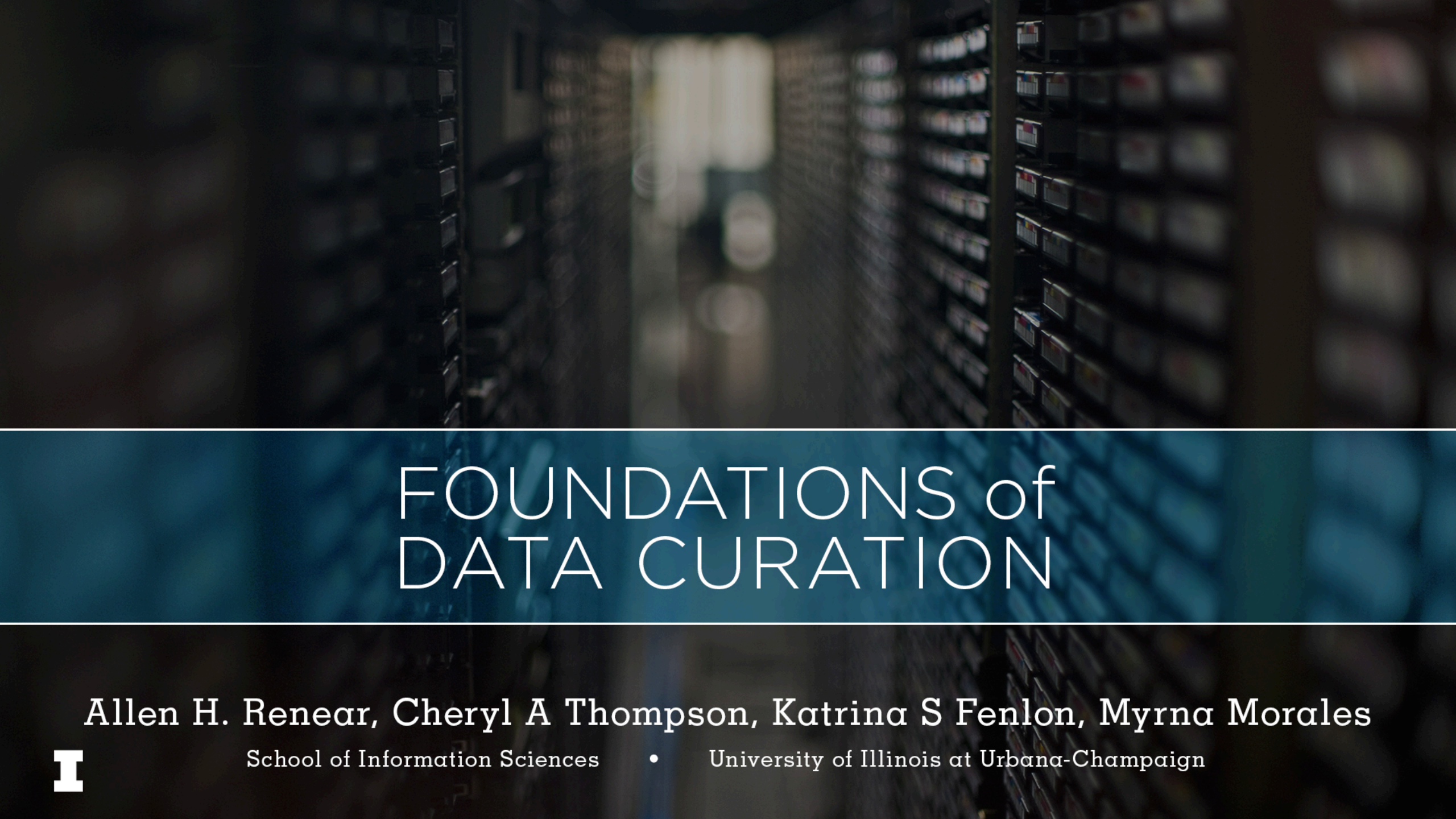


FOUNDATIONS of DATA CURATION

Allen H. Renear, Cheryl A Thompson, Katrina S Fenlon, Myrna Morales

School of Information Sciences • University of Illinois at Urbana-Champaign

DATA MODELS: TREES

6

IMPLEMENTING THE SOLUTION: XML

Implementing the solution: XML

- XML: a schema language (or: a meta-grammar)
    - XML schemas
    - XML documents
- Example of an XML schema (DTD)
- Example of XML processing
- XML tools
- XML languages

XML

An XML document uses a defined set of delimiters with arbitrary element names and attribute value pairs to nest spans of text.

A *well-formed* XML document fits a formal grammar (along with other constraints) that ensures the document can be parsed as a tree by an XML parser.

> NB: a well-formed XML document need not have a schema that defines the element vocabulary and grammar;
>
> > it may use arbitrary element, attribute, and value names and arrange text objects in any way that does not violate the tree data structure.

The two main things in the XML world

Schemas [such as Document Type Definitions (DTDs)]

- One for each document type (class, category, genre)
- Defines a markup language for document structures by specifying its vocabulary and syntax (grammar)
- What elements can occur in documents of a particular type, what patterns these elements may form, what other information can be included about these elements?

Document Instances

- Particular documents, marked up with a markup language that meets well-formedness constraints, and, perhaps, also meets the constraints of a relevant schema.

A well-formed XML document

```
<anthology>
  <poem>
    <heading>THE SICK ROSE</heading>
    <stanza>
      <line>O Rose thou art sick.</line>
      <line>The invisible worm,</line>
      <line>That flies in the night</line>
      <line>In the howling storm:</line>
    </stanza>
    <stanza>
      <line>Has found out thy bed</line>
      <line>Of crimson joy:</line>
      <line>And his dark secret love</line>
      <line>Does thy life destroy.</line>
    </stanza>
  </poem>
  <!-- more poems go here -->
</anthology>
```

Example from the Text Encoding Initiative (P5)

Schemas for Trees

This is an XML Document Type Definition (DTD), defining an XML document type

```
<!ELEMENT anthology (poem+)>
<!ELEMENT poem (title?, stanza+)>
<!ELEMENT title (#PCDATA)
<!ELEMENT stanza (line+) >
<!ELEMENT line (#PCDATA) >
```

This schema specifies element vocabulary and grammar

The DTD schema language is based on (extended) Backus Naur Form (BNF) grammars

Some XML schema languages provide additional validation and constraints on content, including data typing.

Another DTD

```
<!ELEMENT poem    (title, author? verse) >
<!ATTLIST poem
              editor    CDATA             #REQUIRED>
<!ELEMENT verse   (stanza+)>
<!ELEMENT stanza  (line+)>
<!ELEMENT title   (#PCDATA | italic | persname)*>
<!ELEMENT author  (#PCDATA)
<!ATTLIST author
              sex       (male | female)   #IMPLIED
              dates     CDATA             #IMPLIED
              bio       IDREF             #IMPLIED>
<!ELEMENT line    (#PCDATA | italic | persname)
<!ATTLIST line
              lang      CDATA             "ENGLISH">
<!ELEMENT italic  (#PCDATA)>
<!ELEMENT persname(#PCDATA)>
```

Another XML document, with attribute/value pairs

```
<!DOCTYPE text SYSTEM ”poem.dtd">

<poem editor=“Sara Porter”>


<title>Terence</title>
<author person=“N320”>A. E. Houseman</author>

<verse>
<stanza>
<line>Terence this is stupid stuff  </line>
<line>you eat your victuals fast enough</line>
<line>there can’t be much amiss ‘tis clear</line>

</stanza>
<stanza>
[…]
<line  lang=“latin”>The old lie: </line>
<line>  in vino veritas  </line>
</stanza>
</verse>

</poem>
```

Valid XML Documents

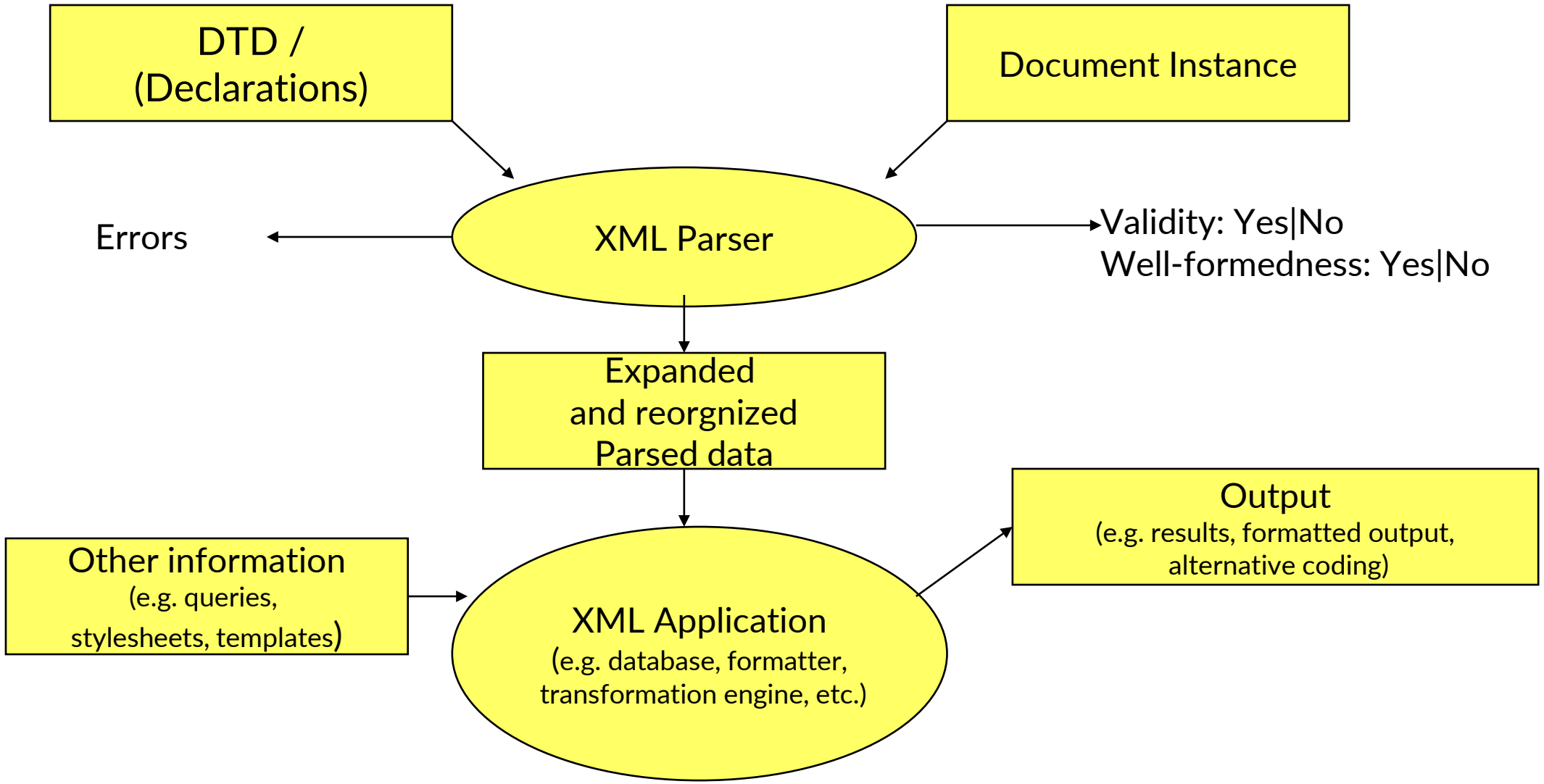
A document instance is *valid* with respect to some schema if it conforms to the declarations in that schema, which is to say, matches the grammar and other constraints.

> . . . nothing out of place, nothing missing, no attributes with values the wrong type, no references that fail, and so on.

A *validating parser* applies an XML schema to an XML document and determines whether or not the document conforms to the constraints specified in the schema.

All valid XML documents are well-formed, but not vice versa.

XML Processing



Some XML tools and schema languages

Two important XML transformation tools

XSLT: “a language for transforming XML documents into other XML documents” https://www.w3.org/TR/xslt

Xquery: “a standardized language for combining documents, databases, Web pages, and almost anything else” https://www.w3.org/XML/Query/

Two other XML schema languages

XML Schema (XSD): “the XML Schema Definition Language…offers facilities for describing the structure and constraining the contents of XML documents”

http://www.w3.org/XML/Schema

- A more complex schema language than DTDs, but does more than validate
- Written in XML
- Common for business applications

RelaxNG: “a schema language for XML” http://relaxng.org/

- Similar expressiveness to XSD, with simpler syntax
- Less commercial application support

Some important XML languages for documents

You should be familiar with these. Please explore the websites.

XHTML: “a family of current and future document types and modules that reproduce, subset, and extend HTML”

https://www.w3.org/TR/xhtml1/

TEI: Text Encoding Initiative – “a standard for the representation of texts in digital form”

http://www.tei-c.org/index.xml

JATS: Journal Article Tag Suite “defines a set of XML elements and attributes for tagging journal articles”

https://jats.nlm.nih.gov/

XML Languages and Interchange

There are an enormous number of XML markup languages.
See https://en.wikipedia.org/wiki/List_of_XML_markup_languages

Most of these languages are not for text, but are interchange and preservation formats for structured data.

XML is an important preservation format. It use of simple ASCII text with inline tags, can be parsed without a schema, and if a schema is available can be validated ensuring a correct grammar (nothing missing, nothing out of place) and data typing.

FOUNDATIONS OF DATA CURATION (IS531)

Allen H. Renear, Cheryl A Thompson, Katrina S Fenlon, Myrna Morales
School of Information Sciences
University of Illinois at Urbana-Champaign

Includes material adapted from work by Carole Palmer, Melissa Cragin, David Dubin, Karen Wickett, Bertram Ludaescher, Ruth Duerr and Simone Sacchi.

Comments and corrections to: renear@illinois.edu.