Grammatical Induction and Recognition of the Documentary Form of Records

William Underwood

Sheila Isbell and Matthew Underwood

Digital Curation Symposium, Chapel Hill, NC

April 19-20, 2007

Documentary Forms: Definitions

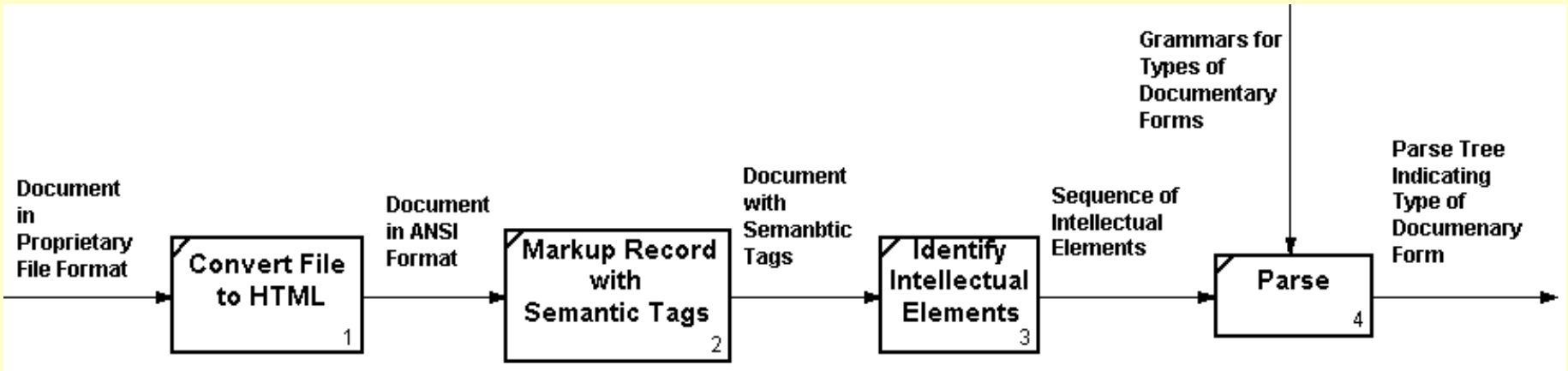
- ***Documentary form*** **is "the rules of representation used to convey a message, that is, the characteristics of a document which can be separated from the determination of the particular subjects, or places it concerns. Documentary form is both physical and intellectual.**
- **The *intellectual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

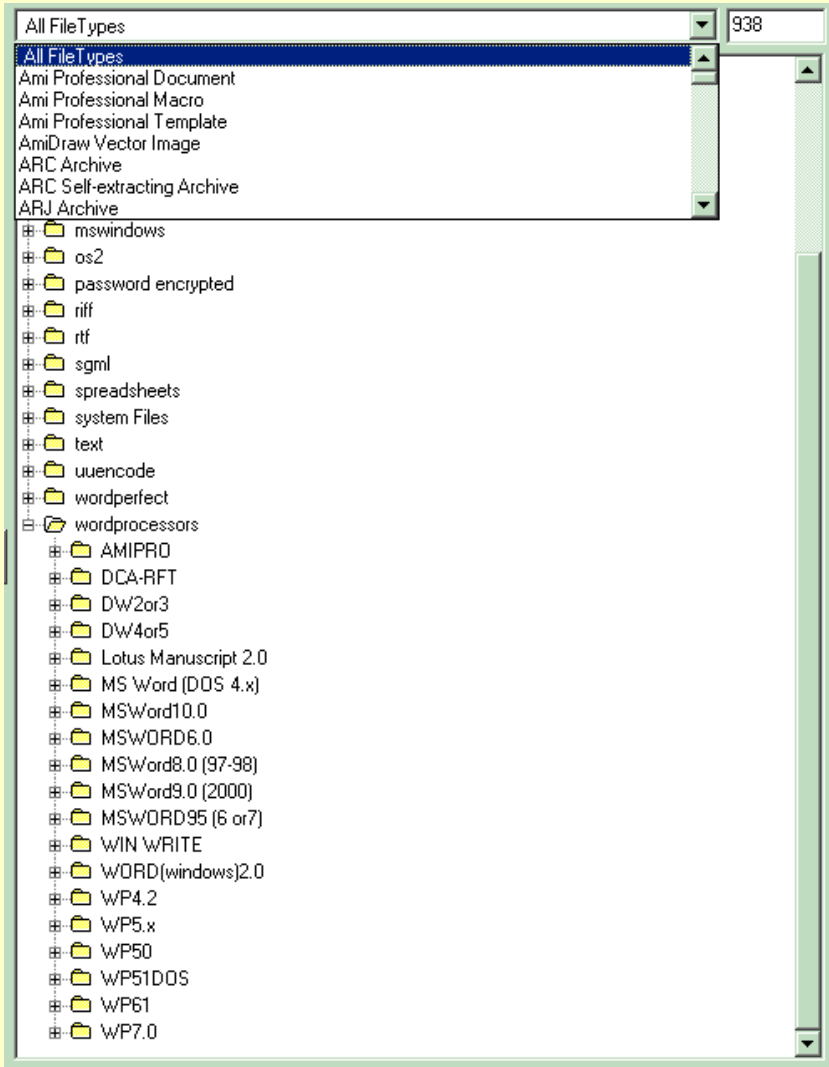
Documentary Forms: Examples from the Bush PC Records

- Agenda
- Attendee List
- Bar Chart
- Biography
- Briefing (Presentation)
- 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
- Newswire
- Nomination to Federal Office
- Notes
- Presidential Statement
- Press Pool Report
- Press Release
- Referral Memo
- Resume
- Schedule
- Signature Memo
- Situation Report
- Summary
- Transcript of Speech
- Staff Register
- Telephone Call Recommendation
- Transcript of News Conference

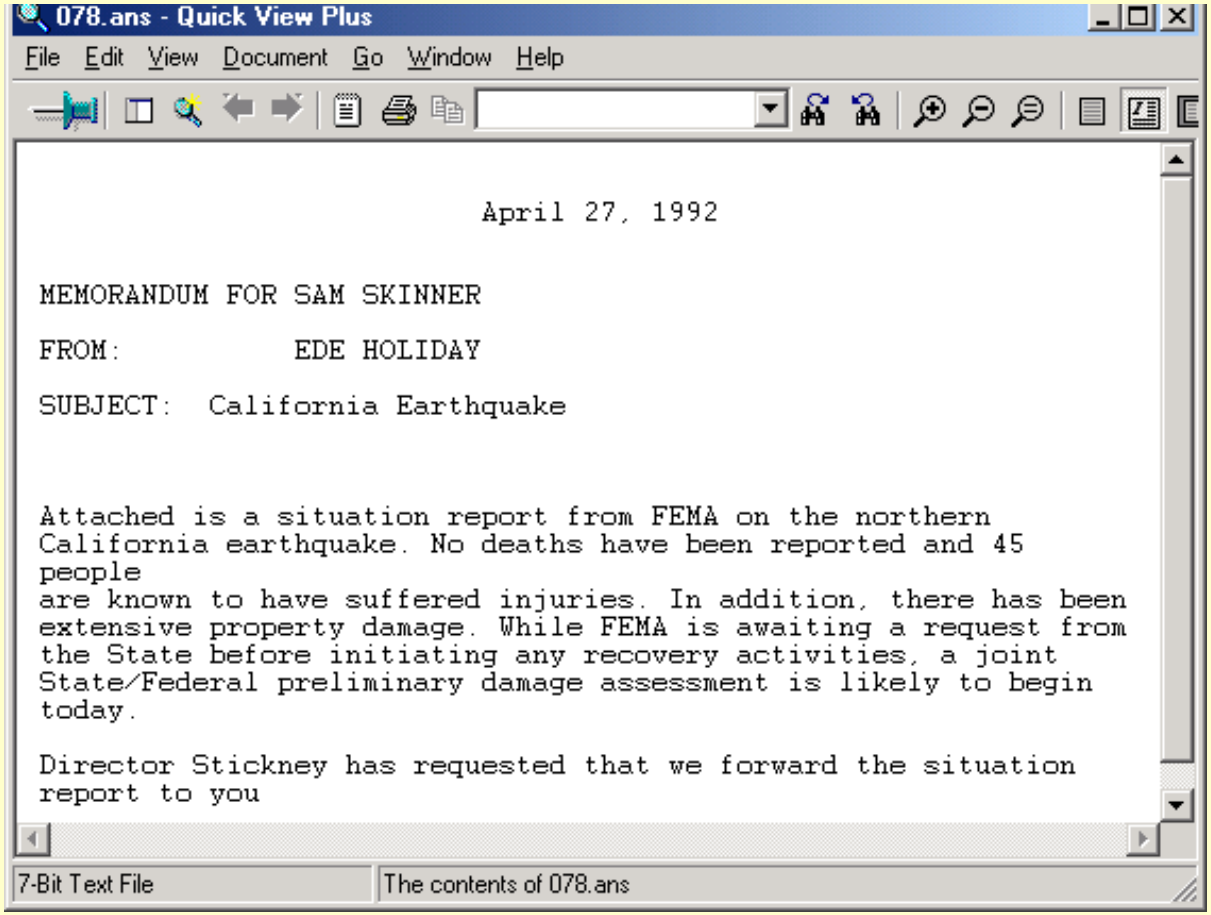
Documentary Form: Automatic Recognition



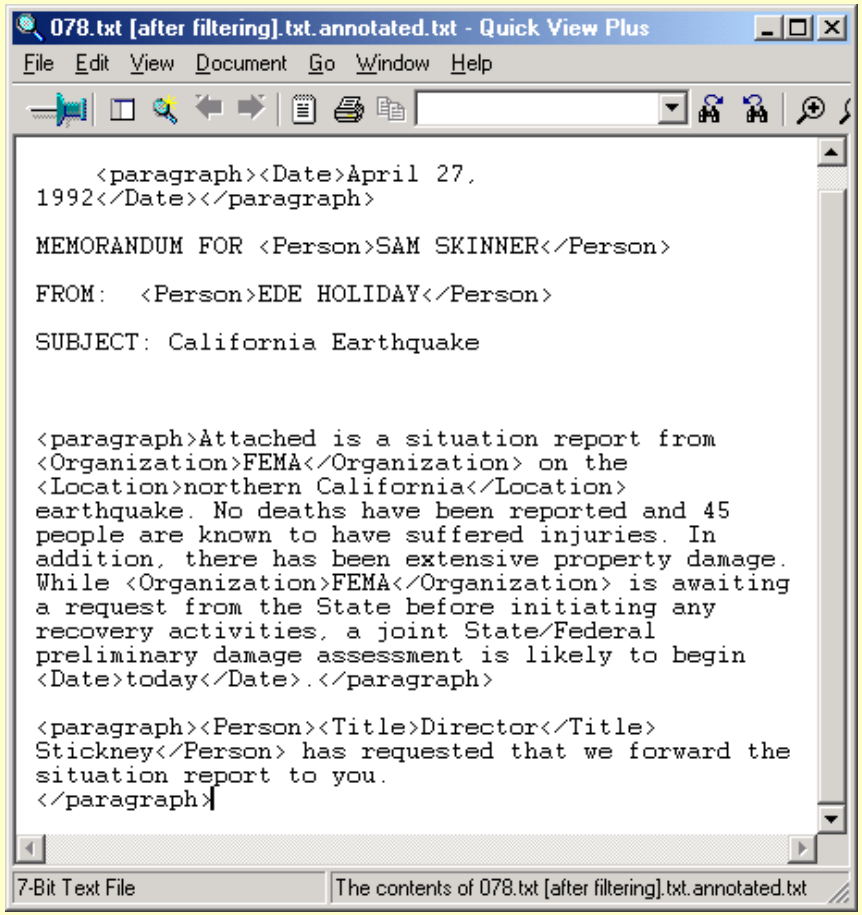
File Fornat Type Identifier



Documentary Form: Document Converted to ANSI Text



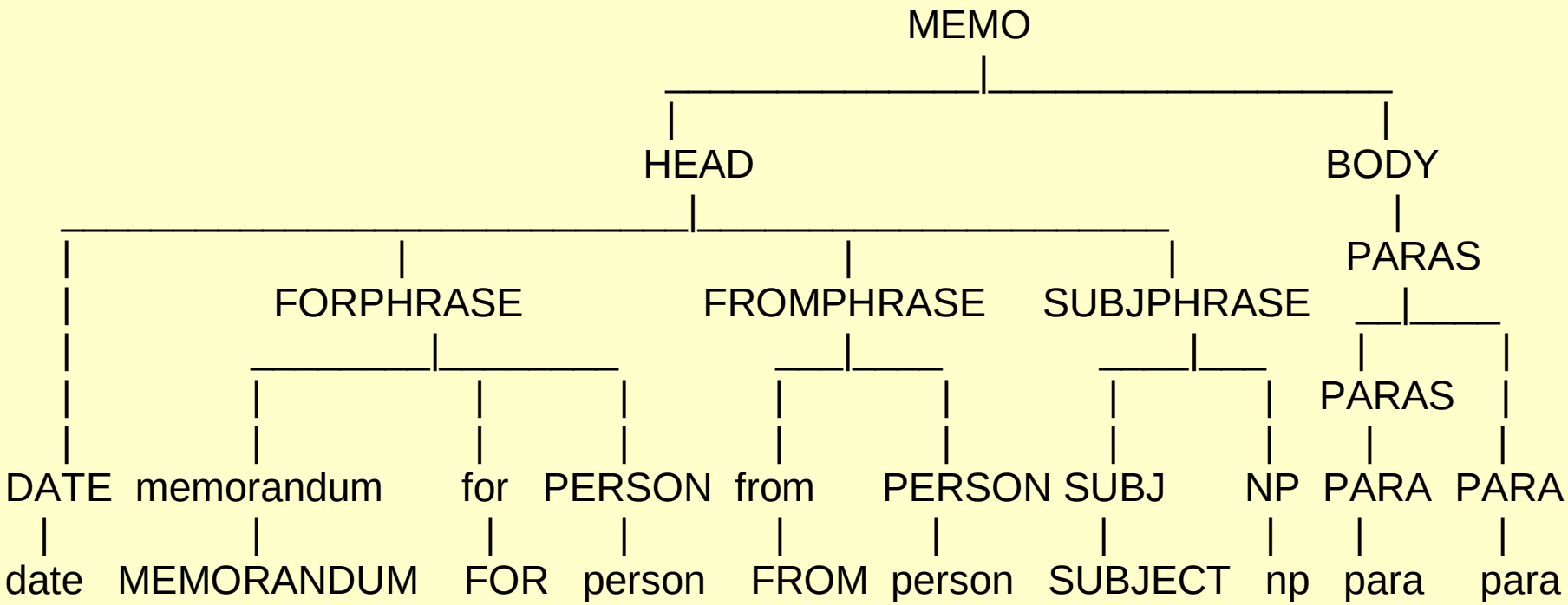
Documentary Form: Annotated Document



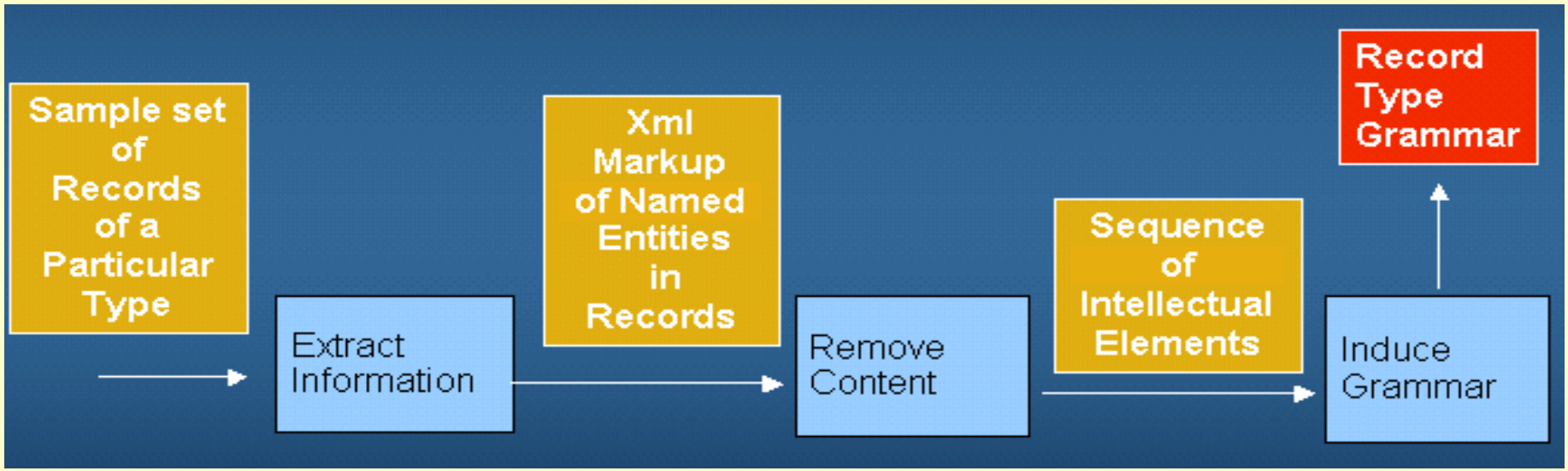
Documentary Form: Grammar for Memoranda

MEMO -> HEAD BODY
HEAD -> DATE FORPHRASE FROMPHRASE SUBJPHRASE
FORPHRASE -> memorandum for PERSON
FROMPHRASE -> from PERSON
SUBJPHRASE -> SUBJ NP
BODY -> PARAS
PARAS -> PARAS PARA
PARAS -> PARA
DATE -> date
memorandum -> MEMORANDUM
for -> FOR
PERSON -> person
from -> FROM
SUBJ -> SUBJECT
NP -> np
PARA -> para

Documentary Form: Parse Tree Indicating a Memo



Documentary Form: Induction of Grammar from Samples of a Document Type



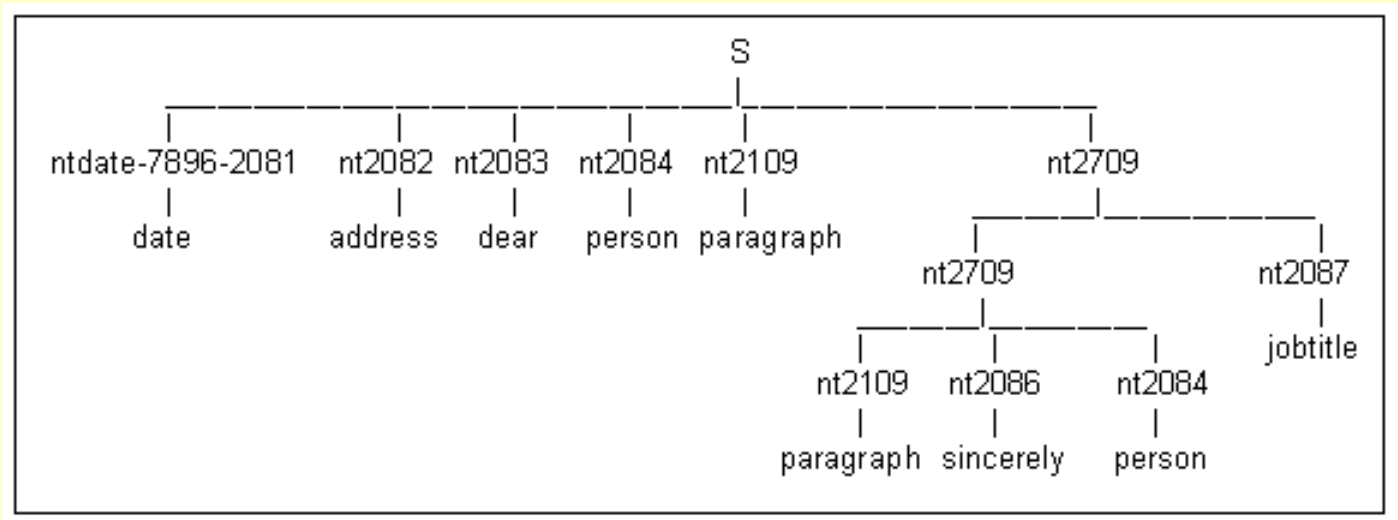
Documentary Form: Sample of Intellectual Forms of Correspondence

(DATE Dear PERSON PARAGRAPH PARAGRAPH Warmly ADDRESS);;C:\testDir\001.txt
(DATE ADDRESS Dear PERSON PARAGRAPH Devotedly PERSON);;C:\testDir\005.txt
(DATE ADDRESS Dear PERSON PARAGRAPH PARAGRAPH PARAGRAPH PARAGRAPH PARAGRAPH PARAGRAPH PARAGRAPH PARAGRAPH PARAGRAPH PARAGRAPH PARAGRAPH PARAGRAPH PARAGRAPH PARAGRAPH Sincerely PERSON);;C:\testDir\007.txt
(DATE ADDRESS Dear PERSON PARAGRAPH PARAGRAPH PARAGRAPH PARAGRAPH PARAGRAPH PARAGRAPH PARAGRAPH PARAGRAPH PARAGRAPH PARAGRAPH With best regards Sincerely PERSON);;C:\testDir\008.txt
(DATE ADDRESS Dear PERSON PARAGRAPH PARAGRAPH PARAGRAPH All Best PERSON);;C:\testDir\009.txt
(DATE ADDRESS Dear PERSON PARAGRAPH PARAGRAPH Sincerely PERSON);;C:\testDir\016.txt
(DATE Dear PERSON PARAGRAPH PARAGRAPH PARAGRAPH With best regards Sincerely PERSON JOBTITLE ADDRESS);;C:\testDir\029.txt
(Dear PERSON PARAGRAPH PARAGRAPH PARAGRAPH PARAGRAPH With best regards Sincerely PERSON JOBTITLE ADDRESS);;C:\testDir\031.txt
(DATE Dear PERSON PARAGRAPH PARAGRAPH PARAGRAPH Sincerely PERSON JOBTITLE ADDRESS);;C:\testDir\037.txt
(DATE ADDRESS Dear PERSON PARAGRAPH PARAGRAPH PARAGRAPH Sincerely PERSON JOBTITLE);;C:\testDir\059.txt
(DATE Dear PERSON PARAGRAPH PARAGRAPH PARAGRAPH Sincerely PERSON JOBTITLE ADDRESS);;C:\testDir\068.txt
(DATE ADDRESS Dear PERSON PARAGRAPH PARAGRAPH PARAGRAPH PARAGRAPH PARAGRAPH PARAGRAPH Sincerely PERSON JOBTITLE);;C:\testDir\070.txt
(DATE Dear PERSON PARAGRAPH PARAGRAPH PARAGRAPH With best regards Sincerely PERSON JOBTITLE ADDRESS);;C:\testDir\072.txt
(DATE Dear PERSON PARAGRAPH PARAGRAPH PARAGRAPH With best regards Sincerely PERSON JOBTITLE ADDRESS);;C:\testDir\076.txt
(DATE Dear PERSON PARAGRAPH PARAGRAPH Sincerely PERSON JOBTITLE ORGANIZATION ADDRESS);;C:\testDir\080.txt
(DATE Dear PERSON PARAGRAPH PARAGRAPH Sincerely PERSON JOBTITLE ADDRESS);;C:\testDir\082.txt
(DATE Dear PERSON PARAGRAPH PARAGRAPH PARAGRAPH Sincerely PERSON JOBTITLE ADDRESS);;C:\testDir\084.txt
(DATE Dear PERSON PARAGRAPH PARAGRAPH PARAGRAPH Sincerely PERSON JOBTITLE ADDRESS);;C:\testDir\086.txt
(DATE Dear PERSON PARAGRAPH Warmly PERSON ADDRESS);;C:\testDir\106.txt
(DATE Dear PERSON PARAGRAPH PARAGRAPH PARAGRAPH Warmly PERSON ADDRESS);;C:\testDir\107.txt
(DATE Dear PERSON PARAGRAPH PARAGRAPH Sincerely PERSON JOBTITLE ADDRESS);;C:\testDir\112.txt

Documentary Form: Induced Grammar for the Documentary Form of Correspondence

```
NT2219     (NT2109 NT2109)                                    1.000     ;43.0000
NT2239     (NT2219 NT2219 NT2219)                             1.000     ;12.0000
NT2692     (NT2692 NT2219)                                    0.325     ;7.00000
           (NTDATE-7896-2081 NT2083 NT2084)                   0.675     ;14.0000
NT2709     (NT2109 NT2086 NT2084)                             0.349     ;13.0000
           (NT2109 NT2709)                                    0.120     ;5.00000
           (NT2121 NT2084)                                    9.143e-2;4.00000
           (NT2239 NT2709)                                    0.206     ;8.00000
           (NT2709 NT2087)                                    0.234     ;9.00000
S        (NT2692 NT2121 NT2082)                               1.316e-2;1.00000
         (NT2692 NT2239 NT2086 NT2084 NT2087 NT2082) 0.171 ;4.00000
         (NT2692 NT2709 NT2082)                               0.434     ;9.00000
         (NTDATE-7896-2081 NT2082 NT2083 NT2084 NT2109 NT2709) 0.382 ;8.0
NT2082       address                                          1.000     ;22.0000
NT2083       dear                                             1.000     ;22.0000
NT2084       person                                           1.000     ;43.0000
NT2086       sincerely                                        1.000     ;17.0000
NT2087       jobtitle                                         1.000     ;13.0000
NT2109       best                                             3.899e-2;5.00000
             paragraph                                        0.883     ;97.0000
             regards                                          3.899e-2;5.00000
             with                                             3.899e-2;5.00000
NT2121       devotedly                                        0.375     ;2.00000
             warmly                                           0.625     ;3.00000
NTDATE-7896-2081    date                                      1.000     ;22.0000
```

Documentary Form: Parse Tree Indicating Form of White House Correspondence



Automatic Description

Item Description

File Unit (Folder) Description

Series Description

Item Description

Date = April 27, 1992

For = SAM SKINNER

From = EDE HOLIDAY

Subject = California Earthquake

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

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.

Use of these Tools in Selection and Appraisal

- Use file format type identifier to identify files that cannot be preserved without conversion to other formats.
- Instead of sampling an e-record series, automatically identify all document types and generate file unit and series descriptions.
- Use grammatical induction for learning documentary form of web pages.
- After transfer, before accession verify that what is received is what is expected.

Use of these Tools in Other Archival Activities

- Review
  - Identifying records subject to restrictions on disclosure
- Description
  - Document type recognition and automatic description enable earlier intellectual control of large volumes of e-records
- Search and Retrieval
  - Index records on document type, and elements such as chronological date, participants in communication, and subject or actions and support search and retrieval on these.

Research Status

- Grammars for about 20 document types
- Need samples of 100 or more of each document type to effectively apply grammatical induction
- Recognizing the communication act performed by a written record – requesting, informing, resigning, appointing, nominating. (about 300 speech acts)
- Extending method to recognize physical form