Tools for File Type and Record Type Identification

William Underwood

Georgia Tech Research Institute

Atlanta, Georgia, USA

75th Annual Meeting of the

Society of American Archivists

Chicago, Illinois

August 25, 2011

Research Motivation

- Archivists need the capability to identify file formats for
  - Insuring compliance with Record Transmittal Agreement
  - Viewing/playing files
  - Conversion to current or standard file formats
  - Archive extraction
  - Password recovery and decryption
  - Repair of damaged files

Definitions

- A file format is a set of rules for encoding and decoding data or computer instructions in a file.
- A *file type* is a class of files with the same file format.
- A *file format signature* is invariant data in a file format that can be used to identify the file type (or format) of a file

External File Format Identifiers

- File Name Extensions
- Metadata stored in the operating system
  - MacOS HFS Creator Code & File Type Code
  - MacOS X Uniform Type Identifier (UTI)
- Multipurpose Internet Mail Extensions (MIME) media types
- PRONOM Persistent Universal Identifier (PUID)

Linux File Command and Magic File

- Unix (Linux) File Command and Magic File are probably the most widely used tool for file format identification.
- Magic number is the term used for the concept of an internal file format signature.
- The file command applies tests for magic numbers contained in the Magic file to files to determine their file type and relevant metadata.

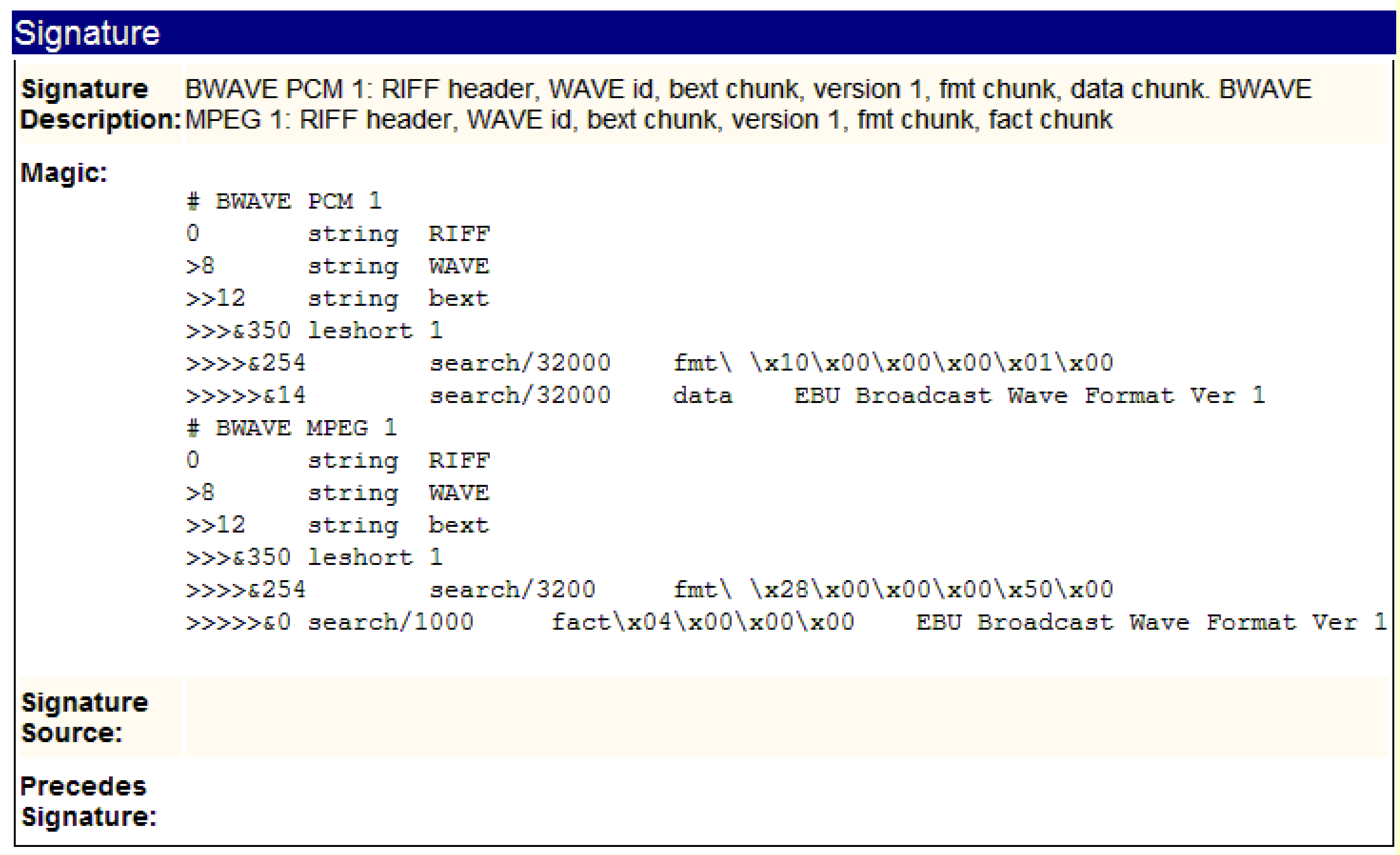
Some Limitations of the file Command and Magic File

- Difficult to update the tests for magic numbers.
- Tests that may give conflicting results must be properly sequenced.
- Tests for magic numbers are not one-to-one with file types.
- Tests output metadata as well as file type.
- Tests for character set and language of text files needs refinement.
- Only a few tests for MS Windows file types.
- Tests for Magic numbers have not been rigorously tested

Extensions of File Command and Magic File to overcome Limitations

- File Format Library
- Magic for individual file formats
- Output of file command/magic file is File Format ID
- Rewriting file command code for identifying Characteristics of Text files and Document Types
- Defined approx. 850 file format signatures
- Collected examples of approx. 700 of the file format types
- Created File Signature Database
- Verified that File Format Identifier with magic file correctly identifies approx. 700 File Types

Magic Test
for Broadcast Wave Format Ver 1



Signature

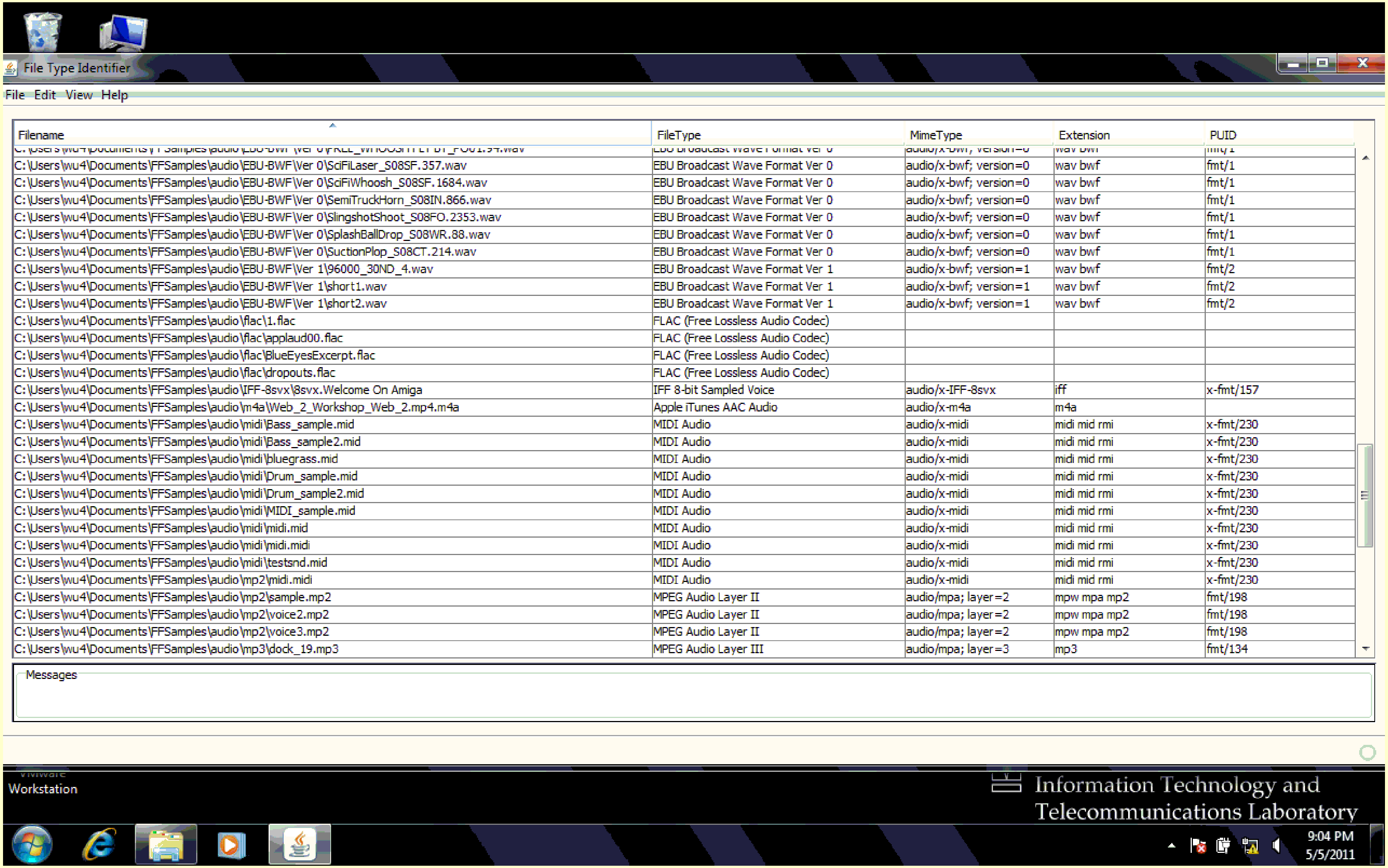
Signature Description: BWAVE PCM 1: RIFF header, WAVE id, bext chunk, version 1, fmt chunk, data chunk. BWAVE MPEG 1: RIFF header, WAVE id, bext chunk, version 1, fmt chunk, fact chunk

Magic:

```
# BWAVE PCM 1
0       string  RIFF
>8      string  WAVE
>>12    string  bext
>>>&350 leshort 1
>>>>&254        search/32000    fmt\ \x10\x00\x00\x00\x01\x00
>>>>>&14        search/32000    data    EBU Broadcast Wave Format Ver 1
# BWAVE MPEG 1
0       string  RIFF
>8      string  WAVE
>>12    string  bext
>>>&350 leshort 1
>>>>&254        search/3200     fmt\ \x28\x00\x00\x00\x50\x00
>>>>>&0 search/1000     fact\x04\x00\x00\x00    EBU Broadcast Wave Format Ver 1
```

Signature Source:

Precedes Signature:



Filename	FileType	MimeType	Extension	PUID
C:\Users\wu4\Documents\FFSamples\audio\EBU-BWF\Ver 0\SciFiLaser_S08SF.357.wav	EBU Broadcast Wave Format Ver 0	audio/x-bwf; version=0	wav bwf	fmt/1
C:\Users\wu4\Documents\FFSamples\audio\EBU-BWF\Ver 0\SciFiWhoosh_S08SF.1684.wav	EBU Broadcast Wave Format Ver 0	audio/x-bwf; version=0	wav bwf	fmt/1
C:\Users\wu4\Documents\FFSamples\audio\EBU-BWF\Ver 0\SemiTruckHorn_S08IN.866.wav	EBU Broadcast Wave Format Ver 0	audio/x-bwf; version=0	wav bwf	fmt/1
C:\Users\wu4\Documents\FFSamples\audio\EBU-BWF\Ver 0\SlingshotShoot_S08FO.2353.wav	EBU Broadcast Wave Format Ver 0	audio/x-bwf; version=0	wav bwf	fmt/1
C:\Users\wu4\Documents\FFSamples\audio\EBU-BWF\Ver 0\SplashBallDrop_S08WR.88.wav	EBU Broadcast Wave Format Ver 0	audio/x-bwf; version=0	wav bwf	fmt/1
C:\Users\wu4\Documents\FFSamples\audio\EBU-BWF\Ver 0\SuctionPlop_S08CT.214.wav	EBU Broadcast Wave Format Ver 0	audio/x-bwf; version=0	wav bwf	fmt/1
C:\Users\wu4\Documents\FFSamples\audio\EBU-BWF\Ver 1\96000_30ND_4.wav	EBU Broadcast Wave Format Ver 1	audio/x-bwf; version=1	wav bwf	fmt/2
C:\Users\wu4\Documents\FFSamples\audio\EBU-BWF\Ver 1\short1.wav	EBU Broadcast Wave Format Ver 1	audio/x-bwf; version=1	wav bwf	fmt/2
C:\Users\wu4\Documents\FFSamples\audio\EBU-BWF\Ver 1\short2.wav	EBU Broadcast Wave Format Ver 1	audio/x-bwf; version=1	wav bwf	fmt/2
C:\Users\wu4\Documents\FFSamples\audio\flac\1.flac	FLAC (Free Lossless Audio Codec)			
C:\Users\wu4\Documents\FFSamples\audio\flac\applaud00.flac	FLAC (Free Lossless Audio Codec)			
C:\Users\wu4\Documents\FFSamples\audio\flac\BlueEyesExcerpt.flac	FLAC (Free Lossless Audio Codec)			
C:\Users\wu4\Documents\FFSamples\audio\flac\dropouts.flac	FLAC (Free Lossless Audio Codec)			
C:\Users\wu4\Documents\FFSamples\audio\IFF-8svx\8svx.Welcome On Amiga	IFF 8-bit Sampled Voice	audio/x-IFF-8svx	iff	x-fmt/157
C:\Users\wu4\Documents\FFSamples\audio\m4a\Web_2_Workshop_Web_2.mp4.m4a	Apple iTunes AAC Audio	audio/x-m4a	m4a	
C:\Users\wu4\Documents\FFSamples\audio\midi\Bass_sample.mid	MIDI Audio	audio/x-midi	midi mid rmi	x-fmt/230
C:\Users\wu4\Documents\FFSamples\audio\midi\Bass_sample2.mid	MIDI Audio	audio/x-midi	midi mid rmi	x-fmt/230
C:\Users\wu4\Documents\FFSamples\audio\midi\bluegrass.mid	MIDI Audio	audio/x-midi	midi mid rmi	x-fmt/230
C:\Users\wu4\Documents\FFSamples\audio\midi\Drum_sample.mid	MIDI Audio	audio/x-midi	midi mid rmi	x-fmt/230
C:\Users\wu4\Documents\FFSamples\audio\midi\Drum_sample2.mid	MIDI Audio	audio/x-midi	midi mid rmi	x-fmt/230
C:\Users\wu4\Documents\FFSamples\audio\midi\MIDI_sample.mid	MIDI Audio	audio/x-midi	midi mid rmi	x-fmt/230
C:\Users\wu4\Documents\FFSamples\audio\midi\midi.mid	MIDI Audio	audio/x-midi	midi mid rmi	x-fmt/230
C:\Users\wu4\Documents\FFSamples\audio\midi\midi.midi	MIDI Audio	audio/x-midi	midi mid rmi	x-fmt/230
C:\Users\wu4\Documents\FFSamples\audio\midi\testsnd.mid	MIDI Audio	audio/x-midi	midi mid rmi	x-fmt/230
C:\Users\wu4\Documents\FFSamples\audio\mp2\midi.midi	MIDI Audio	audio/x-midi	midi mid rmi	x-fmt/230
C:\Users\wu4\Documents\FFSamples\audio\mp2\sample.mp2	MPEG Audio Layer II	audio/mpa; layer=2	mpw mpa mp2	fmt/198
C:\Users\wu4\Documents\FFSamples\audio\mp2\voice2.mp2	MPEG Audio Layer II	audio/mpa; layer=2	mpw mpa mp2	fmt/198
C:\Users\wu4\Documents\FFSamples\audio\mp2\voice3.mp2	MPEG Audio Layer II	audio/mpa; layer=2	mpw mpa mp2	fmt/198
C:\Users\wu4\Documents\FFSamples\audio\mp3\dock_19.mp3	MPEG Audio Layer III	audio/mpa; layer=3	mp3	fmt/134

Research Motivation

- Metadata extraction is a critical aspect of the ingestion of textual e-records into digital archives and libraries.
- Metadata is needed to support description of individual e-records and aggregations of these records and to support search and retrieval of records.

Document Types: Examples in Presidential E-Records

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

Documentary Form

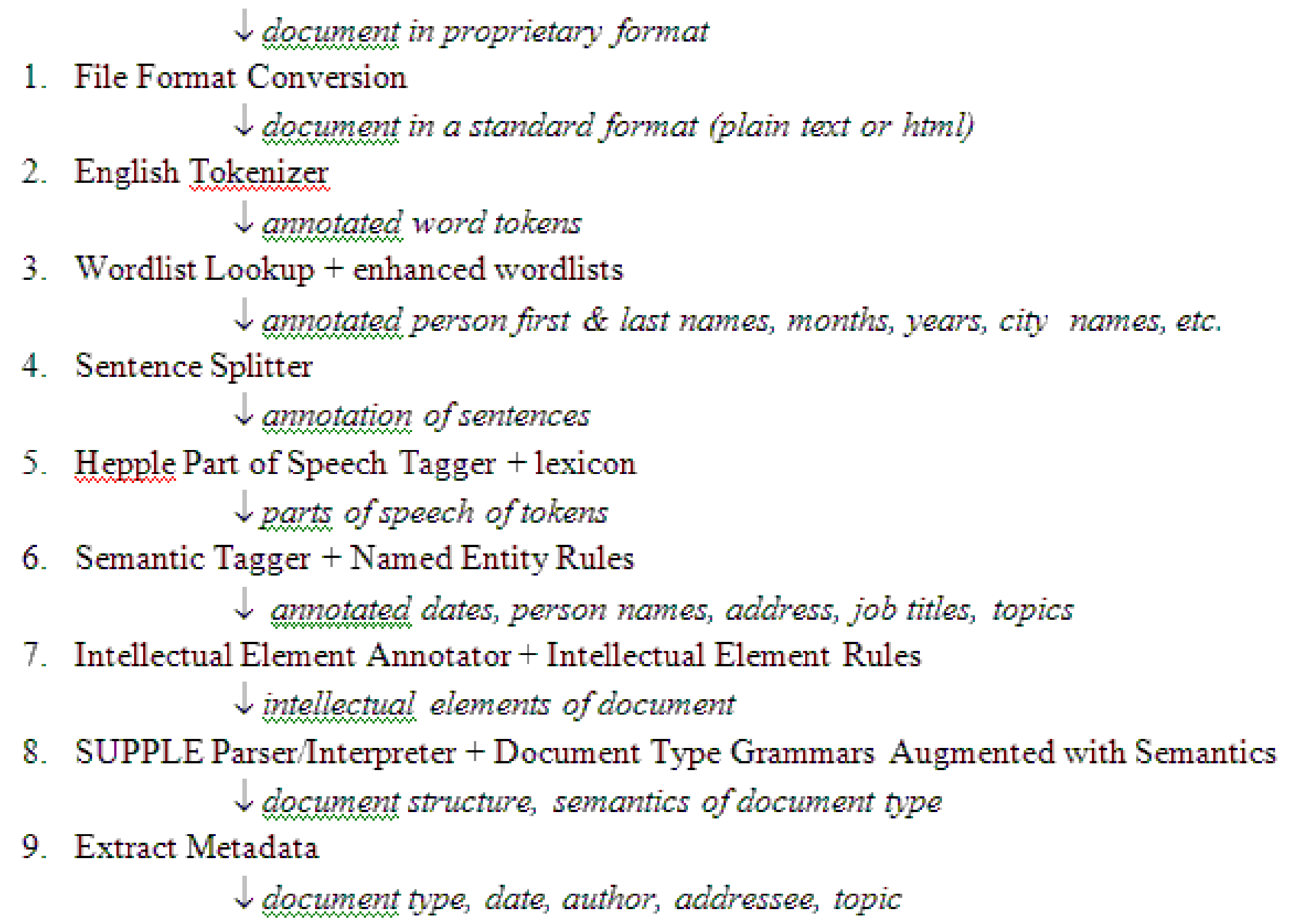
- Documentary form consists of intellectual form and physical form.
- Intellectual elements are those terms or semantic categories that are common to a document type.
- Intellectual form is the rules that characterize the possible combinations of intellectual elements
- Physical elements are the physical attributes of the intellectual elements
- Physical form is the rules that characterize the physical layout of the physical elements.

L. Duranti, Diplomatics: New Uses for an old Science

Documentary Form, Record Types, and Document Type Definitions

- A (documentary) *form* is “a class of documents distinguished on the basis of common physical and/or intellectual characteristics of a document.” [ICA, ISAD(G)]
- A *specific records type* is “the intellectual format of the archival materials.” [NARA, LCDRG]
- An XML *document type definition* (DTD) specifies the intellectual form of a document in terms of its elements, an extended context-free grammar and a tag set used to mark-up a document. [W3C, XML]
- An XSL style sheet specifies the physical elements and physical layout of the elements of a DTD.

Method for Recognizing Document Forms and Extracting Metadata



↓ *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*

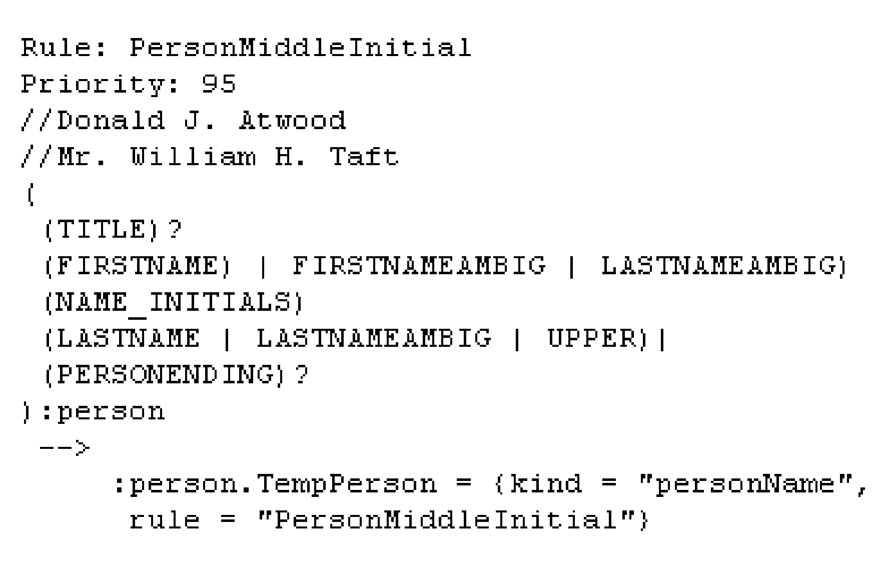
9. Extract Metadata

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

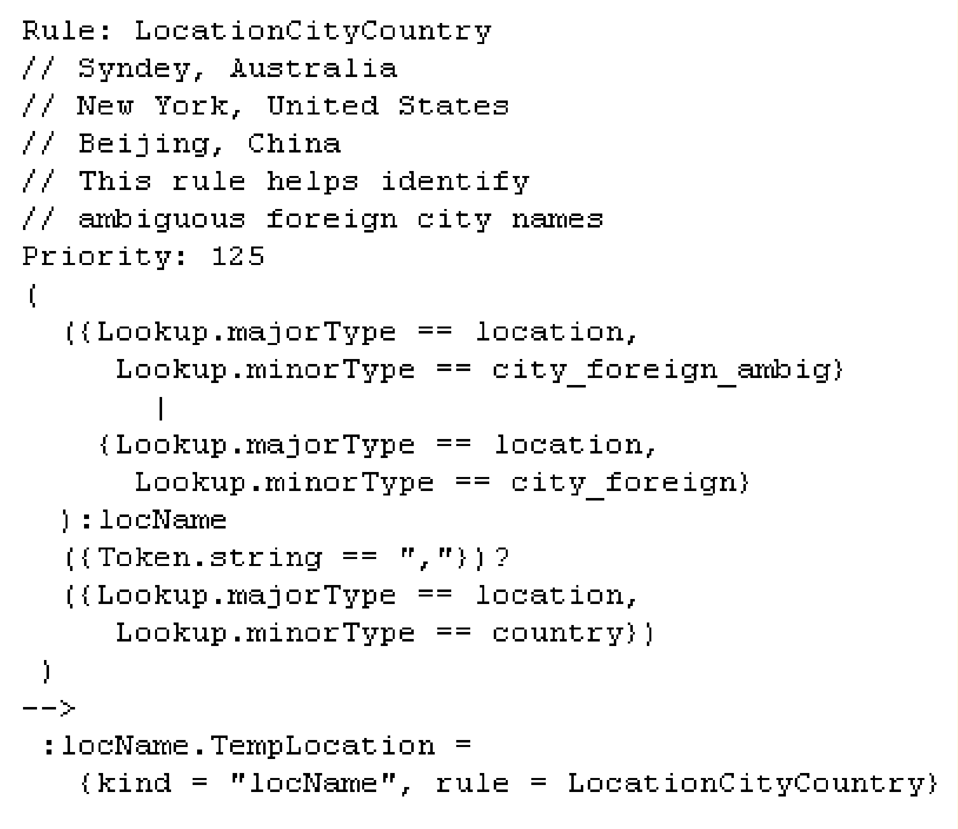
Information Extraction: Wordlists

- Person_female_first.lst (8263)
- Person_male_first.lst (3704)
- Person_surname.lst (83,805)
- Location_city_us.lst (33,017)
- Location_us_county.lst (1,938)
- Location_us_state.lst (50)
- Location_foreign_city.lst (3802)
- Location_country.lst (458)
- Org_noun.lst (915)
- Org_ending.lst (238)
- Org_us_govt_dept_agency.lst (519)

Java Annotation Pattern Engine (JAPE) Rules

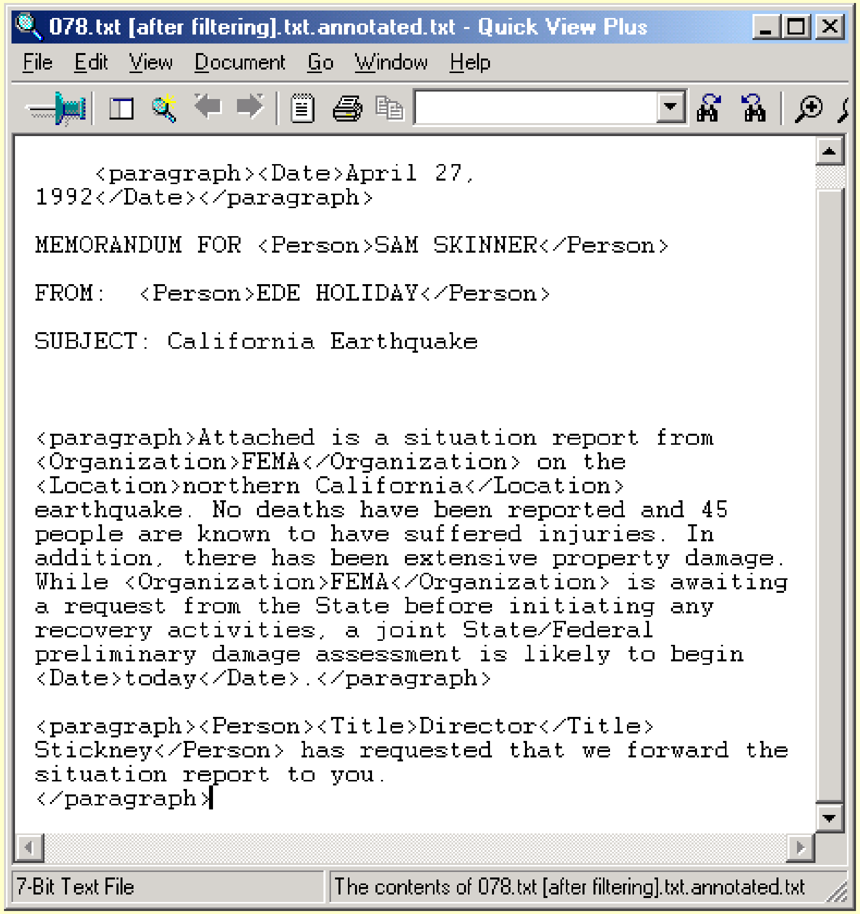


```
Rule: PersonMiddleInitial
Priority: 95
//Donald J. Atwood
//Mr. William H. Taft
(
 (TITLE)?
 (FIRSTNAME) | FIRSTNAMEAMBIG | LASTNAMEAMBIG)
 (NAME_INITIALS)
 (LASTNAME | LASTNAMEAMBIG | UPPER)|
 (PERSONENDING)?
):person
 -->
     :person.TempPerson = {kind = "personName",
      rule = "PersonMiddleInitial"}
```



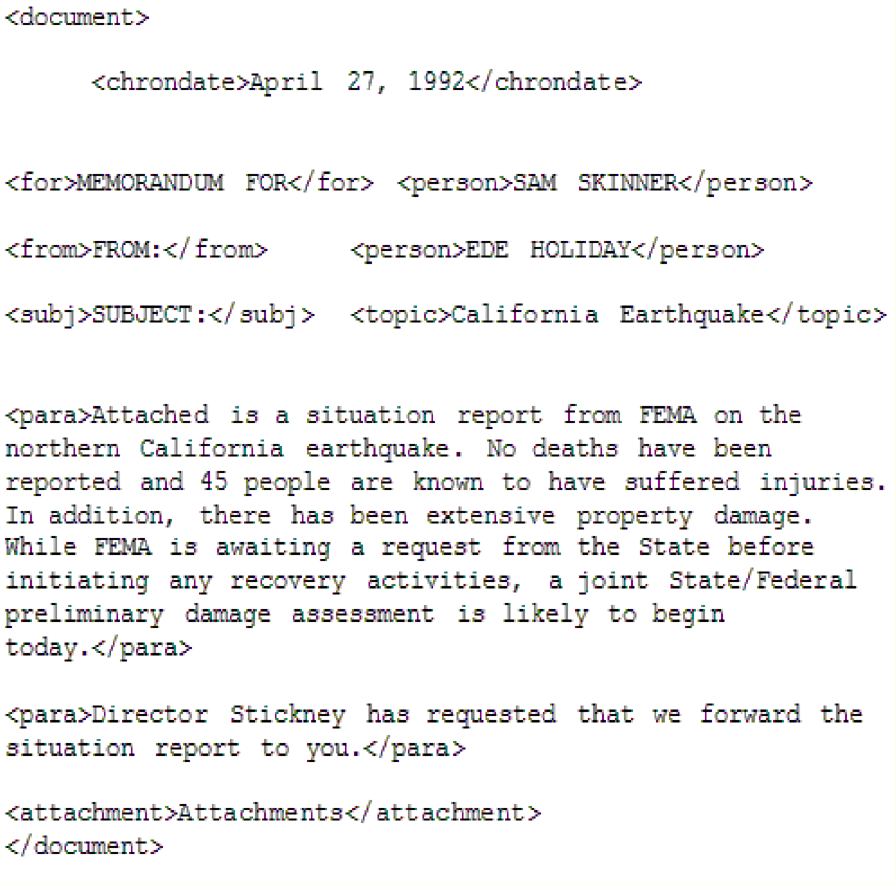
```
Rule: LocationCityCountry
// Syndey, Australia
// New York, United States
// Beijing, China
// This rule helps identify
// ambiguous foreign city names
Priority: 125
(
  ({Lookup.majorType == location,
     Lookup.minorType == city_foreign_ambig}
        |
    {Lookup.majorType == location,
      Lookup.minorType == city_foreign}
  ):locName
  ({Token.string == ","})?
  ({Lookup.majorType == location,
     Lookup.minorType == country})
 )
-->
 :locName.TempLocation =
   {kind = "locName", rule = LocationCityCountry}
```

Documentary Form: Intellectual Element Recognition


078.txt [after filtering].txt.annotated.txt - Quick View Plus
File Edit View Document Go Window Help

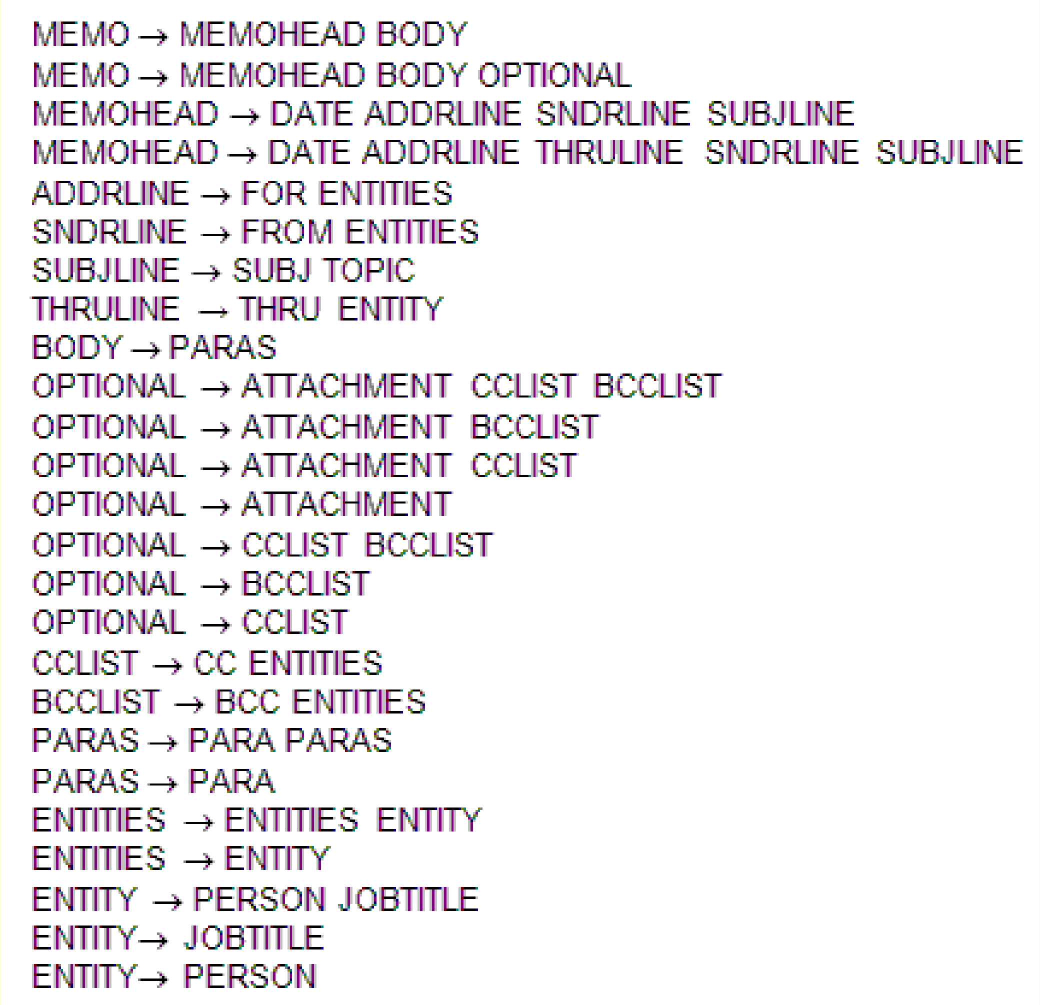
```
    <paragraph><Date>April 27,
1992</Date></paragraph>

MEMORANDUM FOR <Person>SAM SKINNER</Person>

FROM:  <Person>EDE HOLIDAY</Person>

SUBJECT: California Earthquake



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

<paragraph><Person><Title>Director</Title>
Stickney</Person> has requested that we forward the
situation report to you.
</paragraph>
```

7-Bit Text File
The contents of 078.txt [after filtering].txt.annotated.txt



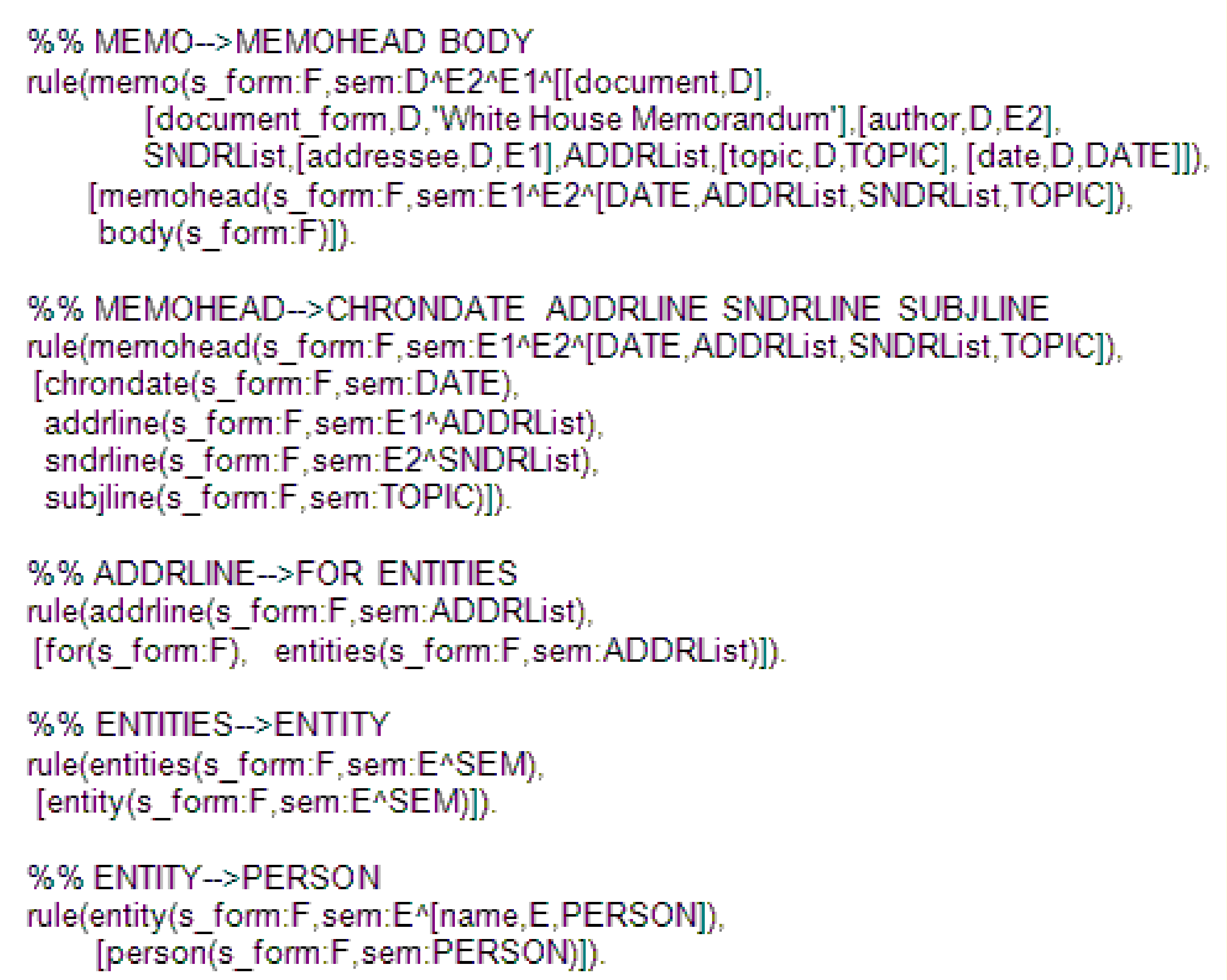
```
<document>

      <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>
```

Document Types: Grammar for a Memorandum



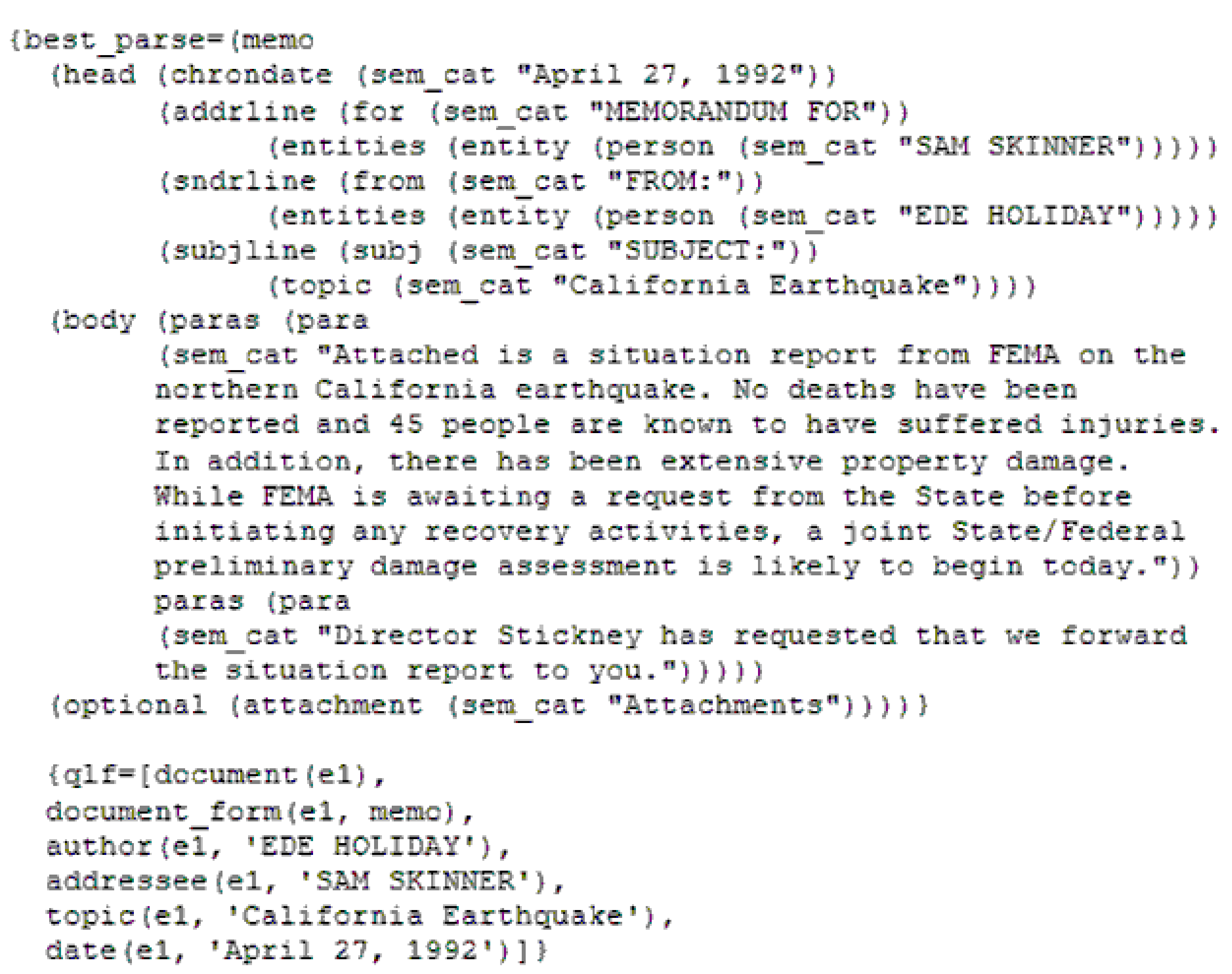
```
MEMO → MEMOHEAD BODY
MEMO → MEMOHEAD BODY OPTIONAL
MEMOHEAD → DATE ADDRLINE SNDRLINE SUBJLINE
MEMOHEAD → DATE ADDRLINE THRULINE SNDRLINE SUBJLINE
ADDRLINE → FOR ENTITIES
SNDRLINE → FROM ENTITIES
SUBJLINE → SUBJ TOPIC
THRULINE → THRU ENTITY
BODY → PARAS
OPTIONAL → ATTACHMENT CCLIST BCCLIST
OPTIONAL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

Grammar for Memorandum Augmented with Semantic Rules



```
%% 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)]).
```

Parse Tree and Semantics of the Document



```
{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")))))}

  {qlf=[document(e1),
  document_form(e1, memo),
  author(e1, 'EDE HOLIDAY'),
  addressee(e1, 'SAM SKINNER'),
  topic(e1, 'California Earthquake'),
  date(e1, 'April 27, 1992')]}
```

Metadata Extracted
for Item Description and Indexing

DocumentType = memo

Date = April 27, 1992

Author = SAM SKINNER

Addressee = EDE HOLIDAY

Topic = California Earthquake

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

Grammars and Semantics for Documentary Forms

Formal Letter

White House Informal Letter

White House Memorandum

Action-Decision Memorandum

White House Referral

Recommended Telephone Call

White House Press Release

Presidential Determination

Executive Order

Presidential Proclamation

National Security Directive

National Security Review

Memorandum of Conversation

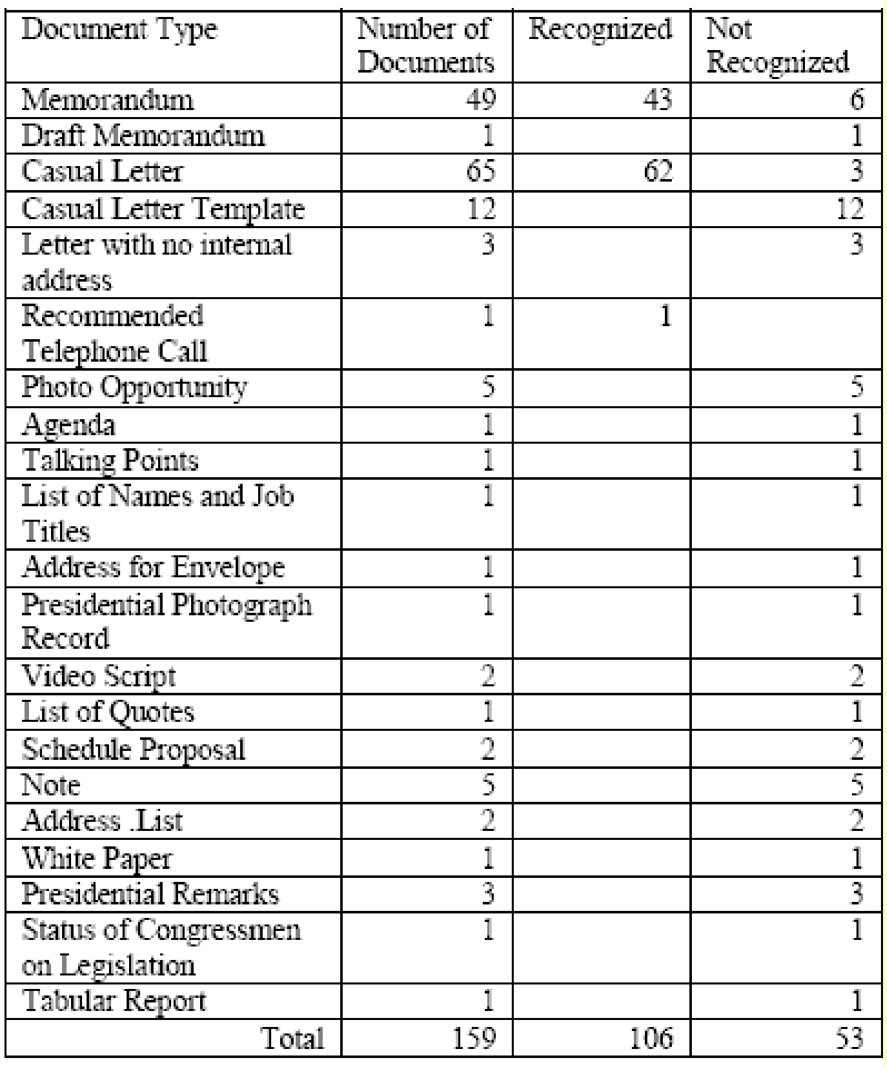
Memorandum of Telephone Conversation

Implementation and Test

Rules were constructed for recognizing the intellectual elements of these 14 documentary forms

- Grammars merged and converted to SUPPLE Parser Notation
- Semantics were added to grammar rules
- Option added to PERPOS for automatically describing contents of containers
- Implemented method interfaced to PERPOS
- Method was successfully tested on 112 documents of the 14 document types in various textual file formats.

Experimental Evaluation



Document Type	Number of Documents	Recognized	Not Recognized
Memorandum	49	43	6
Draft Memorandum	1		1
Casual Letter	65	62	3
Casual Letter Template	12		12
Letter with no internal address	3		3
Recommended Telephone Call	1	1	
Photo Opportunity	5		5
Agenda	1		1
Talking Points	1		1
List of Names and Job Titles	1		1
Address for Envelope	1		1
Presidential Photograph Record	1		1
Video Script	2		2
List of Quotes	1		1
Schedule Proposal	2		2
Note	5		5
Address .List	2		2
White Paper	1		1
Presidential Remarks	3		3
Status of Congressmen on Legislation	1		1
Tabular Report	1		1
Total	159	106	53

Summary of Research Results

- The intellectual elements of documentary forms can be defined in terms of the keywords and semantic categories in a document.
- Documentary forms can be defined using context-free grammars.
- Grammars for documentary forms can be used with a parser/interpreter to automatically recognize the documentary form of textual records and extract metadata.
- Method has been tested using grammars for 14 Document Types
- Method is being experimentally evaluated.

Research Issues

- Can the intellectual elements of documentary forms be learned without a teacher?
- Can grammatical induction be used with examples of a particular document type to induce a grammar automatically?
- Can the recognition method be extended to include physical elements of documentary form and grammatical definition of physical layout?

Additional Information

GTRI url: http://perpos.gtri.gatech.edu

PRONOM url: www.nationalarchives.gov.uk/PRONOM/Default.aspx

DROID url: http://source-forge.net/projects/droid

W. Underwood. Grammar-based Recognition of Documentary Forms and Extraction of Metadata. *The International Journal of Digital Curation*, Vol 5, Issue 1, 2010. www.ijdc.net/index.php/ijdc/article/view/152