# Automatic Metadata Extraction for Archival Description and Access

**WILLIAM UNDERWOOD**
**Georgia Tech Research Institute**

**Abstract:** The objective of the research reported is this paper is to develop techniques for automatically extracting metadata from electronic records that is necessary for automatically describing items, file units and records series and for supporting access to these records. Archival metadata and elements of descriptions include document type, date, author, addressee, and topic. The research results include the definition of documentary forms using context-free grammars, and a method for recognizing the documentary forms of textual e-records while simultaneously identifying document metadata.

## Introduction

This research project is addressing archivist's needs for automated decision support for archival description, access and review of Presidential e-records.[1] The Freedom of Information Act (FOIA) provides that citizens may request Presidential records 5-years after the end of an administration. In responding to FOIA requests, Archivists need to be able to search collections of records with high precision and recall. But at the time of responding to FOIA requests, archivists have not read all of the records, so cannot index the records and search on such attributes as person, organization and location names, topics, dates, author's and addressee's names and document types.

Archivists cannot describe a collection until the collection has been manually read and reviewed. With increasing volumes of electronic records, it may be decades or even centuries before new acquisitions are described.

This paper reports progress in the development of techniques for automatically recognizing document types and extracting metadata from e-records. This metadata can be used for indexing and searching collections of records by person, organization and location names, topics, dates, author's and addressee's names and document types, and for automatically describing items, file units and record series.

## Documentary Form and Record Types

The International Council of Archives in its standard for archival description defines a (documentary) *form* as "A class of documents distinguished on the basis of common physical (e.g., water colour, drawing) and/or intellectual (e.g., diary, journal, day book, minute book) characteristics of a document."[2] The standard also specifies that the names of forms be used in describing record series and titling records.

The National Archives and Records Administration's guideline for cataloging archival materials defines *specific records type* as "the intellectual format of the archival materials." The purpose of the specific records type is that it "enables users to search for archival materials by the types of document represented

[1] This research project is sponsored by the ERA Program of the National Archives and Records Administration and the Army Research Laboratory under Army Research Office Cooperative Agreement W911NF-06-2-0050.

[2] ICA, ISAD(G): General International Standard Archival Description, Second Edition, 1999, p. 11.

in the archival materials."[3] The guidelines also specify that specific records types be used in describing record series.

Figure 1 shows examples of the names of some of the specific documentary forms (record types) discovered in Presidential e-records.

| | | |
|---|---|---|
| Agenda | Job Application | Press Pool Report |
| Bar Chart | Mailing List | Press Release |
| Biography | Memo | Referral Memo |
| Briefing Memo | Minutes of Meeting | Resume |
| Decision Memo | National Security Directive | Schedule |
| Correspondence | Newsletter | Signature Memo |
| Diary | Nomination to Federal Office | Situation Report |
| Executive Order | Notes | Telephone Call Recommendation |
| Information Memo | Presidential Proclamation | Transcript of News Conference |

**Figure 1. Documentary Forms in Presidential Records**

The research questions are "How can documentary forms be more precisely defined so that they may be automatically recognized?" and "How can document metadata such as document date, author's and addressee's names, and topic be automatically extracted from a document?"

**Markup Languages and Document Types**

The Standard Generalized Markup Language (SGML)[4] uses a Document Type Definition (DTD) to define document form. A DTD specifies a set of elements, their relationships, and the tag set to markup the document. The Extensible Markup Language (XML) is a simpler subset of SGML.[5] Figure 2 shows a record from the Bush Presidential Records that has been manually marked up as an XML document.[6]

[3] NARA, Life Cycle Data Requirements Guide (LCDRG), March 3, 2008, p. 131.
[4] International Standards Organization, Standard Generalized Markup Language - ISO 8879.
[5] World Wide Web Consortium, Extensible Markup Language XML 1.0 (Fourth Edition), 16 Aug 2006.
[6] Bush Presidential Library, Bush Presidential Records, WHORM Subject File, Disasters-Natural, ID#324869.

```
<?xml version="1.0" standalone="no">
<!DOCTYPE memo SYSTEM "http://www.site/com/dtds/memo.dtd">
<memo> <header>
                  <date>April 27, 1992</date>

<for type="MEMORANDUM FOR"> <person>SAM SKINNER</person></for>
<from type="FROM:">     <person>EDE HOLIDAY</person></from>
<subject type="SUBJECT:"> <np>California Earthquake</np></subject>
</header>
<body>
<para>Attached is a situation report from FEMA on the northern
California earthquake. No deaths have been reported and 45
people are known to have suffered injuries. In addition, there
has been extensive property damage. While FEMA is awaiting a
request from the State before initiating any recovery activities,
a joint State/Federal preliminary damage assessment is likely to
begin today.</para>

<para>Director Stickney has requested that we forward the situation
report to you.</para>
</body></memo>
```

**Figure 2. A Record Manually Marked Up as an XML Document**

The structure of the document is described by pairs of XML tags that bracket content, for example, <date> April 27, 1992</date>. The second line of the XML document is a Document Type Declaration. It links the document file to a DTD that specifies the structure of the document. Figure 3 shows the DTD for the memorandum in Figure 2. The DTD specifies that a memo consists of a header element followed by a body element. The header consists of a sequence of date, for, from and subject elements. The body consists of a sequence of one or more paragraphs.

```
<!DOCTYPE memo [
<!ELEMENT       memo      (header, body)                                  >
<!ELEMENT       header    (date, for, from, subject)                      >
<!ELEMENT       date      (#PCDATA)                                       >
<!ELEMENT       for       (person)                                        >
<!ATTLIST       for       type      NMTOKENS     "MEMORANDUM FOR"       >
<!ELEMENT       person    (#PCDATA)                                       >
<!ELEMENT       from      (person)                                        >
<!ATTLIST       from      type      NMTOKEN        "FROM:"               >
<!ELEMENT       subject   (np)                                            >
<!ATTLIST       subject   type      NMTOKEN        "SUBJECT:"            >
<!ELEMENT       np        (#PCDATA)                                       >
<!ELEMENT       body      (para+)                                         >
<!ELEMENT       para      (#PCDATA)                                       >
]>
```

**Figure 3. External DTD for the Memorandum**

The concept of document structure as defined by a XML DTD is a formal model of the concept of the intellectual form of a document.

**A Method for Recognizing Documentary Forms and Extracting Document Metadata**

Legacy and current Presidential e-records are not XML documents, but e-records in proprietary file formats. However, it will be shown that it is possible to define, recognize and annotate the intellectual elements of a textual e-record, and that the structure of the intellectual elements of a particular documentary form can be defined with rules similar to those of an XML document type definition. This will enable the recognition of documentary forms and extraction of document metadata.

The process of automatically recognizing the document types of documents in proprietary file formats is outlined in Figure 4. The italicized phrases to the right of the downward pointing arrows indicate inputs and outputs of the numbered processing steps.[7]

↓ *document in proprietary format*

1. File Format Conversion

↓ *document in a standard format (plain text or html)*

2. English Tokenizer

↓ *annotated word tokens*

3. Wordlist Lookup + enhanced wordlists

↓ *annotated person first & last names, months, years, city names, etc.*

4. Sentence Splitter

↓ *annotation of sentences*

5. Hepple Part of Speech Tagger + lexicon

↓ *parts of speech of tokens*

6. Semantic Tagger + Named Entity Rules

↓ *annotated dates, person names, address, job titles, topics*

7. Intellectual Element Annotator + Intellectual Element Rules

↓ *intellectual elements of document*

8. SUPPLE Parser/Interpreter + Document Type Grammars Augmented with Semantics

↓ *document structure, semantics of document type*

9. Extract Metadata

↓ *document type, date, author, addressee, topic*

**Figure 4. The Process of Document Type Recognition and Metadata Extraction**

The first through the sixth steps are a previously implemented method for automatically annotating semantic categories in text such as person's names, job titles, dates, location names, postal addresses and organization names.[8] The input to the method is an e-record in a proprietary file format. The first step

[7] W. Underwood and S. Laib, Automatic Recognition of Documentary Forms, Technical Report ITTL/CSITD 08-02, Georgia Tech Research Institute, May 2008. http://perpos.gtri.gatech.edu.

[8] W. Underwood and S. Isbell, Semantic Annotation of Presidential E-Records, Technical Report ITTL/CSITD 08-01, Georgia Tech Research Institute, May 2008. http://perpos.gtri.gatech.edu.

converts that record to a plain text or html file format. The third step, Wordlist lookup, matches the terms (tokens) in the document against approximately 170,000 terms in 181 wordlists for such classes as person first names, surnames, city names, country names, months, and organizational nouns. If there is a match, the text is annotated with a tag for the name of that class. The sixth step, Semantic Tagger applies rules to the previously annotated text to produce additional annotations, for example, person's full names, locations made up of city and state or country names. Figure 5 shows a document whose paragraphs, dates, times, and person, location and organization names have been annotated by the first six steps of the method.

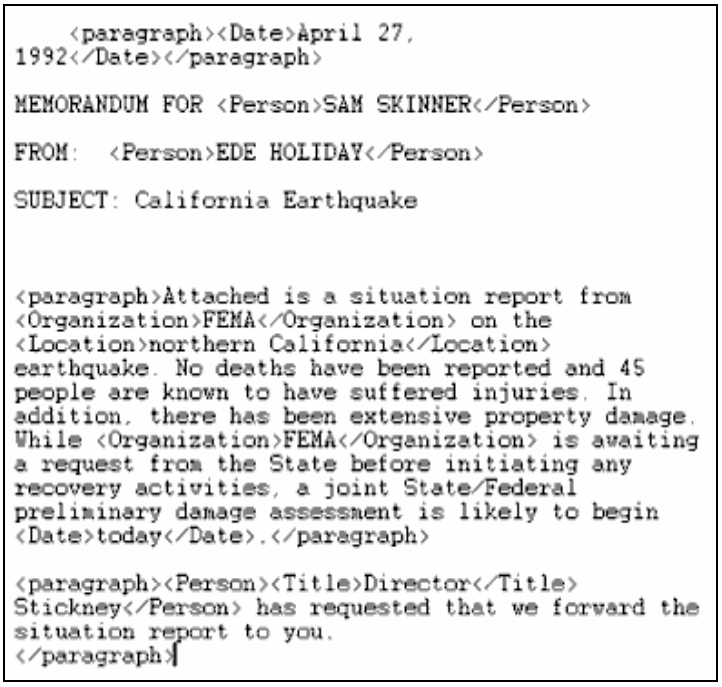


**Figure 5. Document with Annotated Paragraphs and Semantic Categories**

The seventh step, Intellectual Element Annotator, recognizes and annotates the intellectual elements occurring in a document. Currently, there are about 80 intellectual element rules. They apply to the annotated document and identify text strings such as FROM:, SUBJECT:, Attachment, or previously annotated semantic categories such as date, address and person's name as intellectual elements. Figure 6 shows the document in Figure 5 after the annotation of the intellectual elements. The names of the intellectual elements shown in Figure 6 are *chron*(*ological*)*date*, *for*, *person*, *from*, *subj*, *topic*, *para* and *attachment*.

The eighth step, SUPPLE Parser/Interpreter, recognizes the document type using a parse/interpreter with a context-free grammar that characterize the intellectual form of a document type. A *context-free grammar* is a 4-tuple <N, T, R, S> where N is a set of *non-terminal* symbols, T is a set of *terminal symbols*, R is a set of *rules* of the form A→*w* (A is a member of N and *w* is a string of symbols from N or T), and S is a member of N called the *initial symbol*. Linguists use context-free grammars to define the structure of sentences in a language and Computer Scientists use them to define programming languages.

```
<document>

      <chrondate>April  27, 1992</chrondate>


<for>MEMORANDUM  FOR</for>  <person>SAM  SKINNER</person>

<from>FROM:</from>      <person>EDE  HOLIDAY</person>

<subj>SUBJECT:</subj>   <topic>California  Earthquake</topic>


<para>Attached  is  a  situation  report  from  FEMA  on  the
northern  California  earthquake.  No  deaths  have  been
reported  and  45  people  are  known  to  have  suffered  injuries.
In  addition,  there  has  been  extensive  property  damage.
While  FEMA  is  awaiting  a  request  from  the  State  before
initiating  any  recovery  activities,  a  joint  State/Federal
preliminary  damage  assessment  is  likely  to  begin
today.</para>

<para>Director  Stickney  has  requested  that  we  forward  the
situation  report  to  you.</para>

<attachment>Attachments</attachment>
</document>
```

**Figure 6. Annotated Intellectual Elements**

Figure 7 shows the rules of a context-free grammar for the intellectual form of a memorandum. MEMO is the initial symbol of the grammar. The first rule defines a MEMO as consisting of a MEMOHEAD followed by a BODY. The BODY may be followed by OPTIONAL elements. A MEMOHEAD consists of an intellectual element *DATE* followed by an ADDRLINE followed by a SNDRLINE followed by a SUBJLINE. Optionally, there may be a THRULINE between the ADDRLINE and SUBJLINE. An ADDRLINE consists of an intellectual element *FOR* followed by ENTITIES. The SNDRLINE consist of an intellectual element *FROM* followed by ENTITIES. The SUBJLINE consists of an intellectual element *SUBJ* followed by a intellectual element *TOPIC*. ENTITIES consist of a sequence of one or more intellectual elements *PERSON*, *JOBTITLE*, or *PERSON JOBTITLE*. The BODY consists of a sequence of intellectual elements *PARA.* An OPTIONAL element consists of an intellectual element *ATTACHMENT* or a CCLIST or a BCCLIST, or combinations of these. A CCLIST consists of an intellectual element CC followed by ENTITIES. Similarly for a BCCLIST.

MEMO → MEMOHEAD BODY
MEMO → MEMOHEAD BODY OPTIONAL
MEMOHEAD → *CHRONDATE* ADDRLINE SNDRLINE SUBJLINE
MEMOHEAD → *CHRONDATE* ADDRLINE THRULINE SNDRLINE SUBJLINE
ADDRLINE → *FOR* ENTITIES
SNDRLINE → *FROM* ENTITIES
SUBJLINE → *SUBJ TOPIC*
THRULINE → *THRU* ENTITY
ENTITIES → ENTITIES ENTITY
ENTITIES → ENTITY
ENTITY → *PERSON JOBTITLE*
ENTITY → *PERSON*
ENTITY → *JOBTITLE*
BODY → PARAS
PARAS → *PARA* PARAS
PARAS → *PARA*
OPTIONAL → *ATTACHMENT*
OPTIONAL → *ATTACHMENT* CCLIST
OPTIONAL → *ATTACHMENT* BCCLIST
OPTIONAL → *ATTACHMENT* CCLIST BCCLIST
OPTIONAL → CCLIST
OPTIONAL → CCLIST BCCLIST
OPTIONAL → BCCLIST
CCLIST → *CC* ENTITIES
BCCLIST → *BCC* ENTITIES

**Figure 7. Grammar for the Intellectual Form of a Memorandum**

Figure 8 shows the grammar shown in Figure 7 augmented with semantic rules that create an interpretation of the meaning of the documentary form, that is, a representation of the name of document type, its date, author, addressee, and topic. The Intellectual Element Annotator assigns a value to each of the intellectual elements in the grammar. For example, for the document in Figure 6, the intellectual element *PERSON* after the intellectual element *MEMORANDUM FOR* will get the value 'SAM SKINNER'. In Figure 8, the two percent symbols (%%) indicate a comment. A grammar rule such as A $\rightarrow B_1, \ldots B_n$ is represented to the parser by a rule of the form rule(A $[B_1, \ldots B_n]$). The grammar rules are augmented with semantics by the notation included in parentheses after the symbols in the rules, e.g., rule(A( ) $[B_1( ), \ldots B_n( )]$). For instance, the rule

rule(entity(sform:F,sem E^[name,E, PERSON])
[person(s_form:F,sem PERSON)])

shown at the bottom of Figure 8 is used to recognize that a PERSON's name is an ENTITY. The value of the intellectual element *PERSON* is passed to the left-hand side of the rule, ENTITY, and a list [name, E, PERSON] is created whose semantic value is associated with ENTITY. When the rule ENTITIES → ENTITY is used to recognize an ENTITY as an ENTITIES, the semantic value of ENTITY is passed to ENTITIES. When the intellectual element *FOR* followed by ENTITIES is recognized, the semantic value of ENTITIES is passed to ADDRLINE where it is made the value of ADDRList. When *CHRONDATE*, ADDRLINE, SNDRLINE and SUBJLINE are recognized, the semantic value of each of these elements is passed to the variables DATE, ADDRList, SNDRList, and TOPIC and become the semantic values of MEMOHEAD. When MEMOHEAD and BODY are recognized, the semantic values of MEMOHEAD become the semantic values of MEMO.

```
%% MEMO-->MEMOHEAD BODY
rule(memo(s_form:F,sem:D^E2^E1^[[document,D],
        [document_form,D,'White House Memorandum'],[author,D,E2],
        SNDRList,[addressee,D,E1],ADDRList,[topic,D,TOPIC], [date,D,DATE]]),
    [memohead(s_form:F,sem:E1^E2^[DATE,ADDRList,SNDRList,TOPIC]),
     body(s_form:F)]).

%% MEMOHEAD-->CHRONDATE  ADDRLINE SNDRLINE  SUBJLINE
rule(memohead(s_form:F,sem:E1^E2^[DATE,ADDRList,SNDRList,TOPIC]),
 [chrondate(s_form:F,sem:DATE),
  addrline(s_form:F,sem:E1^ADDRList),
  sndrline(s_form:F,sem:E2^SNDRList),
  subjline(s_form:F,sem:TOPIC)]).

%% ADDRLINE-->FOR ENTITIES
rule(addrline(s_form:F,sem:ADDRList),
 [for(s_form:F),   entities(s_form:F,sem:ADDRList)]).

%% ENTITIES-->ENTITY
rule(entities(s_form:F,sem:E^SEM),
 [entity(s_form:F,sem:E^SEM)]).

%% ENTITY-->PERSON
rule(entity(s_form:F,sem:E^[name,E,PERSON]),
     [person(s_form:F,sem:PERSON)]).
```

**Figure 8. Part of the Grammar for the Intellectual Form of a Memorandum Augmented with Semantic Rules**

A parser with grammars for many document types is applied to a document whose intellectual elements have been identified. The parser produces a parse tree representing the documentary form of the document and a logical representation of the semantics of the document. Figure 9 shows the parse tree for the document shown in Figure 6.

The logical representation of the semantics of the sample memo is shown below:

qlf=[document(e1),
document_form(e1, 'White House Memorandum'),
author(e1, e2),
name(e2, 'EDE HOLIDAY')
addressee(e1, e3),
name(e3, 'SAM SKINNER')
topic(e1, 'California Earthquake'),
date(e1, 'April 27, 1992')]

It states that e1 is a document, the document form of e1 is memo, the author of e1 is e2, the name of e2 is 'EDE HOLIDAY', the addressee of e1 is e3, the name of e3 is 'SAM SKINNER', the topic of e1 is 'California Earthquake', and the date of e1 is 'April 27, 1992'.

```
(best_parse=(memo
  (head (chrondate (sem_cat "April 27, 1992"))
        (addrline (for (sem_cat "MEMORANDUM FOR"))
              (entities (entity (person (sem_cat "SAM SKINNER")))))
        (sndrline (from (sem_cat "FROM:"))
              (entities (entity (person (sem_cat "EDE HOLIDAY")))))
        (subjline (subj (sem_cat "SUBJECT:"))
              (topic (sem_cat "California Earthquake"))))
  (body (paras (para
        (sem_cat "Attached is a situation report from FEMA on the
        northern California earthquake. No deaths have been
        reported and 45 people are known to have suffered injuries.
        In addition, there has been extensive property damage.
        While FEMA is awaiting a request from the State before
        initiating any recovery activities, a joint State/Federal
        preliminary damage assessment is likely to begin today."))
        paras (para
        (sem_cat "Director Stickney has requested that we forward
        the situation report to you."))))))
  (optional (attachment (sem_cat "Attachments")))))
```

**Figure 9. Parse Tree for the Sample Memorandum**

In the ninth step, Extract metadata, the document metadata is extracted from this representation. The metadata for this document can be used for creating item titles or item descriptions such as the following:

> A memorandum dated April 27, 1992 from Ede Holiday to Sam Skinner regarding California Earthquake.

The metadata can also be used to provide access points for document search and retrieval.

**Summary of Results and Current Research**

The results of this research are that: (1) the intellectual elements of documentary forms can be defined in terms of keywords and semantic categories in a document, (2) documentary forms (record or document types) can be defined using context-free grammars, and (3) grammars for documentary forms can be used with a parser/interpreter for context-free grammars to automatically recognize the documentary form of textual records while simultaneously identifying document metadata including date, author, addressee, and topic.

Context-free grammars have been constructed for twenty-two of the documentary forms that occur in Presidential e-records. Rules were constructed for recognizing the intellectual elements occurring in nine of these documentary forms—Memoranda, White House Letters, Formal Letters, White House Press Releases, Recommended Telephone Calls, Executive orders, National Security Directives, National Security Reviews, and Minutes of Cabinet Meetings. Nine grammars for the forms were translated into context-free attribute grammars that were used with a parser to parse and interpret the intellectual elements of Presidential e-records. The resulting semantic representation can be used to extract metadata needed for archival description and for record search and retrieval.

Parsing of intellectual elements using manually constructed grammars for documentary forms will not recognize all examples of these document types because people do not always follow exactly the documentary form prescribed in style manuals. In prior research, the automatic induction of grammars for characterizing documentary forms of e-records was investigated.[9] It is planned to collect samples of e-records of the document types considered in this paper and additional document types, and to automatically induce grammars from these samples.

The research described in this paper addressed only the intellectual form of documents. In further research, rules will be formulated for recognizing the physical elements of the physical form of a document. These are elements such as the fonts, font sizes, underlining, horizontal bars, bold and italics. These features are important for recognizing the layout and appearance of a document and for defining additional intellectual elements such as headings.

**References**


Bush Presidential Library, Bush Presidential Records, WHORM Subject File, Disasters-Natural, ID#324869.

International Council on Archives (ICA). ISAD(G): General International Standard Archival Description, Second Edition, 1999.

International Standards Organization. Standard Generalized Markup Language – ISO 8879.

National Archives and Records Administration (NARA). Life Cycle Data Requirements Guide (LCDRG), March 3, 2008.

World Wide Web Consortium. Extensible Markup Language XML 1.0 (Fourth Edition), 16 Aug. 2006.

Underwood, W. and S. Laib, Automatic Recognition of Documentary Forms, Technical Report ITTL/CSITD 08-02, Georgia Tech Research Institute, May 2008. http://perpos.gtri.gatech.edu.

Underwood, W. and S. Isbell, Semantic Annotation of Presidential E-Records, Technical Report ITTL/CSITD 08-01, Georgia Tech Research Institute, May 2008. http://perpos.gtri.gatech.edu.


[9] W. E. Underwood and B. Harris. Inferring and Recognizing the Documentary Form of Record Types. PERPOS TR ITTL/CSITD 05-8, Georgia Tech Research Institute, August 2006. http://perpos.gtri.gatech.edu.