preserving transformations is applied to submitted "raw data" => *ingestion network*
- Migration:
  - ... as time goes by ...
  - ... migrate to new physical media, maybe data formats, information model ...
  - "easy migration" <=> "good" archival format & model
- Instantiation/Access:
  - revive/reanimate the archive => *queryable collection/database*
- Goal: preserve information!

(ok: just "*records*" ...)

Src: [Ludaescher-et-al, RIDE'01]

4

Archival Example: Senate Collection

- What you see:

```
**** S. 345
                                        DATE INTRODUCED: 02/03/1999
SPONSOR: Allard
                                OFFICIAL TITLE
A bill to amend the Animal Welfare Act to remove the limitation
that permits interstate movement of live birds, for the purpose
of fighting, to States in which animal fighting is lawful.
                                LATEST STATUS
Feb  3, 1999 Read twice and referred to the Committee on
             Agriculture.
```

- … is maybe **NOT** what you get (a not so well documented format):

^@^@y^K^@^@\206^K^@^@Ê^K^@^@Ô^K^@^@^@^L^@^@^N^L^@^@ü^L^@^@\202^L^@^@È^L^@^@Ò^L^@^@ÿ\
^L^@^@^M^M^@^@j^M^@^@w^M^@^@»^M^@^@Æ^M^@^@ô^M^@^@^B^N^@^@\203^N^@^@÷ëßÓëßÇÂ¹Â¹Â®˜®Â\
Â\230Â\230Â Â\230Â\230Â®˜®Â Â\230Â\230Â Â\230Â\230Â Â\230Â\230Â Â\230Â\230Â Â\230Â\
^N6^H\201OJ^C^@QJ^C^@]^H\201^@^N5^H\201OJ^C^@QJ^C^@\^H\201^@^K^B^H\201OJ^C^@QJ^C^@^\
...
^
ction sent to the House.^M^M**** S. 345^MDATE INTRODUCED: 02/03/1999^MSPONSOR: Alla\
rd^MOFFICIAL TITLE^MA bill to amend the Animal Welfare Act to remove the limitation\
that permits interstate movement of live birds, for the purpose of fighting, to St\
ates in which animal fighting is lawful.^MLATEST STATUS^MFeb 3, 1999 Read twice \
and referred to the Committee on Agriculture.^M^M**** S. 387^MDATE INTRODUCED: 02/0\
8/1999^MSPONSOR: McConnell^MOFFICIAL TITLE^MA bill to amend the Internal Revenue Co\
d

Src: [Ludaescher-et-al, RIDE'01]

5

Senate Collection Example

- Rich Text Format (a ***documented*** Microsoft format) :

\pard\par^M
\pard\b **** S. 345\b0\par^M
\pard\qr DATE INTRODUCED: 02/03/1999\par^M
\pard SPONSOR: Allard\par^M
\i\qc OFFICIAL TITLE\i0\par^M
\pard A bill to amend the Animal Welfare Act to remove the limitation that permits \
interstate movement of live birds, for the purpose of fighting, to States in which \
animal fighting is lawful.\par^M
\i\qc LATEST STATUS\i0\par\pard^M
\pard\plain \fi-1900\li1900\nowidctlpar\adjustright{Feb 3, 1999\tab Read twice and\
referred to the Committee on Agriculture.\par}^M
\pard^M

- … can be wrapped into XML:

**** S. 345
DATE INTRODUCED: 02/03/1999
SPONSOR: Allard
OFFICIAL TITLE
A bill to amend the Animal Welfare Act to remove the lim\
itation that permits interstate movement of live birds, for the purpose of fighting\
, to States in which animal fighting is lawful.
LATEST STATUS
<string>Feb 3, 1999&tab;Read twice and referred to the Committee on Agriculture\
.</string>

Src: [Ludaescher-et-al, RIDE'01]

6

Senate Collection Example

- … the XML can be ***lifted*** from the ***presentation*** level:

```
<p bold="off">**** S. 345</p>
<p align="right" bold="off">DATE INTRODUCED: 02/03/1999</p>
<p bold="off">SPONSOR: Allard</p>
<p align="center" bold="off" italic="off">OFFICIAL TITLE</p>
<p bold="off" italic="off">A bill to amend the Animal Welfare Act to remove the lim\
itation that permits interstate movement of live birds, for the purpose of fighting\
, to States in which animal fighting is lawful.</p>
<p align="center" bold="off" italic="off">LATEST STATUS</p>
<p><string>Feb  3, 1999&tab;Read twice and referred to the Committee on Agriculture\
.</string></p>
<p></p>
```

- … to the ***information*** **level**:

```
<bill name="S.345">
   <committees>
     <committee>SENATE: AGRICULTURE</committee>
   </committees>
   <date_introduced>02/03/1999</date_introduced>
   <latest_status_list>
        <latest_status>   <ls_date>Feb  3, 1999</ls_date>
                          <ls_txt>Read twice and referred to the Committee on Agriculture</ls_txt>
        </latest_status>
   </latest_status_list>
   <official_title>A bill to amend the Animal Welfare Act to remove the limitation that permits interstate movement of live birds, for
the purpose of fighting, to States in which animal fighting is lawful.</official_title>
   <sponsor>Allard, Wayne [CO]</sponsor>
</bill>
```

Src: [Ludaescher-et-al, RIDE'01]

7

XML as an Archival Format

- Information level "schema" as an XML DTD:

```
<?xml version="1.0" encoding="UTF-8"?>
<!ELEMENT bills (bill*)>
<!ELEMENT bill  ( abstract?, committees?, congressional_record?, cosponsors?, date_introduced?,
                  digest?, latest_status_list?, official_title?, sponsor?, statement_of_purpose?,
                  submitted_by?, submitted_for?)>
<!ATTLIST  bill_name CDATA #REQUIRED>
<!ELEMENT committees (committee*)>
<!ELEMENT cosponsors (cosponsor*)>
<!ELEMENT digest (#PCDATA)>
<!ELEMENT latest_status_list (latest_status*)>
<!ELEMENT latest_status (ls_date, ls_txt)>
<!ELEMENT abstract (#PCDATA)>
<!ELEMENT committee (#PCDATA)>
<!ELEMENT congressional_record (#PCDATA)>
<!ELEMENT cosponsor (co_name)>
<!ELEMENT co_name (#PCDATA)>
<!ATTLIST  co_name a-date CDATA #IMPLIED>
<!ELEMENT date_introduced (#PCDATA)>
…
<!ELEMENT  statement_of_purpose (#PCDATA)>
<!ELEMENT submitted_by (#PCDATA)>
<!ELEMENT submitted_for (#PCDATA)>
```

Src: [Ludaescher-et-al, RIDE'01]

8

Open Archival Information System (OAIS) Information Model

An AIP (archival information package) contains

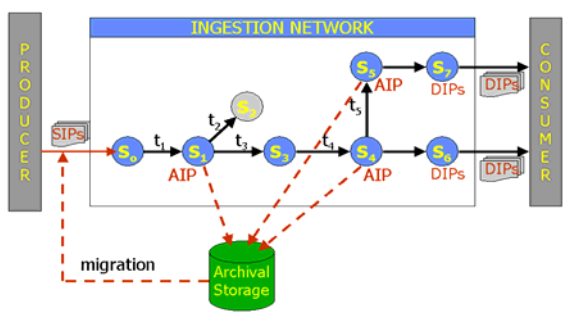
- content information (CI) (represented as info_objects), and
- preservation description information (PDI)

(A)IP *(archival) information package =*
[DI *descriptive information*
[PI *packaging information (ISO-9660 for CD directories)*
[CI *content information*
PDI *preservation description information =*
[PR *provenance (origin, processing history)*
CON *context (relation to external information)*
REF *reference (identifies the CI, e.g., ISBN, URI)*
FIX *fixity (e.g., checksum over CI)*
]]]]

Src: [Ludaescher-et-al, RIDE'01]

9

Archival Ingestion Networks



Transformation ***t*** is ***information preserving***, if it is *reversible*, i.e., if there is an inverse ***t_inv***, s.t., for all ***d*** in ***dom(t)***:

t_inv(t(d)) = d .

Example:

- {d1, d2, …} ⊆ HTML ==***wrapper***==> {d1', d2', …} ⊆ XML
- {d1', d2', …} ==***inverse wrapper*** (XSLT)==> {d1'', d2'', …} ⊆ HTML

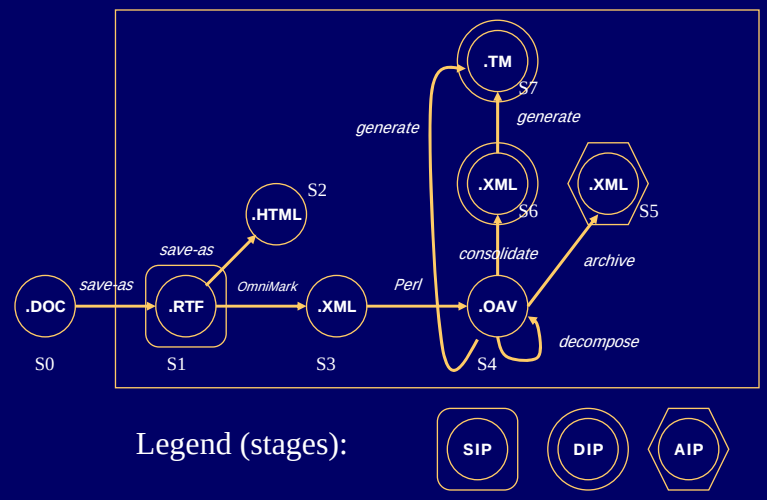
• asking for exact inverse often not practical

=> consider e.g. normalized HTML or restrict to higher level representations

Src: [Ludaescher-et-al, RIDE'01]

10

Ingestion Network: Senate Collection



Src: [Ludaescher-et-al, RIDE'01]

11

From XML-Based to Knowledge-Based Archives

- Collection-based archival with XML: save data "as is" plus...
  - ... separate content from presentation
  - ... tag your data (take a lift in the info hierarchy)
  - ... use a self-describing, semistructured data format (XML)
- Knowledge-based archival: now add ...
  - ... conceptual level information
  - ... integrity constraints
  - ... explanations/derivation rules:
    - archiving only results y=f(x) vs. archiving the rules/function "f" (e.g. f = “the Florida procedure”...)

  => employ knowledge representation languages

Src: [Ludaescher-et-al, RIDE'01]

12

Knowledge-Based Archival: Senate Example

Data provider says:

"Please archive all records of legislative activities of the 106th senate!"

Integrity constraints, eg:

(1) *{senators_with_file} = UNION (sponsor, cosponsors, submitted_by)*

(2) *{senators} = {sponsors} = {co-sponsors}*

Violation:

– the *rhs* is a SUPERSET of the *lhs* !

Exceptions:

– (Chafee, John), (Gramm, Phil), (Miller, Zell)

(Possible) Explanations:

– senators who joined (Zell), passed away (Chafee), were forgotten (Gramm)!?

Checking ICs:

IF *sponsor(X), not senator(X)* **THEN** *ADD(exception_log, missing_senator_info(X))*

IF ***condition*** **THEN** ***action***

Action = LOG, WARN, ABORT, ...

Src: [Ludaescher-et-al, RIDE'01]

13

Maximizing "Self-Containedness" ...

- Self-validating archives: add ...
  - ... "*executable* knowledge" (=rules)
  - *"helping (bugging?) the data provider"*

  => add the functionality and meaning of DTD (+Schema+IC+...) validation to the AIP

  => package the validator!
- Self-instantiating archives: add ...
  - ... "*executable* ingestion process"
  - *"helping the archival engineer (aka archivist)"*
  - *…here is: looking over your shoulder…*

  => add the functionality of database transformations to the AIP

  => package the transformers!
- BUT packaging validators and transformers increases infrastructure **dependence**!

Src: [Ludaescher-et-al, RIDE'01]

14

Maximize "Self-Containedness" ...
…While Minimizing Infrastructure Dependence

Basic Idea: use a language of *executable specifications* for self-validation and self-instantiation!

=> Use *"Bootstrapping"* for Self-Validating & Self-Instantiating Archives

Example: DTD Validator in Logic (F-Logic, Datalog,…)

```
% specify <!ELEMENT X   (Y,Z)>
false IF P:X, not (P1.X):Y.
false IF P:X, not (P2.X):Y.
false IF P:X, not P[_-> _].
false IF P:X[N->_], not N=1, not N=2.
...
```

Src: [Ludaescher-et-al, RIDE'01]

15

XML Extensions as General Constraint Languages

Assume an archival language *A* for IPs (e.g. *A*=XML)

Def. *C* is a *constraint language* for *A*, if for all $\varphi \in C$, the set of *valid archives* $V(\varphi) = \{a \in A \mid a \models \varphi\}$ is decidable.

Example: C = `XML_DTD`, φ = `Senate_DTD`

Def. *C'* *subsumes* *C* (*C'* ϕ *C*) w.r.t. *A*, if for all $\varphi \in C$ there is an encoding $enc(\varphi) \in C'$ s.t. for all $a \in A$:

$$a \models \varphi \quad \text{iff} \quad a \models enc(\varphi)$$

Proposition:

- `XML_Schema` ϕ `XML_DTD`
- `{F_Logic',Datalog'}` ϕ `XML_DTD`

Src: [Ludaescher-et-al, RIDE'01]

16

Summary: Towards Bootstrapping Knowledge-Based Archives

Baron von Münchhausen, pulling himself out of the swamp

- enable addition of *semantic annotations* ("knowledge") via ***logic rules*** to AIPs
- add ***executable specifications*** of **semantics**
  => AIP += KP (*knowledge package*, i.e., *logic ules*)

  => **self-validating archive**
- add ***executable specifications*** of the **ingestion network**
  => AIP += IN (*ingestion network, ...more logic rules*)

  => **self-instantiating archive**

=> ***bootstrapping knowledge-based archive*** with DTD/Schema/IC validation and ingestion transformations **all expressed in a declarative logic program**

- Outlook from the 2do list: build a prototype **BARON** = **B**ootstrapping **A**rchive of **R**ules, **O**ntologies, and Ingestion **N**etworks

Src: [Ludaescher-et-al, RIDE'01]

17

References

- Towards Self-Validating Knowledge-Based Archives, Bertram Ludäscher, Richard Marciano, Reagan Moore, *11th Workshop on Research Issues in Data Engineering* (RIDE), Heidelberg, IEEE Computer Society, April 2001, SDSC TR-2001-1, January 18, 2001.
- Knowledge-Based Persistent Archives, Reagan Moore, SDSC TR-2001-7, January 18, 2001
- The Senate Legislative Activities Collection (SLA): a Case Study Infrastructure Research to Support Preservation Strategies, Richard Marciano, Bertram Ludäscher, Reagan Moore, SDSC TR-2001-5, January 18, 2001
- Reference Model for an Open Archival Information System (OAIS), Draft Recommendation, Consultative Committee for Space Data Systems, CCSDS 650.0-R-1, May 1999.
- Digital Rosetta Stone: A Conceptual Model for Maintaining Long-term Access to Digital Documents, Alan R. Heminger, Steven B. Robertson

Src: [Ludaescher-et-al, RIDE'01]

18

ADDITIONAL MATERIAL AHEAD …

Src: [Ludaescher-et-al, RIDE'01]

19

Collection-Based Archival with XML

- Archival Formats Desiderata
  - standardized, open, as simple as possible, …

  ==> self-contained and self-describing

  ==> XML provides a good framework for archival
- Data/Instance Level: records/objects/tuples

  ==> content information (CI)
- Schema/Class Level:
  - collection structure & metadata, types

  ==> packaging information (PI) and descriptive information (DI)
- Missing in Action...
  - conceptual level information: relationships between collection attributes/classes, integrity constraints, derived knowledge, …
  - parts in CON, PI, but need for knowledge packages (KPs)

=> Knowledge-Based Archival

Src: [Ludaescher-et-al, RIDE'01]

20

Getting your hands dirty with logic rules

- Some logic rules for reassembling the doc structure (lexical scopes) from the OAV (or rather AOV):

```
attr_interval(Attr, SID, Attr_val, LN, LN1) :-
        oav(Attr, (SID, LN), Attr_val),
        oav(Attr, (SID, LN1), _),
        LN1 > LN,
        not attr_between(Attr,SID,LN,LN1).

attr_between(Attr,SID,LN,LN1) :-
        oav(Attr, (SID, LN), _),
        oav(Attr, (SID, LN1), _),
        oav(Attr, (SID, LN2), _),
        LN < LN2,  LN2 < LN1.
```

Src: [Ludaescher-et-al, RIDE'01]

21

Summary: what is the declarative (logic) approach?

- Use of declarative database and knowledge representation formalisms for...
  - adding knowledge packages to AIPs:
    - capture context known at the time of archival using conceptual models of collections, integrity constraints, virtual relations, …
  - applying them at *ingestion* (aka: *bringing-in*), *migration*, and *instantiation/access* time

    (= wrapping, transforming, querying collections)

Src: [Ludaescher-et-al, RIDE'01]

22