Symposium on Document Engineering

November 9–10, 2001
Doubletree Hotel Buckhead, Atlanta, Georgia, USA

Document engineering is an emerging discipline within computer science that investigates systems for documents in any form and in all media. Document engineering is concerned with principles, tools and processes that improve our ability to create, manage and maintain documents. An analogy to software engineering both apt and deliberate, but document engineering looks at computer systems for documents in general.

Because documents are so pervasive in modern life, it might seem difficult to put define them meaningfully. A document is a representation of information that is designed to be read or played back by a person. It may be presented on paper, on a screen, or played through a speaker or some other output device and its underlying representation may be in any form and include data from any medium. A document may be stored in final presentation form or it may be generated on-the-fly, undergoing substantial transformations in the process. A document may include extensive hyperlinks and be part of a large web of information. Furthermore, apparently independent documents may be composed, so that a web of information may itself be considered a document.

Document technology is pervasive and the growth of the World Wide Web is only increasing its importance. Web standards such as HTML, XML, XSL, and RDF are the modern embodiments of prior document research, much of which was presented at the ancestor meetings of this symposium.

The ACM Symposium on Document Engineering was organized by a international steering committee of document researchers with the strong support of the ACM CIKM conference organization. This first Symposium was organized as a workshop at CIKM 2001. The symposium's goals are to bring document engineering researchers together to present and discuss their research work and to increase the field's visibility within computer science. The steering committee plans for the Symposium to run on an annual basis.

The Symposium had 35 participants from seven countries. 55 full papers had been submitted, of which 18 were accepted. Unfortunately, the Symposium was held a short time after the events of September 11, 2001 and several authors were not able to attend. The thirteen presentations covered a wide variety of document engineering topics and many valuable discussions resulted. The proceedings have been published by ACM Press and are available in the ACM Digital Library.

The Symposium's keynote speaker was Rob Akcsyn, who is well known for his work on hypermedia document systems in both academic and industrial settings and

has also been active in many ACM organizational activities. Akcsyn discussed the many lessons that he has learned in the process of developing these systems and his views of how the document engineering community should move forward. Some of the important system insights Akcsyn presented were: that good system performance has many unexpected effects; that large scale requires simple designs; and that systems should be built using "uphill representations", where the representations are more general and abstract than the final product. His recommendations for the field of document engineering were numerous. He noted that document engineering is basically the intersection of engineering and documents and he recommended that we should be liberal in our notion of documents, but narrow in our notion of engineering, particularly avoiding issues of management or pure science, focusing instead on the design tradeoffs in our systems. He also recommended that we should identify grand challenges for the field.

Management of document structure is a central topic for document engineering. Pietriga described VXT, a new system for defining XML transformation visually. A key feature of the VXT system is that the visual elements correspond directly to elements of a textual transformation language, Circus. Pietriga's presentation included a convincing demonstration of the system. Tozawa described a method for type-checking XSLT style sheets. In particular, he showed how to determine whether a style sheet can be proven to transform of a document of type A into a type B in XSLT0, a subset of XSLT. He used tree automata to model the DTDs, because they are more expressive than DTDs. Villard showed how XSLT transformations could be used to create adaptable multimedia presentations in a multiple-view system based on the Kaomi multimedia toolkit. Villard's presentation included a short demonstration of the system and showed the use of direct manipulation to edit XSLT transformation sheets.

Several presenters were interested in issues for hypermedia document systems. Na described context-aware Trellis (caT), which enhances the Trellis Petri-net-based hypermedia model with colored tokens and a hierarchical structure. With a clear example, they showed how caT can be used to create flexible, adaptive hypermedia systems. The caT system can also be used to create Petri nets for software specification. Muchaluat-Saade discussed the relationships between architectural definition languages (ADLs) and hypermedia authoring languages, which have a number of naturally analogous characteristics. She described the many differences and presented a structural metamodel that subsumed both classes of languages. Phelps described the Multivalent Browser, a Web browser with a novel architecture, designed for use in research. He argued that the plug-in architectures of existing Web browsers are inadequate for researchers because they only allow a limited set of browser extensions. For example, it is not possible for a researcher to replace the document parser or formatter in commercial browsers. The Multivalent Browser is structured to allow any part of the system to be replaced, using a model based on behaviors, hubs and a variety of internal protocols.

Another important area is document analysis. McKechnie described a system for helping human analysts correctly classify civil engineering documents. This system exploits several interesting heuristics to provide a good user interface while suggesting likely categorizations of documents. For instance, it lists first those documents for which the automated analysis could only find a poor match, so that the human analysts effort could be focused on likely problem areas. Pimentel described a system that used Latent Semantic Indexing to automatically deduce suitable hypermedia links.

Based on results from using their software with a repository of multimedia documents in an electronic classroom repository, they showed that the quality of the automatically generated links was quite high and were able to discover sensible links that human analysts had not identified. Duong described a technique for extracting text from images of printed documents. The technique focuses on identifying the regions of a page that contain text, rather than on recognizing the characters themselves. Their method involved several steps beginning with the identification of zones of interest, followed by binarization and extraction of the text zones.

Other presentations showed the diversity of document engineering topics. Document researchers have long been interested in constraint systems for specifying layout and appearance. Zhou presented CGLIB, a constraint-based graphics library implemented in Prolog. CGLIB is intended as the starting point for a higher-level graphics library and includes constraints for grid, table and tree layout of objects. Silberhorn's TabulaMagica is a system for managing complex tables displaying information described by as many as eight categories. TabulaMagica has separate models for structure and presentation that are coupled by object sharing, references, and constraints. Both models are DAGS with constraints being used to maintain the validity of their organization. Salminen discussed requirements for XML document databases. They found many areas in which current database and SGML systems are not adequate including the lack of a definition of document equivalence and the need for mechanisms to manage schemas as well as documents. Sampaio described an environment for SMIL 2.0 documents that uses a formal design methodology to ensure the consistency of synchronization constraints. The key difficulty faced by the system is a state space explosion in its logic system due to parallel scheduling constraints. The solution involves defining special states to represent common abstractions of parallel scheduling.

The attendees were enthusiastic about both the quality and the topics of the symposium and they discussed organizational issues in creating a successful symposium series. The Symposium on Document Engineering 2002 will be held in conjunction with CIKM 2002 in McLean, VA on November 8 and 9, 2002. Planning has begun for the 2003 symposium, which will be held in Grenoble, France in the fall of that year.

Ethan V. Munson
Symposium Chair, DocEng 2001