The International Journal of Digital Curation
Issue 1, Volume 2 | 2007

"The Naming of Cats": Automated Genre Classification

Yunhyong Kim, Seamus Ross
Digital Curation Centre (DCC)
&
Humanities Advanced Technology and Information Institute (HATII)

June 2007

"The Naming of Cats is a difficult matter, It isn't just one of your holiday games; You may think at first I'm as mad as a hatter, When I tell you, a cat must have three different names." - T.S. Eliot, The Naming of Cats

Abstract

This paper builds on the work presented at the ECDL 2006 in automated genre classification as a step toward automating metadata extraction from digital documents for ingest into digital repositories such as those run by archives, libraries and eprint services (Kim & Ross, 2006b). We have previously proposed dividing features of a document into five types (features for visual layout, language model features, stylometric features, features for semantic structure, and contextual features as an object linked to previously classified objects and other external sources) and have examined visual and language model features. The current paper compares results from testing classifiers based on image and stylometric features in a binary classification to show that certain genres have strong image features which enable effective separation of documents belonging to the genre from a large pool of other documents.

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.

Background and Objective

In (Kim & Ross, 2006b) we summarised the valuable role of automated metadata extraction in the cost-effective efficient management of digital collections: metadata play a key role in management processes (Hedstrom et al., 2003; Ross & Hedstrom, 2005) and the manual creation of metadata is expensive (DELOS/NSF Working Groups, 2003; Hedstrom et al., 2003; PREservation Metadata: Implementation Strategy Working Group (PREMIS), n.d.). As we pointed out in (Kim & Ross, 2006b), ERPANET's (Electronic Resources Preservation Access Network, n.d.) Packaged Object Ingest Project (ERPANET: Packaged Object Ingest Project, n.d.) identified automatic extraction tools for technical metadata (e.g. National Archives UK Digital Object Identification (DROID, n.d.; National Library of New Zealand: Metadata Extraction Tool, n.d.), and there has been substantial work on descriptive metadata extraction within specific domains (e.g. Automatic Metadata Generation, n.d.; DC-dot, n.d.; Giuffrida, Shek, & Yang, 2000) and also in (MetadataExtractor, n.d.; Thoma, 2001) along with other work in information extraction from text (Arens & Blaesius, 2003; Bekkerman, McCallum, & Huang, 2004; Breuel, 2003; Ke, Bowerman, & Oakes, 2006) and also (Sebastiani, 2002; Shafait, Keysers, & Breuel, 2006; Shao & Futrelle, 2005; Witte, Krestel, & Bergler, 2005). However, a general tool has yet to be developed to extract metadata from digital objects of varied types and genres. This paper develops concepts of genre classification, i.e. the automatic grouping of documents into distinctive document structures, to be followed by focused metadata extraction from single structural types, as a means of creating a tool capable of extracting metadata across many domains at different semantic levels. To reiterate the argument in (Kim & Ross, 2006b), identifying the genre first provides a mechanism to limit the scope of document forms from which to extract other metadata. Within a single genre, metadata such as author, title, keywords, identification numbers or references can be expected to appear in a specific style and region, and independent methods have been developed for genre-specific extraction of such metadata for some classes of documents (e.g. scientific papers). Note also that different institutions focus on collecting and managing digital materials in different genres; genre classification will support automating the identification, selection, and acquisition of materials in keeping with local collecting policies.

A review of Biber (1995), Karlgren and Cutting (1994), Kessler, Nunberg and Schuetze (1997), Rauber and Müller-Kögler (2001), Bagdanov and Worring (2001), Boese (2005), Finn and Kushmerick (2006) and Santini (2004a) exemplifies the lack of consensus on the definition of genre. Biber's analysis of document genres employed five functional dimensions (information, narration, elaboration, persuasion, abstraction) to characterise text, while Karlgren et al. and Boese concentrated on more popularly accepted genre classes such as FAQ, Job Description, Editorial or Reportage. Kessler et al. tried to address both types, while Finn et al. studied binary classifications of two aspects (fact versus opinion, positive versus negative reviews). Santini discussed general genre facets and Bagdanov limited his task to detecting specific journals and brochures. Others (Barbu, Heroux, Adam, & Trupin, 2005; Rauber & Müller-Kögler, 2001) attempted the clustering of documents rather than classification. An overview of the various efforts in genre analysis can be found in a technical report by Santini (2004b). A broader review of metadata extraction and genre classification is also being prepared by the DELOS NoE Digital Preservation Cluster and is expected to be completed before the publication of this paper.

The variety of definitions adopted by these researchers illustrates a confused interplay of two notions: one of structure and one of function. Structure is defined by the visual layout and is expected to be distinguishable mostly by measurable features such as amount of white space; the length of the document, sentences, or words; and, the presence or absence and location of headers, delimiters, images, or links. Function, on the other hand, is defined by the intended role of the document and is expected to be characterised mostly by linguistic models and semantic analyses of the documents. The two notions are closely linked together by medium, process or event. For example, a scientific research article is usually sructured so that a title is present on the first page followed by author, affiliation, a body of text consisting of sections, and finally a list of references. It has the function of communicating, arguing or describing research. The interrelationship of structure and function are represented by the formatting requirements of journals or conventions in the community or event for which the document was created. The requirements and conventions evolve to optimise the communicative intentionality within the context; other communities or events may find different structures of documents to optimise the same function. Just as biologists study DNA as the building blocks of living organisms to understand the classes into which they have evolved within their environment, it seems important to identify documents by their structure and their function separately as building blocks to infer their genre class within a standardised schema. We seek to be able to achieve this by a full analysis of five document feature types: image features, syntactic features, stylistic features, semantic structure, and domain knowledge features. We aim to build a system which models the five feature sets for a schema of approximately seventy genres (Table 1).

The genres in Table 1 are not meant to be static: the schema has been evolving as we develop and incorporate well-structured classification standards and as we become aware of digital genres we had not encountered before or which have just emerged in the digital domain. The experiments in this paper have initially limited the study to the image and stylistic feature sets on the nineteen most prolific genres in our experimental dataset. Along with the results in (Kim & Ross, 2006b), the results here are intended to be another step towards a full analysis.

The experimental data in this paper is from the pool of 570 PDF (Adobe Acrobat PDF specification, n.d.) files that were sampled randomly from the Internet as described in (Kim & Ross, 2006b). As explained in (Kim & Ross, 2006b), by confining the work to studying PDF files, we hope to put a boundary on the problem space, while working with a widely used portable format for digital objects ingested into digital repositories.

This paper, along with (Kim & Ross, 2006a, 2006b), is intended to show the promise of combining separate classifiers trained on different types of features for genre classification. Also note that the bottom-up approach of starting from genre-specific extraction may result in several tools which are overly dependent on the structures of the documents in the domain, with no obvious means of interoperability: the top-down approach of creating a tool which looks across genres, to be refined further within the domain, will enable us to avoid this problem.

Groups	Genres
Book	Academic book, Fiction, Poetry, Handbook, Other book
Article	Abstract, Scientific research article, Other research article, Magazine article, News report
Short Composition	Fictional Piece, Poems, Dramatic Script, Essay, Short Biographical Sketch, Review
Serial	Periodicals (Newspaper, Magazine), Journals, Conference Proceedings, Newsletter
Correspondence	Email, Letter, Memo, Telegram
Treatise	Thesis, Business/Operational report, Technical report, Misc. report
Information Structure	List, Catalogue, Raw Data, Table Calendar, Menu, Form, Programme, Questionnaire, FAQ
Evidential Document	Minutes, Legal proceedings, Financial Record, Receipt, Slips, Contract
Visually Dominant Document	Artwork, Card, Chart, Graph, Diagram, Sheet Music, Poster, Comics
Other Functional Document	Guideline, Regulations, Manual, Grant/Project Proposal, Legal Appeal/Proposal/Order, Job/Course/Project Description, Product/Application Description, Advertisement, Announcement, Appeal/Propaganda, Exam/Worksheet, Fact sheet, Forum Discussion, Interview, Notice, Résumé/CV, Slides, Speech transcript

Table 1 Scope of genres

Classifiers

The experiments described in this paper involve the use of six classifiers. Each classifier is defined by one of two feature types and one of three statistical modelling methods. These statiscal methods include Naïve Bayes (NB), Support Vector Machine (SVM) and Random Forest (RF) which are all available with the Weka machine learning toolkit (Witten & Frank, 2005). The two feature types, to be considered in combination with these three statistical methods, are denoted **image features** and **stylometric features**, and are described below.

Image features: this depends on features extracted from the PDF document when handled as an image. It uses the module pdftoppm from XPDF (Noonberg, n.d.) to extract the first page of the document as an image. The resulting image is divided into a sixty-six by sixty-six grid[1]. Then Python’s Image Library (PIL) (Python, n.d.; Python Imaging Library, n.d.) is employed to extract pixel values in each region. Each region is given a value of 0 or 1 depending on the amount of non-white pixel values it contains.

Stylometric features: this looks at the frequency of selected words, number of font changes, the difference between the largest font size and smallest font size, length of the document , average length of words, and number of words in the front page of the document. The font information was extracted on the level of words using a modified version of *pdftohtml* (PDFTOHTML, n.d.), developed by Volker Heydegger at the University of Cologne. The modified version converts a PDF document to a XML file with all the font information for each word in the document. A word list was automatically constructed containing all words which appear in more than half of the

1 The choice of the dimension reflects the fact that it seemed to produce the best results at the time but further analysis may be necessary.

files in any one genre from a dataset which had been set aside. For each file, the frequency of each word was recorded as a vector then augmented by length and font information.

This paper expresses the view that the image along with the stylistic features will capture the structural elements of genres while the language model combined with the stylistic and semantic features will help to separate documents of distinct functional categories. Involving the image of a document in the classification also enables the management of documents without violating security, maximises the viability of a language-independent tool, frees the process from being solely dependent on text-processing tools with encoding requirements and problems relating to special characters[2], and makes the method applicable to paper documents that are digitally imaged (i.e. scanned).

Experiments

There are three main experiments described in this paper:

Clustering experiment: this experiment compared the cluster resolution for two sets of features: the image features and the stylometric features. We grouped the data in nineteen genres into two clusters using the Weka Machine Learning Toolkit's (Witten & Frank, 2005) Estimation-Maximisation algorithm. The purpose was to see how well the files in each genre group into one cluster. The result is expressed in terms of the major percentage of files within each genre which have been grouped into one cluster.

Periodicals versus Thesis: in this experiment, we took documents in the genres Periodicals and Thesis. We used the image features with Naïve Bayes to classify the documents in a 10-fold cross-validation experiment.

Periodicals versus Non-periodicals: we expanded on the experiment above to group 271 examples from eighteen additional genre classes with the examples from the genre Thesis as one group of 289 documents labelled Non-periodicals. The eighteen supplementary classes include Academic Book, Other Book, Business Report, Factsheet, Fictional Book, Forms, Instruction Guideline, Job description, List, Minutes, Newsletter, Magazine Article, Scientific Research Article, Other Research Article, Product Description, Slides, Technical Report and seventeen other miscellaneous documents. The result was examined with six classifiers (the three statistical methods NB, SVM and RF in combination with the image and stylometric features) in a 10-fold cross-validation experiment.

2 *pdftohtml* failed to extract information from seventeen percent of the documents. The image processing did not fail on any documents.

Results

Table 2 shows the results of the clustering experiment. The key finding in this experiment is that the genres for which image features fail to cluster are the genres for which stylometric features cluster very well. For instance, note that Scientific Research Articles divide half and half into each cluster with no preference when using the visual features while ninety-two percent group into one cluster when using stylometric features. The opposite is true of Periodicals.

The results described in Tables 3 and 4 use three standard indices in classification tasks: accuracy, precision and recall. Let N be the total number of documents in the data, N_C the number of documents in the dataset which are in class C, T the total number of correctly labelled documents in the dataset independent of the class, $TP(C)$ the number of true positives for class C, and $FP(C)$ the number of false positives for class C. Accuracy is defined to be

$\frac{T}{N}$, and,

precision and recall for class C is defined to be

$\frac{TP(C)}{(TP(C)+FP(C))}$ and $\frac{TP(C)}{N_C}$,

respectively.

Genre Group	Genre	Visual	Stylistic
Book	Academic Book	87.5	60
	Fiction	87.5	83.3
	Other Book	75	82.4
Article	Scientific Research	50	92
	Other Research	90	73.7
	Magazine	62.5	84.6
Serial	Periodicals	94.7	62.5
	Newsletter	74.1	83.3
Treatise	Thesis	100	90
	Business Report	66.7	90.9
	Technical Report	68.2	72.2
Information Structure	List	68.4	85.7
	Form	68.8	69.2
Evidential Document	Minutes	94.7	76.9
Other Functional Document	Instruction/ Guideline	90.5	50
	Job/Course/Project Description	50	66.7
	Product/Application Description	61.1	68.8
	Factsheet	53.3	78.6
	Slides	60	91.7

Table 2 A Comparison of Visual and Stylometric Clusters (percentage of files in one cluster)

Table 3 illustrates the result when the dataset was confined to Periodicals and Theses. To check if the high accuracy found in this experiment actually reflects the distinctiveness of image features in periodicals, the experiment was repeated with eighteen additional classes (described in the section "Experiments" above) of non-periodicals added to Thesis to form a class Non-periodicals. The classifiers based on image showed an overall decrease (c.f. Naïve Bayes performance of Table 4 and figures in Table 3), but the recall rate (indicated in bold-face), with respect to Periodicals, of the image Naïve Bayes classifier is still quite high. The best overall performer is the classifier on stylometric features and SVM, however, the precision of the classifier based on Random Forest with stylometric features is excellent. The recall rates of the classifiers based on stylometric features are consistently poor across all three statistical methods.

10-fold Cross Validation with the Image classifier, Overall accuracy: 97.26 %

Genres	Precision	Recall
Periodicals (19 items)	1	0.947
Thesis (18 items)	0.947	1

Table 3 Distinguishing Periodicals from Thesis using image features

The results in Tables 2 and 3, indicate that Periodicals and Thesis have strong distinguishing image features in our dataset. The recall of the image Naïve Bayes classifier (Table 4) suggests that the image features of periodicals, though shared by other documents, are more distinctly associated to the genre class peridodicals. The low recall rate of all three stylometric classifiers, with respect to Periodicals, suggest that the characterising stylometric features of periodicals captured by these models do not generalise to other periodicals in our dataset.

Naïve Bayes (NB), Support Vector Machine (SVM) and Random Forest (RF),10 fold cross validation. Overall accuracy:

Statistical Method	Genres	Precision image	Recall image	Precision style	Recall style
NB	Periodicals (19 items)	0.298	**0.875**	0.148	0.25
	Non-Periodicals (289 items)	0.991	0.759	0.957	0.92
SVM	Periodicals (19 items)	0.333	0.188	0.714	0.313
	Non-Periodicals (289 items)	0.956	0.979	0.963	0.993
RF	Periodicals (19 items)	0.571	0.25	**1**	0.125
	Non-Periodicals (289 items)	0.96	0.99	0.954	1

Table 4 Distinguishing Periodicals from Non-periodicals comparing image and stylometric features

We have not provided the overall accuracies in Table 4, because, it is our belief that, on a skewed dataset, in which one class completely overpowers the other class in number, the overall accuracy is of secondary importance as a performance measure. For instance, suppose only one document, in a set of 100 documents, belongs to a pre-defined class A, and the objective is to construct a system which can automatically locate the document belonging to A. A system which labels all documents as not belonging to A has an overall accuracy of 0.99 as a binary classifier, but does not bring us closer to finding the document of interest. A system which labels thirty of the documents, including the relevant document, as belonging to A, will only have an accuracy of 0.71 but will have narrowed down the search to thirty documents.

Conclusion

The results in (Kim & Ross, 2006b) and the results in this paper indicate the promise of using many different sets of features to examine characteristics of genre classes. By examining the variation of feature strengths in genre classes, we expect to be able to create a more robust and purposeful system. If documents with different feature strengths are processed for classification by one classifier, the statistical model can be misled by non-distinguishing features. Trained on sufficient data, this is not a problem; the non-distinguishing features will be filtered out as noise. It is, however, very difficult to have *sufficient data*: for example, in (Kim, 2004), the CANDC part-of-speech tagger (Curran & Clark, 2003), reputed to have performed well elsewhere, was employed to tag words in Astronomy research articles, and shown to tag the term *He*, the chemical element Helium, as a pronoun for all instances which propagated further errors on subsequent words. Separating features into smaller groups will minimise the impact of such artefacts, by trying to exclude the noise from the start.

Improvement of each classifier in this document can be envisioned by refining the word list compiled to represent stylometric features, extending the scope of the image features to include many pages of the document, and looking at alternative ways to capture the common word frequency and topology of the image. We can also integrate more features for comparison such as that built on the N-gram of part-of-speech tags (tags which denote whether a word is a verb, noun or preposition) or chunk tags (tags indicating noun phrases, verb phrases or prepositional phrases), and features built on counts or patterns of subjective and objective noun phrases (cf. Riloff, Wiebe, & Wilson, 2003) and latent semantic analysis. Later we would also like to examine contextual classifiers built on source information of the document such as the name of the journal or address of the web page, and anchor texts or domain information (e.g. prior probabilities of different genres being found in the designated source).

The longer term aim, once a genre classifier with performance comparable to an average human labeller has been developed, will be to integrate the method with other tools which extract author, title, date, identifier, keywords, language, summarisations and other compositional properties of files within a single genre, and combine the tool with ingest models developed elsewhere.

Acknowledgements

This research is a part of the Digital Curation Centre's (DCC) (n.d.) research programme. The DCC is supported by a grant from the United Kingdom's Joint Information Systems Committee (JISC) (n.d.) and the e-Science Core Programme of the Engineering and Physical Sciences Research Council (EPSRC) (n.d.), (grant GR/T07374/01) provides the support for the research programme. Additional support comes from the DELOS: Network of Excellence on Digital Libraries (G038-507618) (n.d.) funded under the European Commission's IST 6th Framework Programme. We would also like to thank Volker Heydegger at the Historisch-Kulturwissenschaftliche Informationsverarbeitung (HKI) (n.d.), University of Cologne, for his programming expertise. HKI at the University of Cologne is a participant in the DELOS Digital Preservation Cluster led by the University of Glasgow.

References

Adobe Acrobat PDF specification. (n.d.). Retrieved November 28, 2006 from http://partners.adobe.com/public/developer/pdf/index_reference.html

Aiello, M., Monz, C., Todoran, L., & Worring, M. (2002). Document understanding for a broad class of documents. *International Journal Document Analysis and Recognition, 5*(1) 1–16.

Arens, A., & Blaesius, K. H. (2003). Domain oriented information extraction from the Internet. *SPIE Document Recognition and Retrieval, Vol 5010,* p. 286.

Automatic Metadata Generation. (n.d.). Retrieved November 28, 2006 from http://www.cs.kuleuven.ac.be/hmdb/amg

Bagdanov, A. D., & Worring, M. (2001). Fine-grained document genre classification using first order random graphs. *International Conference on Document Analysis and Recognition, 2001,* p. 79.

Barbu, E., Heroux, P., Adam, S., & Trupin, E. (2005). Clustering document images using a bag of symbols representation. *International Conference on Document Analysis and Recognition*, *2005,* pp. 1216–1220.

Bekkerman, R., McCallum, A., & Huang, G. (2004). *Automatic categorization of email into folders. Benchmark experiments on Enron and SRI corpora* (Tech. Report IR-418). University of Massachusetts: Center for Intelligent Information Retrieval.

Biber, D. (1995). *Dimensions of register variation: A cross-linguistic comparison*. Cambridge University Press.

Boese, E. S. (2005). Stereotyping the web: genre classification of web documents. (Master's thesis, Colorado State University, Computer Science Department, 2005).

Breuel, T. M. (2003). An algorithm for finding maximal whitespace rectangles at arbitrary orientations for document layout analysis. *7th International Conference on Document Analysis and Recognition*, *2003,* pp. 66–70.

Curran, J., & Clark, S. (2003). Investigating GIS and smoothing for maximum entropy taggers. *Proceedings of the 11th Meeting of the European Chapter of the Association for Computational Linguistics (EACL-03)*, pp.91-98. Budapest, Hungary.

Digital Curation Centre. (n.d.). Retrieved November 28, 2006 from http://www.dcc.ac.uk

DC-dot. (n.d.). UKOLN Dublin Core metadata editor. Retrieved November 28, 2006 from http://www.ukoln.ac.uk/metadata/dcdot/

DELOS Network of Excellence on Digital Libraries. (n.d.). Retrieved November 28, 2006 from http://www.delos.info/

DELOS/NSF Working Groups. (2003, June). *Reference models for digital libraries: Actors and roles*. Final report. Retrieved November 28, 2006 from http://delos-noe.iei.pi.cnr.it/activities/internationalforum/Joint-WGs/actors/Actors-Roles.pdf

Dublin Core Initiative. (n.d.). Retrieved November 28, 2006 from http://dublincore.org/tools/#automaticextraction

Electronic Resources Preservation Access Network (ERPANET). (n.d.). Retrieved November 28, 2006 from http://www.erpanet.org

Engineering and Physical Sciences Research Council (EPSRC). (n.d.). Retrieved November 28, 2006 from http://www.epsrc.ac.uk/

ERPANET: Packaged Object Ingest Project. (2003). Retrieved November 28, 2006 from http://www.erpanet.org/events/2003/rome/presentations/ross_rusbridge_pres.pdf

Finn, A., & Kushmerick, N. (2006). Learning to classify documents according to genre. *Journal of American Society for Information Science and Technology, 57*(11), 1506-1518.

Giuffrida, G., Shek, E., & Yang, J. (2000). Knowledge-based Metadata Extraction from PostScript File. *Proceedings of the 5th ACM International Conference on Digital Libraries, 2000,* pp. 77–84.

Han, H., Giles, L., Manavoglu, E., Zha, H., Zhang, Z., & Fox, E. A. (2000). Automatic Document Metadata Extraction using Support Vector Machines. *Proceedings of the 3rd ACM/IEEE-CS Joint Conference on Digital Libraries,* pp. 37–48. New York: ACM Press.

Hedstrom, M., Ross, S., Ashley, K., Christensen-Dalsgaard, B., Duff, W., Gladney, H., Huc, C., Kenney, A. R., Moore, R., & Neuhold, E. (2003). *Invest to Save: Report and recommendations of the NSF-DELOS Working Group on digital archiving and preservation*. (Report of the European Union DELOS and US National Science Foundation Workgroup on Digital Preservation and Archiving). Retrieved November 28, 2006 from http://delos-noe.iei.pi.cnr.it/activities/internationalforum/Joint-WGs/digitalarchiving/Digitalarchiving.pdf

Historisch-Kulturwissenschaftliche Informationsverarbeitung (HKI). (n.d.). Retrieved November 28, 2006 from the University of Cologne (Universität zu Köln) website: http://www.hki.uni-koeln.de/

Joint Information Systems Committee. (n.d.). Retrieved November 28, 2006 from http://www.jisc.ac.uk/

Karlgren, J. & Cutting, D. (1994). Recognizing text genres with simple metric using discriminant analysis. *Proceedings of the 15th Conference on Computational Linguistics, Vol. 2,* 1071–1075.

Ke, S. W., Bowerman, C. & Oakes, M. (2006). PERC: A personal email classifier. In *Proceedings of 28th European Conference on Information Retrieval, ECIR 2006*, 460–463.

Kessler, B., Nunberg, G., & Schuetze, H. (1997). Automatic detection of text genre. In *Proceedings of the 35th Annual Meeting on Association for Computational Linguistics,* 32–38.

Kim, Y. (2004). Anaphora resolution for automatic citation linking. (Master's thesis, MSc. for Speech and Language Processing). University of Edinburgh.

Kim, Y., & Ross, S. (2006a). Automating Metadata Extraction: Genre Classification. In *Proceedings of the UK e-Science All Hands Meeting*, *Nottingham, UK*. ISBN 0-9553988-0-0. Retrieved November 28, 2006 from http://www.allhands.org.uk/2006///proceedings/papers/663.pdf

Kim, Y., & Ross, S. (2006b). Genre classification in automated ingest and appraisal metadata. In J. Gonzalo, (Ed.), *Proceedings European Conference on advanced technology and research in Digital Libraries (ECDL),* in *Lecture Notes in Computer Science*, *Vol. 4172* (pp. 63-74). Berlin, Germany: Springer Verlag.

MetadataExtractor. (n.d.). Retrieved November 28, 2006 from http://pami.uwaterloo.ca/ (follow the link for Text Mining)

National Archives UK: DROID (Digital Object Identification). (n.d.). Retrieved November 28, 2006 from http://www.nationalarchives.gov.uk/aboutapps/pronom/

National Library of Medicine US. (n.d.). Retrieved November 28, 2006 from http://www.nlm.nih.gov/

National Library of New Zealand: Metadata Extraction Tool. (n.d.). Retrieved November 28, 2006 from http://www.natlib.govt.nz/en/whatsnew/4initiatives.html#extraction

Noonberg, D. B. (n.d.). XPDF PDF document viewer. Retrieved November 28, 2006 from http://www.foolabs.com/xpdf/

PDFTOHTML. (n.d.). PDF to HTML converter. Retrieved November 28, 2006 from http://pdftohtml.sourceforge.net/

PREMIS (PREservation Metadata: Implementation Strategy) Working Group. (n.d.). Retrieved November 28, 2006 from http://www.oclc.org/research/projects/pmwg/

Python. (n.d.). Retrieved November 28, 2006 from http://www.python.org

Python Imaging Library. (n.d.). Retrieved November 28, 2006 from http://www.pythonware.com/products/pil/

Rauber, A., & Müller-Kögler, A. (2001). Integrating automatic genre analysis into digital libraries. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries, Roanoke, VA*, 1-10.

Riloff, E., Wiebe, J., & Wilson, T. (2003). Learning subjective nouns using extraction pattern bootstrapping. In *Proceedings of the Seventh Conference on Natural language learning at HLT-NAACL 2003, Vol. 4* (pp.25–32). Morristown, NJ: Association for Computational Linguistics.

Ross, S., & Hedstrom, M. (2005). Preservation research and sustainable digital libraries. *International Journal on Digital Libraries, 5,*(4) 317-324. DOI: 10.1007/s00799-004-0099-3.

Santini, M. (2004a). A shallow approach to syntactic feature extraction for genre classification. In *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).

Sebastiani, F. (2002). Machine learning in automated text categorization. In *ACM Computing Surveys, Vol. 34,* 1-47.

Shafait, F., Keysers, D., & Breuel, T., M. (2006). Performance comparison of six algorithms for page segmentation. *7th IAPR Workshop on Document Analysis Systems (DAS),* 368–379.

Shao, M., & Futrelle, R. (2005). Graphics Recognition in PDF document. *6th IAPR International Workshop on Graphics Recognition (GREC2005)*, 218–227.

Thoma, G. (2001). *Automating the production of bibliographic records*. (R&D report of the Communications Engineering Branch, Lister Hill National Center for Biomedical Communications, National Library of Medicine).

Witte, R., Krestel, R. & Bergler, S. (2005). ERSS 2005:Coreference-based summarization reloaded. In *Proceedings of DUC 2005 Document Understanding Workshop*, Vancouver, B.C., Canada.

Witten, I. H., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques*. (2nd ed.). San Francisco: Morgan Kaufmann.