

Automatic Metadata Extraction for Archival Description and Access

William Underwood

Georgia Tech Research Institute

Society of American Archivists

Research Forum

San Francisco

August 26, 2008

Research Motivation

- **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.**

Research Objectives

- **Techniques for automatically recognizing document types and extracting metadata from electronic records that 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.**

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

Document Types: Definitions

Documentary form is “the rules of representation used to convey a message, that is, the characteristics of a document which can be separated from the determination of the particular subjects, or places it concerns. Documentary form is both physical and intellectual.”

The ***intellectual form*** of a document is "the sum of a record's formal attributes that represent and communicate the elements of the action in which the record is involved and of its immediate context, both documentary and administrative."

The ***physical form*** of a document is “the overall appearance, configuration, or shape, derived from its material characteristics and independent of its intellectual content.”

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

Method for Recognizing Document Types and Extracting Date, Author, Addressee and Topic

1. Document Reader
2. English Tokenizer
3. Wordlist Lookup + *enhanced wordlists*
4. Sentence Splitter
5. Hepple POS Tagger + lexicon
6. Semantic Tagger + *Named Entity Rules*
7. *Intellectual Element Annotator + Intellectual Element Rules (DER)*
8. SUPPLE Parser/Interpreter + *Document Type Grammars augmented with Semantics*
9. *Extract Metadata*

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
// Sydney, 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}
```


Examples of Intellectual Element Rules

for → MEMORANDUM FOR

from → FROM:

subj → SUBJECT:

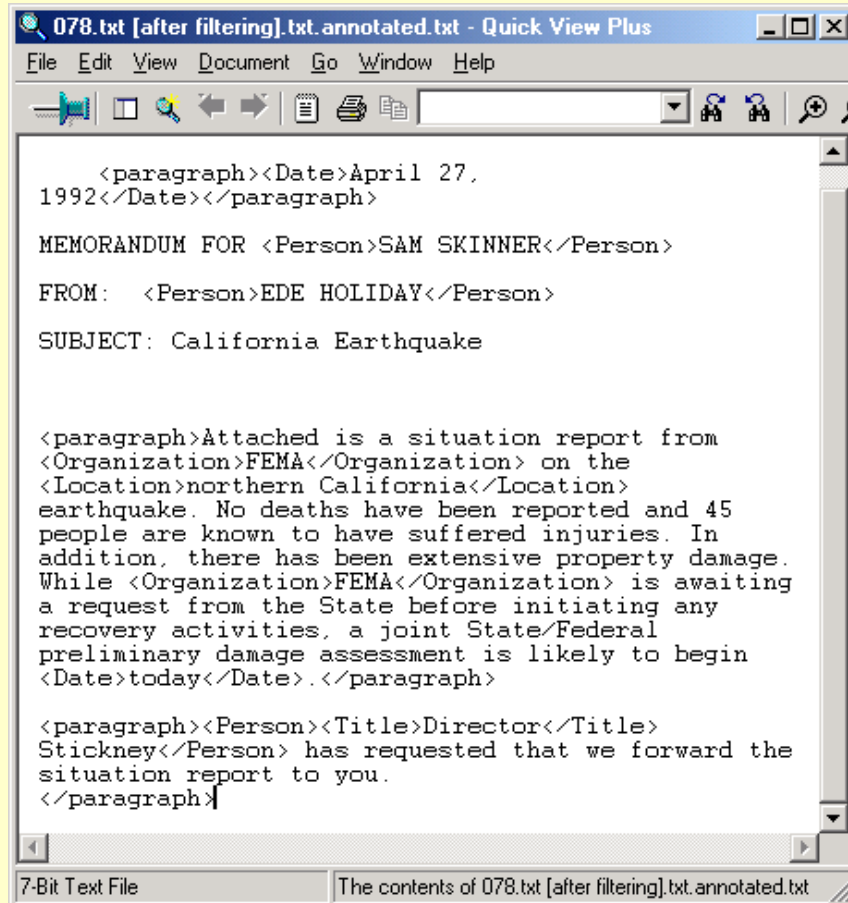
para → <paragraph> </paragraph>

date → <Date> </Date> *and not in a paragraph*

attachment → Attachments

attachment → Attachment

Documentary Form: Intellectual Element Recognition



The screenshot shows a text editor window titled "078.txt [after filtering].txt.annotated.txt - Quick View Plus". The window contains the following text with XML annotations:

```
<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>
```

```
<document>

    <chrontdate>April 27, 1992</chrontdate>

<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 (chrodate (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.

Summary of Research Results & Current Research

- **Results**
 - A method for automatic document type recognition and metadata extraction.
 - Demonstrated it on 10 document types
- **Current Research**
 - Inducing grammars for documentary form from samples
 - Automatic Description of items, file units and record series
 - Automatic Recognition of the topics of records

Additional Information

Website: perpos.gtri.gatech.edu

W. Underwood and S. Isbell, Semantic Annotation of Presidential E-Records, Technical Report ITTL/CSITD 08-01, May 2008

W. Underwood and S. Laib. Automatic Recognition of Documentary Forms, Technical Report ITTL/CSITD 08-02, May 2008

Backup

...

ARC - Archival Research Catalog

Search ARC for [Archival Descriptions](#) [Digital Copies](#) [People](#) [Organizations](#) [View My List](#) [Glossary](#) [Help](#)

Archival Descriptions Advanced Search

Find results: with **all** of the words [Search Tips](#)

with the **exact phrase**

with **at least one** of the words

without the words

Search **title only**

by **description identifier**
e.g 28803 OR "111-M-1342"

Limit results to [Basic Search](#)

Highlight Search Terms *(on the results page)*

▶ **Type of Archival Materials**

▶ **Location of Archival Materials**

▶ **Level of Description**

▼ **Date Options**

Date range:

Between and
YYYY YYYY

Descriptions Created or Updated Since
MM DD YYYY

Example of an Item Title and Description

Archival Description

56 results retrieved for **President bush** with filters applied

 [Refine Search](#)  [Highlight Search Terms](#)  [Return to Results](#)

Description 50 of 56 Descriptions page: [46](#) [47](#) [48](#) [49](#) [50](#) [51](#) [52](#) [53](#) [54](#) [55](#)

Letter from George H. W. Bush to His Children on New Year's Eve 1990, 12/31/1990  [Email](#)  [Print](#)

ARC Identifier 595134  [Add to My List](#)

Item from Collection GB-GBPP: George H. W. Bush Papers, ca. 1942 - ca. 2004

[Details](#) [Scope & Content](#) [Archived Copies](#) [Digital Copies](#) [Hierarchy](#)

This letter was typewritten by President George H. W. Bush and addressed to his children: George, Jeb, Neil, Marvin, and Doro. He expresses his happiness at their Christmas celebration held at Camp David, then writes concerning his conflicted feelings as he prepares for the possibility of war with Iraq.

Archival Description Hierarchy: Series, File Unit, Item

The screenshot displays a web-based archival description hierarchy. At the top, there are tabs for 'Details', 'Scope & Content', 'Archived Copies', 'Digital Copies', and 'Hierarchy'. Below the tabs are buttons for 'Open all', 'Close all', and 'Print Hierarchy'. The hierarchy is as follows:

- Collection:** GB-GBPP: George H. W. Bush Papers, ca. 1942 - ca. 2004
ARC ID: 595138
Creator: Bush, George, 1924-
- Series:** Presidential Daily Files, compiled 01/20/1989 - 12/31/1992
ARC ID: 595141
Contact(s): George Bush Library (NLGB), 1000 George Bush Drive West, College Station, TX, 77845. PHONE: 979-691-4000; FAX: 979-691-4050; EMAIL: bush.library@nara.gov.
- File Unit:**

File Unit	Container Count
Monday, December 31, 1990, 12/31/1990 - 12/31/1990 ARC ID: 595156	
- Item:**

Item	Container Count
Letter from George H. W. Bush to His Children on New Year's Eve 1990, 12/31/1990 (Textual Records) ARC ID: 595134	2 page(s)

Method for Extracting Person, Organization and Location Name Metadata

- Document Reader
- English Tokenizer
- Wordlist Lookup + Wordlists
- Sentence Splitter
- Hepple POS Tagger + Lexicon
- Named Entity Transducer + rules
- Extract Metadata

Annotation and Metadata Extraction: Performance

Annotation Type	Correct	Partially Correct	Missing	Spurious	Precision	Recall	F-Measure
Person	515	11	42	57	0.8928	0.9164	0.9044
Location	270	15	54	24	0.8981	0.8186	0.8565
Organization	509	31	31	50	0.889	0.9186	0.9035
Date	456	1	1	1	0.9967	0.9967	0.9967
Money	28	1	0	8	0.7703	0.9828	0.8636
Percent	6	0	0	0	1.0	1.0	1.0

Overall average precision: 0.9178 Overall average recall: 0.9282 Overall average F-measure: 0.9108

Annotated Document and Metadata Extracted For Indexing

<Date>April 27, 1992</Date>
<Time>7:00 a.m. EDT</Time>

SITUATION REPORT #3
PETROLIA EARTHQUAKE

DATE AND TIME

OF OCCURRENCE: <Date>April 25, 1992</Date>, <Time>11:06 a.m.
PDT</Time>

LOCATION: <Location>Northern California</Location>, 30 miles
southeast of <Location>Eureka</Location>

1. SITUATION:

<paragraph>At <Time>2:06 p.m. EDT</Time> on <Date>April
25</Date>, a 6.9 Richter magnitude earthquake occurred in
<Location>Northern California</Location>, 30 miles
southeast of <Location>Eureka</Location> near the town of
<Location>Petrolia</Location>.</paragraph>

<paragraph>At <Time>3:42 a.m. EDT</Time> on <Date>April
26</Date> a strong aftershock (6.2 magnitude) occurred in
the vicinity of <Location>Ferndale</Location> and
<Location>Petrolia</Location> causing additional damage in
<Location>Ferndale</Location>, <Location>Fortuna</Location>
and <Location>Scotia</Location>-- but no report of
additional injuries. The shock was felt as far away as
<Location>San Francisco</Location>.</paragraph>

<paragraph>At <Time>7:18 a.m. EDT</Time> on <Date>April
26</Date> a second aftershock registering 6.5 on the
Richter scale occurred approximately 23 miles west of
<Location>Cape Mendocino</Location>.</paragraph>

PERSON = (Nick Nikas, Roy Kite)

LOCATION = (Northern California, Eureka, Petrolia,
Ferndale, Fortuna, Scotia, San Francisco, Cape
Mendocino, Highway 254, Avenue of the Giants, San
Andreas, Mendocino Fracture Zone, Cascadia Subduction
Zone, Rio Dell, Route 101, Humboldt County, Fresno,
Garberville)

ORGANIZATION = (Pacific Gas and Electric, PG&E,
California Office of Emergency Services, FEMA, CAL OES,
National Guard, Pacific Lumber Company, CALTRANS,
Conservation Corps, Department of Water Resources, Air
National Guard, Emergency Support Team, US Army Corps
of Engineers, US Coast Guard, Salvation Army, Red Cross,
Humboldt County Board of Supervisors)

File Unit Description

A memorandum dated June 7, 1990 from John Niehuss to Stephen Janzansky regarding World Bank Green Fund.

A memorandum dated August 16, 1990 from Greg Petersmeyer to Nicholas Brady, Richard Jarman, and Michael Boskin regarding Charitable Deductions.

A memorandum dated September 18, 1990 from Ede Holiday to John Sununu regarding DOE's concerns on White House Process

This file unit contains Cabinet Documents including memoranda relating to the World Bank Green Fund, Charitable Deductions and DOE's concerns on White House Process.

Series Description

This file unit contains Cabinet Documents including memoranda relating to the World Bank Green Fund, Charitable Deductions and DOE's concerns on White House Process.

This file unit contains materials relating to the 1992 Petrolia, California Earthquake. It includes memoranda, situation reports and correspondence.

This series consists of Cabinet Documents including memoranda relating to the World Bank Green Fund, Charitable Deductions and DOE's concerns on White House Process. This series also consists of memoranda, situation reports and correspondence relating to the 1992 Petrolia, California Earthquake.