The International Journal of Digital Curation
Issue 1, Volume 5 | 2010

Grammar-Based Recognition of Documentary Forms and Extraction of Metadata

William Underwood,
Principal Research Scientist,
Georgia Tech Research Institute

Abstract

Metadata extraction is a critical aspect of ingestion of collections into digital archives and libraries. A method for automatically recognizing document types and extracting metadata from digital records has been developed. The method is based on a method for automatically annotating semantic categories such as person's names, job titles, dates, and postal addresses that may occur in a record. It extends this method by using the semantic annotations to identify the intellectual elements of a document's form, parsing these elements using context-free grammars that define documentary forms, and interpreting the elements of the form of the document to identify metadata such as the chronological date, author(s), addressee(s), and topic. Context-free grammars were developed for fourteen of the documentary forms occurring in Presidential records. In an experiment, the document type recognizer successfully recognized the documentary form and extracted the metadata of two-thirds of the records in a series of Presidential e-records containing twenty-one document types.[1]

[1] This paper is based on the paper given by the author at the 5th International Digital Curation Conference, December 2009; received November 2009, published June 2010.
The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. ISSN: 1746-8256 The IJDC is published by UKOLN at the University of Bath and is a publication of the Digital Curation Centre.

Introduction

The increasing volume of digital records being acquired by archives and libraries poses significant challenges to archivist's manual procedures for processing records. Archivists traditionally describe records at record group (or collection), series, file unit and item levels. This provides the archive's intellectual control over its holdings and supports access to the records. Archival descriptions (summaries or metadata) include the names of the types of records that occur in a record series, for example, correspondence, memoranda or agenda. Record descriptions also include author's and addressee's names as well as the topics of records. Archivists cannot completely 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. An automated method of metadata extraction and description is needed.

The next section of this paper reviews the concept of documentary form and related concepts. The related research in document type (or genre) identification is summarized. Then the method for recognizing documentary forms and extracting document metadata is described. An implementation and experimental evaluation of the method is described. Finally, the results of the research are summarized with a discussion of open research issues.

Documentary Form, Record Types and Document Type

The International Council of Archivists (1999) 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". 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" (NARA, 2008). The purpose of the specific records type is that it "Enables users to search for archival materials by the types of document represented in the archival materials". The guidelines also specify that specific records types be used in describing record series.

The science of diplomatics defines *documentary form* as "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" (Duranti, 1998). 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" (Duranti, 1998).

The Standard Generalized Markup Language (SGML) uses a Document Type Definition (DTD) to define document form (International Standards Organization, 1986). A DTD specifies a set of elements, their relationships, and the tag set is used to markup the document. The Extensible Markup Language (XML) is a simpler subset of SGML (World Wide Web Consortium, 2006). 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.

The concept of genre is similar to that of documentary form but includes classes of documents that are not characterized by their intellectual or physical form, but by pragmatic or rhetorical features. Examples of written genre include academic prose, biography, instructional material and newspaper reports. See Santini (2004b) for a discussion.

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.

Related Research

The reader is referred to Santini (2004b) for a survey of state-of-the-art approaches to genre identification of digital documents. Santini (2004a) also describes a method based on part-of-speech trigrams for classifying ten genres including conversations, interviews, public debate, biography and reportage. The objective of the research of Kim and Ross (2007a, 2007b) is the recognition of genre for the purpose of metadata extraction from digital records ingested into digital archives or libraries. Their approach is to identify features of documents that will allow them to automatically classify documents by genre. The features they have identified include: image features, syntactic features, stylistic features, semantic structure, and domain knowledge features. These features are used with an image classifier, an n–gram model classifier and a stylo-metric classifier. Our research differs from that of Kim and Ross, and of other researchers in genre identification, in that our objective is to recognize a document's form by parsing its intellectual elements using grammars characterizing document types. However, there are document types for which it is necessary to use pragmatic features to recognize the genre, for example, white papers and biography.

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 2. The italicized phrases to the right of the downward pointing arrows indicate inputs and outputs of the numbered processing steps (Underwood & Laib, 2008).

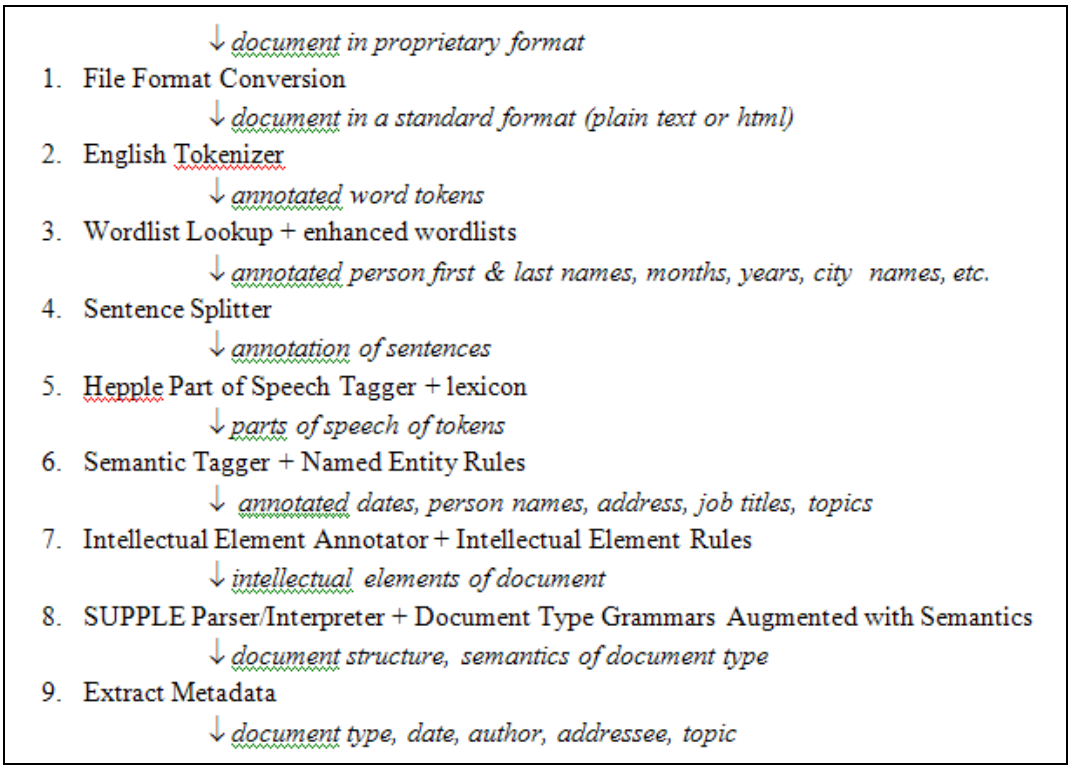

Figure 2. The Process of Document Type Recognition and Metadata Extraction.

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 (Underwood & Isbell, 2008). The input to the method is an e-record in a proprietary file format. The first step 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 3 shows a document whose paragraphs, dates, times, and person, location and organization names have been annotated by the first six steps of the method.

```
    <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>
```

Figure 3. 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 100 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 4 shows the document in Figure 3 after the annotation of the intellectual elements.

```
<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 4. Annotated Intellectual Elements.

The names of the intellectual elements shown in Figure 4 are *chron*(*ological*)*date*, *for*, *person*, *from*, *subj, topic*, *para* and *attachment*.

The eighth step, SUPPLE Parser/Interpreter (Gaizauskas, Hepple, Saggion, Greenwood, & Humphreys, 2005), recognizes the document type using a parse/interpreter with a context-free grammar that characterizes 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 natural language and Computer Scientists use them to define programming languages.

Figure 5 shows some of 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 an 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 5. Grammar for the Intellectual Form of a Memorandum.

Figure 6 shows the grammar shown in Figure 5 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.

```
%% 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 6. Part of the Grammar for the Intellectual Form of a Memorandum Augmented with Semantic Rules.

The Intellectual Element Annotator assigns a value to each of the intellectual elements in the grammar. For example, for the annotated document in Figure 4, the intellectual element *PERSON* after the intellectual element *MEMORANDUM FOR* will get the value 'SAM SKINNER'.

In Figure 6, the two percent symbols (%%) indicate a comment. A grammar rule such as A $\rightarrow$ B_1, …B_n is represented to the parser by a rule of the form rule(A [B_1, … 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( ), …B_n( )]). For instance, the rule shown at the bottom of Figure 6 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 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.

A parser with grammars for many document types is applied to a document whose intellectual elements are 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 7 shows the parse tree for the document shown in Figure 4.

```
(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 7. Parse Tree for the Sample Memorandum.

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'.

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.

Implementation and Test of the Method

The method for recognizing documentary forms and extracting metadata has been implemented. Grammars have been developed for the following 14 document types.

Formal Letter	Presidential Determination
White House Casual Letter	Executive Order
White House Memorandum	Presidential Proclamation
Action-Decision Memorandum	National Security Directive
White House Referral	National Security Review
Recommended Telephone Call	Memorandum of Conversation
White House Press Release	Memorandum of Telephone Conversation

A corpus of 112 documents with examples of each of these 14 document types was constructed from paper records of Presidential records that are public records or have been reviewed and disclosed to the public. They include records from presidential administrations from Reagan to Obama. The records were scanned, OCRed and converted to file formats typical to the period in which they were created e.g. DisplayWrite, Word Perfect 5, Word 98). This corpus simulates the digital records created by Presidential administrations. The method has been applied to this corpus and it correctly identifies the documentary forms and extracts the associated metadata.

Experimental Evaluation of the Method

The National Archives and Records Administration has a collection of e-records from the administration of President George H. W. Bush. The collection consists of personal computer records from White House staff members and offices. These legacy computer records have not been reviewed for possible restrictions on disclosure, so are not yet available to the public.

A system called PERPOS has been prototyped that supports accession, archival processing, and storage and retrieval of such records. This prototype provides an environment for experimental evaluation of new techniques for preserving, describing reviewing and retrieving e-records (Underwood, Laib and Hayslett-Keck, 2006).

The method described in this paper is being experimentally evaluated by applying it to series of presidential e-records that have been accessioned into PERPOS. PERPOS provides the facility for converting files in legacy file formats to plain text or html and for associating the metadata with records to which the method has been applied.

The table in Figure 8 shows the results of one of the early experiments on a series of e-records. The document types and number of documents of each type were manually determined. The series contains twenty-one document types including eighteen for which grammars have not yet been constructed. The method recognized the document type and extracted the metadata for the three document types whose form had been defined and for two-thirds of the records in the series. The records are predominantly White House Memoranda and White House Casual Correspondence. Those memoranda not recognized include a memo without a subject and memos through two people, rather than a single person as specified by the grammar.

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

Figure 8. Results of the method applied to record series 113.

The White House casual letters that were not recognized were due to the semantic category annotator failing to recognize a postal address, a person's name or job title. This problem can be addressed by improving the performance of the semantic annotator.

Conclusions

The results of this research are that: (1) the intellectual elements of documentary forms can be defined in terms of the 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 fourteen of the documentary forms that occur in Presidential e-records. Rules were constructed for recognizing the intellectual elements of these documentary forms. These grammars 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.

The intellectual elements of a documentary form were identified either by reference to a style manual for the form or by comparing examples of a document type to identify those elements of the examples that did not change from one example to another. The question arises, can the intellectual elements of a particular documentary form be learned from examples without a teacher? The question also arises, could grammatical induction be used with samples of a particular documentary form to induce a grammar automatically rather than manually? This would eliminate the manual effort needed to construct grammars from large samples, and could provide a method for automatically refining the grammar when examples of a documentary form were encountered that did not fit the current grammatical model. It would also facilitate the extension of documentary form recognition to a larger number of documentary forms.

Underwood and Harris (2006) demonstrated that it is possible to induce a grammar for the documentary form of White House memoranda and correspondence from sequences of intellectual elements occurring in samples of these document types. One of the obstacles to progress in this research was that samples of document types were created from OCRed paper documents and the intellectual element recognizer had not been created. Now, the intellectual element recognizer and document type recognizer have been interfaced to PERPOS. There are hundreds of thousands of e-records in the PERPOS repository that can be used in grammatical induction experiments. The intellectual element recognizer is being modified to output the intellectual elements of a record for use in grammatical induction rather than in recognition. This research has not progressed to the point that there are experimental results to report.

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.

Acknowledgements

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.

References

Duranti, L. (1998) *Diplomatics: New uses for an old science* (pp.134). Society of American Archivists and Association of Canadian Archivists in association with Scarecrow Press.

Gaizauskas, R., Hepple, M., Saggion, H., Greenwood, M.A. & Humphreys, K. (2005, October). SUPPLE: A practical parser for natural language engineering applications. *International Workshop on Parsing Technologies,* Vancouver.

International Council of Archivists. (1999). *ISAD(G): General International Standard Archival Description* (pp.11), Second Edition.

International Standards Organization. (1986). *Standard Generalized Markup Language - ISO 8879:1986.*

Kim, Y. & Ross, S. (2007a, March). Detecting family resemblance: automated genre classification. *Data Science Journal, 6*(Supplement), 172-183.

Kim, Y. and Ross, S. (2007b, June). "The naming of cats": Automated genre classification. *The International Journal of Digital Curation, 2*(1), 49-61.

NARA. (2008). *Life Cycle Data Requirements Guide (LCDRG)* (pp.131).

Santini, M. (2004a). A shallow approach to syntactic feature extraction for genre classification. *Proceedings of the 7th Annual Colloquium of the UK Special Interest Group for Computational Linguistics*.

Santini, M. (2004b). *State-of-the-art on automatic genre identification* (Technical Report ITRI-04-03). University of Brighton, UK: Information Technology Research Institute (ITRI).

Underwood, W.E. & Harris, B. (2006, August). *Inferring and recognizing the documentary form of record types* (PERPOS TR ITTL/CSITD 05-08). Georgia Tech Research Institute. Retrieved from http://perpos.gtri.gatech.edu/publications/PERPOS%20TR%2005-08.pdf

Underwood, W & Isbell, S. (2008, May). *Semantic annotation of presidential e-records* (Technical Report TR ITTL/CSITD 08-01). Georgia Tech Research Institute. Retrieved from http://perpos.gtri.gatech.edu/publications/TR%2008-01.pdf

Underwood, W. & Laib S. (2008, May). *Automatic recognition of documentary forms* (Technical Report ITTL/CSITD 08-02). Georgia Tech Research Institute. Retrieved from http://perpos.gtri.gatech.edu/publications/TR%2008-02.pdf

Underwood, W., Laib, S. & Hayslett-Keck, M. (2006, September). *Reference manual for PERPOS: An electronic records repository and archival processing system*, version 3.1. (PERPOS TR ITTL/CSITD 06-2). Georgia Tech Research Institute. Retrieved from http://perpos.gtri.gatech.edu/publications/PERPOS%20TR%2006-02.pdf

World Wide Web Consortium. (2006). *Extensible Markup Language XML 1.0* (Fourth Edition).