D R A F T FOREWORD

Imagine a world without documents! No books, magazines, email, laws, and recipes. For better or worse, in our world documents outnumber every other kind of artifact. The Library of Congress contains over twenty million books plus another one hundred million items in the special collections. Google indexes about twenty *billion* web pages (July 2010). Both of these are growing like topsy. There are far more documents than houses, cars, electronic gadgets, socks, and even rubber bands.

Since the dawn of antiquity documents have played an essential role in fostering civilizations. One could make a plausible argument that material and social progress are proportional to document density. The web rivals the importance of the printing press in making documents widely available rather than kept under lock and key in a monastery or the royal library.

The variety of physical embodiments of paper documents (multivolume encyclopedias, pocket books, newspapers, magazines, passports, driver's licenses) is easily matched by the number of file types, generally known by acronyms like *DOC, PDF, HTM, XML, TIF, GIF….* The suffix indicates the type of processing appropriate for each file type. Differences between file types include how they are created, compressed, decompressed, and rendered in a human-readable format. Equally important is the balance between ease of modification of textual content (*linguistic or logical components*) and appearance *(physical* or *layout components*).

Important paper documents like deeds and stock certificates are written or printed in indelible ink on watermarked stock, and then locked into a vault. Digital files stored on personal computers or transmitted over the internet must be similarly protected against malware launched by curious or malicious hackers. Perhaps surprisingly, many of the elaborate and recondite measures used to ensure the authenticity, secrecy, and traceability of digital documents are rooted in Number Theory. This is one of the oldest branches of pure mathematics, with many counterintuitive theorems related to the factorization of integers into prime numbers. The difficulty of factorizing large numbers is of fundamental importance in modern cryptography. Nevertheless, some current systems also incorporate symbol substitutions and shuffles borrowed from ancient ciphers.

Because of its applications to computer security, cryptography has advanced more in the last forty years than in the previous three thousand. Among widely used systems based on the *secret (symmetric) key* and *public (asymmetric) key* paradigms are the Data Encryption Standard (*DES*), Rivest Ciphers (*RCn*), the Rivest, Shamir, Adleman (*RSA*) algorithm, and the Digital Signature Algorithm (*DSA*). *One-way encryption* is often used to encrypt passwords and to generate

a *digital fingerprint* that provides proof that a file has not been altered. Other methods can provide irrefutable evidence of the transmission of a document from one party to another.

There is a piquant contrast between the role of national governments in promoting the use of secure software to facilitate commerce and to grease the wheels of democracy, and its duty to restrict the propagation of secure software in order to protect its military value and to maintain the ability of law enforcement agents to access potentially criminal communications.

While we are not yet ready to declare all written material that exists only on paper as *legacy documents*, that moment cannot be very far. For scanned documents, the transition from document image analysis to content analysis requires optical character recognition (*OCR*). Even born-digital documents may require OCR in the absence of software for reading intermediate file formats. Some documents such as musical scores, maps and engineering drawings are based primarily on long-established and specialized graphic conventions. They make use of application-specific file types and compression methods. Others, like postal envelopes, bank checks and invoice forms, are based on letters and digits but don't contain a succession of sentences.

The value of digital documents transcends ease of storage, transmission and reproduction. Digital representation also offers the potential of the use of computer programs to find documents relevant to a query from a corpus and to answer questions based on facts contained therein. The corpus may contain all the documents accessible on the World Wide Web, in a digital library, or in a domain-specific collection (e.g. of journals and conference reports related to digital document processing). For text-based documents, both information retrieval (*IR*) and query-answer (*QA*) systems require natural language processing (*NLP*).

Procedures range from establishing the relative frequency, morphology and syntactic role of words to determining the *sense*, in a particular context, of words, phrases and sentences. Often simple relationships with arcane names, like *synonymy*, *antinomy*, *hyperonymy*, *hyponymy*, *meronymy* and *holonymy*, are sought between terms and concepts. Fortunately the necessary linguistic resources like lexicons, dictionaries, thesauri, and grammars are readily available in digital form. To avoid having to search through the entire collection for each query, documents in large collections – even the World Wide Web - are indexed according to their putative content. *Metadata* (data about the data, like catalog information) is extracted and stored separately.

Professor Ferilli is a key member of a research team that has made steady progress for upwards of thirty years on the conversion of scanned paper documents to digital formats. He presents a useful overview of the necessary techniques, ranging from the low-level preprocessing functions of binarization and skew correction to complex methods based on first-order logic for document classification and layout analysis. Many of the algorithms that he considers best-in-class have been incorporated, after further tuning, into his group's prototype document analysis systems.

The first chapter highlights the rapid transition of juridical and commercial records from paper to electronic form. The rest of the book traces the contemporaneous evolution of digital document processing from esoteric research into a pervasive technology. Professor Ferilli offers an up-to-date tour of the architecture of common document file types, lists their areas of applicability, and provides detailed explanations of the notions underlying classical compression algorithms.

He expertly guides the reader along the often tortuous paths from the basic mathematical theorems to the resulting security software. Next, he reviews recent legislation with emphasis on relevant Italian, British and European Community laws. OCR is discussed mainly from the perspective of open source developments because little public information is available on commercial products. Handwritten and hand-printed documents are outside the scope of this work. Except for some sections on general image processing techniques and color spaces, the focus of this book is on printed documents comprising mainly natural language. The relevant NLP, IR and QA techniques (based on both statistical methods and formal logic) and the management of large collection of documents are reviewed in the last two chapters.

The study of documents from the perspective of computer science is so enjoyable partly because it provides, as is evident from this volume, many opportunities to bridge culture and technology. Extensive references, especially to early seminal contributions, facilitate further exploration of this fascinating topic. The material presented in the following pages will be most valuable to the many researchers and students who already have a deep understanding of some aspect of document processing. It is such scholars who are likely to feel the need, and to harvest the benefits, of learning more about the growing gamut of techniques necessary to cope with the entire subject of digital document processing.

George Nagy
Professor,
Rensselaer Polytechnic Institute