2005－DD－50（3）
2005／5／27

デジタルドキュメント研究 10 年の傾向

斎藤　伸雄†　　三田　虎史†

インターネットの隆盛や，XML の普及，オフィス文書の電子化の促進など，近年，情報処理技術においてデジタルドキュメントを取り巻く環境は急速に変化してきた．当研究会でも，状況の変化に併せて様々な研究報告が行われており，研究報告対象の変化の流れは，デジタルドキュメント環境の変化に同調しているはずである．そこで，研究報告の流れを確認すべく，デジタルドキュメント研究会発足以来これまでの研究報告 252 件の抄録情報に出現してきた言葉について，データマイニングツールを用いて解析を行い，年代ごとに研究報告のトピックを俯瞰したので，それらの傾向について報告する．

Tendency of digital document research 10 years

Nobuo SAITO†　　Takeshi MITA†

The Internet becomes popular, XML spreads, the office document is made electronic, and the environment that surrounds a digital document that relates to information processing technology has changed rapidly in recent years. Various research reports were done changing in the situation.

The change to be researched is sure to be tuned to the change in the digital document environment.

Then, to confirm the flow to be researched, the word included in 252 abstract information that had been reported in the digital document society was analyzed with a data mining tool. It takes a general view of the topic that appears every age, and it reports on those tendencies.

1．はじめに

1996 年に発足した情報処理学会デジタルドキュメント研究会は，2005 年に 10 年目を迎え，これまで(2005.1 月開催の研究会まで)に約 250 件の研究発表が実施されている．この 10 年の間には，オフィス文書を始めとしたドキュメントの電子化が急速に進み，インターネットの隆盛に伴う web ドキュメントの増加や，XML の誕生など，デジタルドキュメント研究の重要な対象となる文書環境に大きな変化が生まれている．そこで，文書環境の変化に伴う過去の研究内容の変遷を捉え，これからのデジタルドキュメントに関する未来予測の足がかりとすべく，この 10 年間の研究発表の内容を振り返り，研究対象となったトピックの抽出を行った．

†凸版印刷株式会社　情報ビジネス開発本部
Info-Communication Business Division,
TOPPAN Printing Co.,Ltd.

2．分析の方法

2．1　分析対象資料

過去の研究報告は，情報処理学会電子図書館にて閲覧が可能である．デジタルドキュメント 10 年の軌跡を振り返るにあたって，当該図書館に掲載されている資料[1]をもとに調査を行った．

<http://fw8.bookpark.ne.jp/cm/ipsj/select_signotes2.asp?category2=DD>(参照 2005-4-27)

情報処理学会電子図書館では，各研究報告資料の書誌情報までを無料で閲覧可能となっており，研究会の登録会員であれば，本文の PDF ファイルも無料で閲覧が可能である．今回の調査では，全研究報告の抄録を含めた書誌情報までを閲覧し，検討を行った．書誌情報には，研究報告の発行年月情報も含まれており，大抵の場合は研究会やシンポジウムに合わせて研究報告が行われるので，研究報告の発行年月において

研究会等が開催されたものと推定した．それによると，最初のデジタルドキュメント研究会が開催されたのは，1996 年 5 月であったようであり，初回の研究報告は，３件であったと思われる．

また，年度毎の抄録情報に出現するキーワードの分析には，テキストマイニングツールを用いた．

2．2　テキストマイニングツール

各研究報告の抄録情報に出現するキーワードの分析には，テキストマイニングツールである TRUE TELLER[2]を用いた．TRUE TELLER は，顧客からのアンケート情報などのテキストデータを分析するマーケティングツールとして活用されるテキストマイニングツールである．分析対象として与えられたテキストデータは，形態素に分解され，出現している単語や，関係(係り受け)によって分析されることにより，テキストデータのグループに出現する単語を出現件数順に並べる「単語ランキング」や，グループ毎の単語や単語間の関係を２次元にプロットして表示する「マッピング」などの解析結果を得ることができる．

3．研究報告の分析

3．1　報告数による分析

1996 年 5 月の研究会を皮切りに，2005 年 1 月の研究会までに 48 回の研究会が開催され，252 件の研究報告が実施されている(表１)．

表１　研究報告件数の状況

年	研究会回数	発表件数	平均報告数	所属の区分		
				企業	学校	共同
1996 年	4 回	12 件	3.0	7	5	0
1997 年	6 回	27 件	4.5	15	11	1
1998 年	5 回	21 件	4.2	16	4	1
1999 年	6 回	35 件	5.8	24	10	1
2000 年	5 回	20 件	4.0	16	4	0
2001 年	5 回	20 件	4.0	14	5	1
2002 年	5 回	36 件	7.2	13	22	1
2003 年	5 回	31 件	6.2	14	14	3
2004 年	6 回	45 件	7.5	17	25	3
2005 年	1 回	5 件	5.0	1	4	0
合計	48 回	252 件	5.3	137 54.4%	104 41.3%	11 4.4%

年毎の研究報告数の変化を見ると，初年である 1966 年には 4 回の研究会で 12 件の報告が行われており，それ以降は毎年 20 件～35 件程度の報告が行われているが，図１や図２に示したように，特徴的な傾向はみられない．しかし，2004 年には過去最高の 45 件の報告が行われており，近年は研究会１回あたりの報告数も増加していることもあり，活発な研究活動が行われつつある感がある．

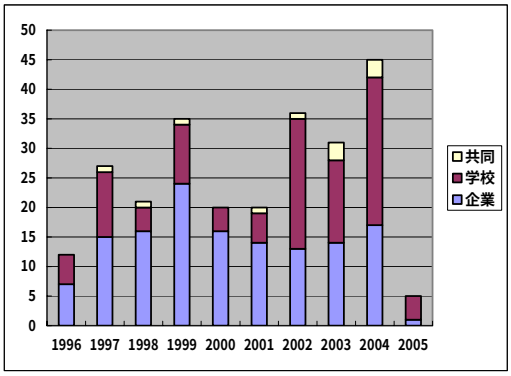

図１　研究報告数の変化

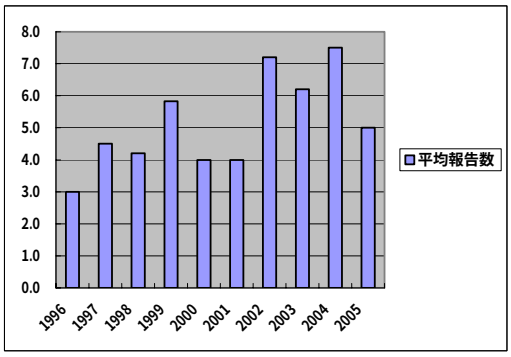

図２　研究会１回あたり報告数の変化

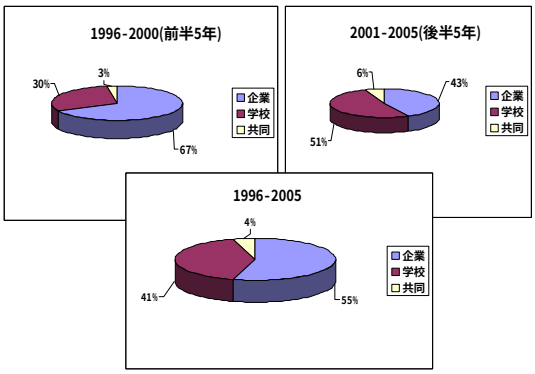

図３　報告者の所属の割合

また，各報告の発表者の所属を，企業，学校，共同の３種類に分類しその割合を見てみると，全体では企業 55%,学校 41%,共同 4%となっており(図３)，企業所属の研究報告がやや多めとなっているが，前半 5 年と後半５年に分けてみると，前半５年では，企業 67%,学校 30%に対して，後半５年では，企業 43%,学校 51%に変化しており，学校からの研究報告が増加傾向にあることがわかる．年に 1 件程度であった産学共同研究の報告も，ここ２年間はそれぞれ３件みられ，産学協同研究が盛んになりつつある．

3．2　コレスポンデンス分析

各年のトピックの類似度，関係の深さを把握するために，テキストマイニングツールを用いてコレスポンデンス分析を行った．集計済みのクロス集計結果を用い，グループに出現している単語の要素を使って，それらの相関関係が最大になるように数量化し，グループの要素を多次元空間(散布図)に表現する．ここでは２次元で表現した(図４)．

第１軸，第２軸ともに共起で自動算出され，白い四角が中心であり，中心に近いほど全体に出現している話題に近く，遠いほど特徴的である．上方に特出しているのが 1999 年，下方が 2002 年，右に特出しているのが 2004 年である．図１のグラフでもわかるように，この３つの年は報告件数が多かったこともあり雑音の多さによって特徴的な結果を示してしまったと思われる．

また，1998 年に XML1.0 が勧告され，1999 年にはそれを契機に XML に関する研究が広まりを見せた事も特徴的な結果の要因となっていると思われる．2002 年には，セマンティック web，オントロジ，メタデータといったテーマが急増し，2004 年には web サービス，SOA，モバイルといったキーワードが盛んに使われたことも一因であろう．

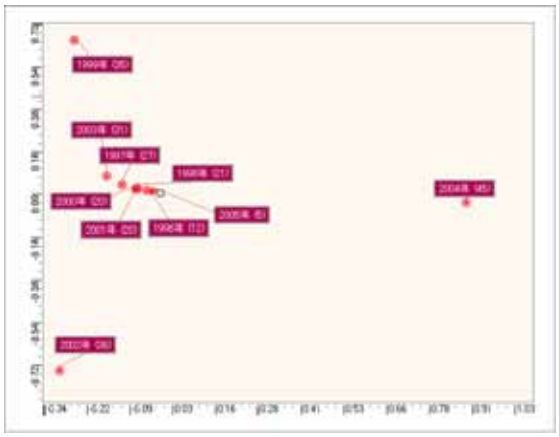

図４　年毎のコレスポンデンス分析結果

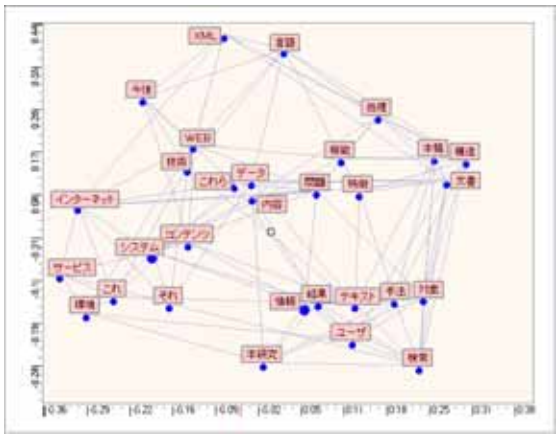

図５　全件単語の主成分分析

3．3　主成分分析

共起による第１軸，第２軸を用いて主成分分析を行った．名詞の上位 30 個をプロットし，100 件以上出現する語は，大きい青い丸，それ以外が小さい丸で表現されている(図５)．上位 30 語の中には，出現数 10 件以下の語は含まれなかった．白い四角が中心である．

右下のクラスタには，「検索」を中心に，「テキスト」を「対象」とした「手法」が提案され，その「結果」が示される研究分野が存在していることがわかる．左下のクラスタでは，「システム」は「インターネット」上で「サービス」として展開され，そこに載せる「コンテンツ」に対しても議論されている．また，上のクラスタでは，「XML」という「言語」が，「Web」の「技術」を支えており，「今後」もキーとなる．

また，これらは左下のクラスタともつながっていることから，インターネット上でのサービスを基礎に「XML」技術が存在していることがわかる．さらに，右上のクラスタでは，「文書」「構造」を対象とした，「特徴」的な「機能」を備える「処理」が研究されていることがわかる．ここも左上のクラスタともつながっており，「処理」には「XML」「言語」が強く関係していることがわかる．

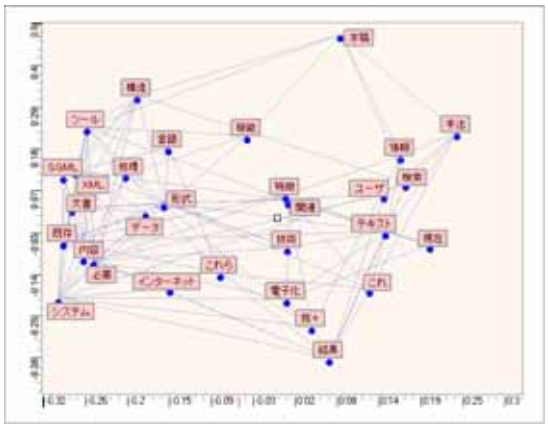

図６　前半５年の主成分分析

図７　後半５年の主成分分析

また，前半５年(図６)と後半５年(図７)に分けて主成分分析を行った．前半５年では，「テキスト」,「電子化」された「情報」「検索」の「手法」や，「SGML」,「ＸＭＬ」形式の「構造」「文書」や「データ」の「処理」が議論され，「インターネット」や「システム」への適用が研究されている．一方，後半５年では，「情報」「検索」の「問題」解決が「目的」の一つとなり，「対象」は「特徴的」な「文書」「構造」の処理であり，「XML」言語であることがわかる．また，「XML」は「Web」「技術」とも関連が深く，「データ」「管理」といった言葉との関連も強い．さらに，「XML」は，「携帯」電話や「インターネット」,「システム」,「サービス」,「コンテンツ」との関連を強めている．

4. トピックの変遷

4.1 1996 年

デジタルドキュメント研究会が発足して最初の研究会では，JAVA に関する研究報告が 2 件行われている．抄録の中では「今後，新たな用途が広がって来ると考えられる」[3]などと記述されているように，www の普及が促進されつつある状況の中で JAVA が注目されていた．そのほか，SGML に関する話題が多く報告されており，主に，SGML を利用した情報システムの構築法に関する研究報告や，既存文書の SGML 化に関する研究報告，DTD が変更になった場合のデータ移行に関する研究報告などが行われ，SGML の普及期であったことが伺える．

また，SGML という言葉と同時に，CALS/EDI といった言葉も使われており，文書処理システムの適用先として CALS が注目されていた．

4.2 1997年

依然として，SGML/CALS といった言葉が出現しているが，この年の研究報告の中で頻出するのは「電子マニュアル」という言葉である．マニュアルの電子化が促進されつつある状況の中，電子マニュアルの見易さに関する研究や，マニュアル電子化の事例紹介といった報告が行われている．他には，SGML データの入力や，データ変換に関する研究報告が実施されており，ドキュメントの電子化にあたっては，電子化文書の生成に課題があった年代であったこと

がわかる．また，「検索」に関する研究報告も数件実施されている．大量の文書が電子化された場合に，ある語句が含まれる箇所を高速に探し出すための検索に関する研究のほか，ハイパーテキストを実現するために，相互参照箇所を検索する研究，最適な検索タームを抽出するための研究，コールセンター等での活用に注目した類似事例文書の検索などである．

4.3 1998 年

この年の最大のトピックは，W3C によって XML Ver.1.0 が 2 月に勧告となったことである．これを受けて 3 月の研究会では早速 XML という言葉を含んだ報告が実施されており，後半に向けて XML に関する話題が増加している．しかし，SGML に関する研究も継続して報告されており，SGML/XML エディタの開発に関する報告や, SGML/HTML 変換に関する研究の報告が行われている．また「www」,「 インターネット」に関する話題も活発になり，HTML 形式のデジタルドキュメントの生成, ハイパーリンクの実現に関する研究が活発になっている事がわかる．

翌年にも 1 件の報告が実施されているが, この年には文字コード・外字に関する研究報告が 2 件実施されている．ドキュメントの電子化が進んだ結果，電子化が困難な文字の存在が課題となり，文字コードや外字に関する研究も行われるようになってきていたのだと推察できる．

4.4 1999 年

引き続き SGML・XML・HTML といったデジタルドキュメントの生成や処理に関する研究報告が多く行われているが，この年には，いくつかの新たなトピックが生まれている．

ひとつは，ドキュメントのセキュリティに関する問題である．この頃，行政文書の電子化が推進されつつあり，情報公開の必要性と，プライベート情報の機密性といった矛盾した情報操作を実現するための研究が始まっていた．また，情報の組織化や，情報を中心とした人間のコミュニケーションを論じる研究なども見られ，情報と社会の関係に関する研究も行われている．ドキュメントのデジタル化が社会に与える影響が危惧され始めてきた．

また，カーナビ・PDA といった情報端末向けのコンテンツとしてのデジタルドキュメントに関する研究もはじまり，地理情報システムへの研究や，XML を活用した，3 次元グラフィックスデータとのドキュメント統合に関する研究の報告が行われ，これまでは，電子マニュアルや，www への適用について論じられる事が多かったデジタルドキュメントの対象に，広がりが見られるようになってきている．

4.5 2000 年

さらに，デジタルドキュメントの対象が広がり，この年には携帯電話やモバイル機器への応用について論じられ始めている．www への応用や, 地理情報システムへの応用，デジタル署名に関する話題も依然として議論されており，研究会での研究報告の対象は，非常に広範囲になってきている．

応用事例についてもより具体的なものとなっており，行政の電子化はもとより，投資情報への応用や，アンケートシステムへの応用が報告されている．さらに，EC/EDI への XML 活用の研究報告と，web サイトを活用したマーケティングの研究報告では,「E ビジネス」というキーワードが使われ始めている．ビジネスを支える要素としてデジタルドキュメントの役割が大きくなりつつある．

4.6 2001 年

21 世紀に入り，研究報告の中に多くに XML という言葉が多く見られるようになってきている．2001 年の研究報告 20 件のうち，16 件の抄録情報に XML,もしくは XHTML 等のキーワードが含まれている．

その多くが応用に関する研究であり．ひとつは設計支援システムであり，ひとつは行政電子化であり，携帯端末への適用や，住所録への適用などが報告されている．

XML 文書検索や，データ変換に関する研究も依然として実施され続けており，広範囲のドキュメントが電子化される事によって，効率の良い検索を実現するための研究テーマや，ドキュメント生成に関わる研究テーマは尽きる事が無い．

4.7 2002 年

この年のトピックは，2001 年の頻出単語ランクの 1 位は「XML」であったにも関わらず(図８)，2002 年には，「XML」は 10 位以内に含まれていない(図９)点である．

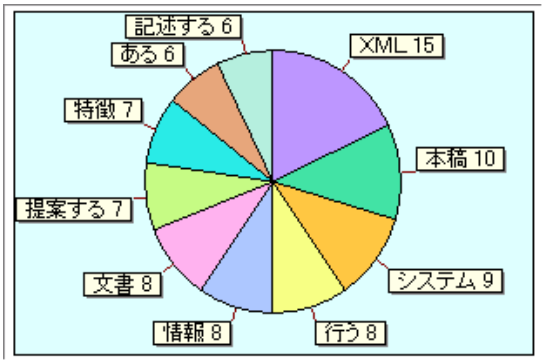

図８　2001 年の頻出単語

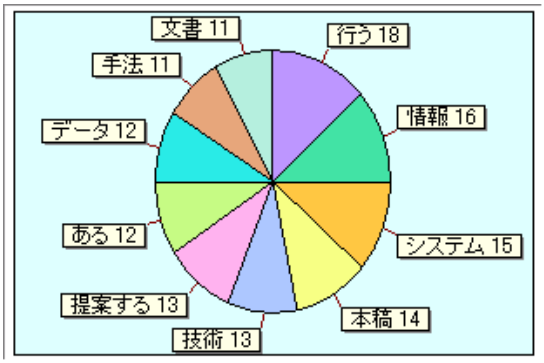

図９　2002 年の頻出単語

ところが，研究報告の内容を見ると，後半では，Xlink，XSL に関する報告があり，XML データベースについての報告も行われていることから，XML に関する研究が実施されなくなった訳ではない．もはや XML は当たり前の技術となり，あえてその言葉を使う必要は無くなってきたようである．

2001 年 11 月にその技術動向を紹介する研究報告が実施された「セマンティック web」に関して，2002 年には，課題や可能性が論じられている．トピックマップやオントロジに関する研究報告も実施され，デジタルドキュメント研究も新たな概念へ広がりはじめた．

また，言語理解・知識表現・自然言語理解に関する研究報告や，コンテンツ管理，文書のクラスタリングといった，文書の内容に深く踏み込んだ研究も活発となり，ドキュメントのデジタル化がすすみ，さらに深い文書の利活用が研究されるようになってきたと考えられる．

4.8 2003 年

この頃より，新しいトピックとして「メタデータ」というキーワードが目立つようになってきた．この背景には，次世代 web として提唱されたセマンティック web 概念の浸透が進んできた事がある一方，これまでは大量文書の効率的な活用のために全文検索を中心とした研究が行われてきたが，メタデータを活用して大量文書の効率的な管理を実現しようとする動きが活発になってきたという背景もある．

2002 年にも話題になっていた，教育への応用に関して，この年には４件の研究報告が行われており，教材のデジタル化，データベース化，教育現場でのインターネット活用等の研究が活発になっていることが伺える．

また，無線ＬＡＮスポットが増え，携帯電話の普及にあわせ，モバイル環境でのデジタルコンテンツ利用に関しても研究報告が行われている．

コンテンツ管理に関する研究や，情報視覚化手法の研究報告も複数行われ，新たな領域への広がりが見られる．

4.9 2004 年以降

すでに前年に兆候が見られた様々な分野への研究対象の広がりは，より顕著となる．

デジタル化されたドキュメントの内容にまで深く踏み込み，文書の要約についての研究報告が行われ，知識管理・コンテンツ管理に関する研究報告も行われている．

情報技術の福祉や医療への活用についても報告されており，点字楽譜に関する研究報告や，テキスト読み上げツールに関する研究報告，電子カルテに関する研究報告も行われている．

従来どおり検索や文書生成に関する研究報告も行われているが，より構造化文書に特化した形の研究へと発展している．また，ペン入力に関する研究報告など，文書システムとユーザのインタラクションに関する研究報告も見られるようになってきている．

4.10 キーワードの傾向

ここまで，各年毎のトピックに関して概観してきたが，前半 5 年と後半 5 年に分けて単語の出現頻度を比較してみると(図 10,11)，「文書」という言葉は，前半・後半とも同数にも関わらず，前半の 2 位から後半は 10 位と後退しており，後半では代わりに「システム」「XML」「データ」といった言葉が上位になる．

前半 5 年は，文書のデジタル化や構造化など，文書自体に関する研究が盛んに実施されたのに対して，後半 5 年は，文書自体を研究対象とするよりも，XML web サービスや，SOA に代表される，XML システム処理といった，デジタルドキュメントの応用や，サービスに関する話題が多くなってきていることがわかる．

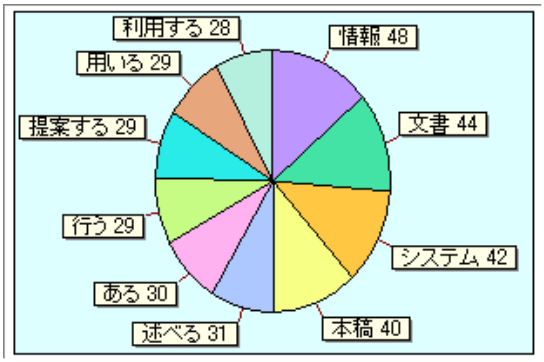

図 10 前半 5 年の頻出単語

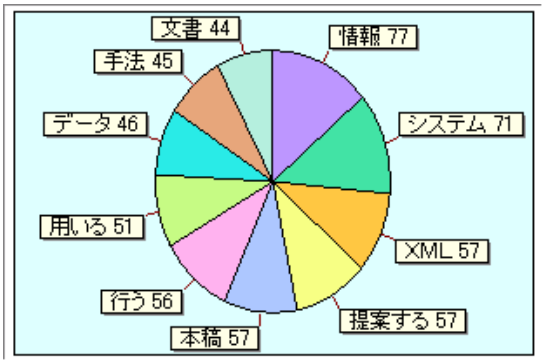

図 11 後半 5 年の頻出単語

5. まとめ

10 年目を迎えたデジタルドキュメント研究会での，これまでの研究報告を概観し，その傾向を分析した．

当初は企業による研究報告が多かったが，近年では大学の研究室や，産学協同の研究報告の割合が増加傾向にある．

この 10 年間の研究報告の傾向を俯瞰すると，一貫して SGML/XML というキーワードがトピックの中心に流れており，1998 年の XML 勧告以降は，XML が中心となっている．しかし，XML の周辺で研究されている内容が，時代の変遷とともに変化してきていることがわかった．前半では，構造化文書の入力・生成，編集，または既存データからの変換など，対象文書自体に関する研究が盛んに行われていたのに対し，後半では構造化文書の処理，応用システム，メタデータの付与など，構造化文書をとり

まく仕組みに関する研究が行われるようになってきていることがわかった．

また，インターネットの爆発的な普及などインフラの変化に合わせ，www やインターネットへの適用に関する研究報告も増加傾向にあり，さらには，携帯電話やＰＤＡといったモバイル端末の普及に合わせて，モバイル端末を利用したシステムの研究や，実験の報告が行われてきている．研究対象も大きく広がりを見せ，福祉サービスや，医療，防災，経済情報への応用の研究報告も見られる．いずれにしても，ドキュメントの電子化は当たり前の技術として浸透し，電子化されたドキュメントの有効活用に関する研究が盛んになりつつある．

今回の報告では，252 件の研究報告の抄録情報に現れる単語を基にデジタルドキュメント研究の流れを俯瞰した．しかし，抄録情報に含まれる単語の集合だけでは研究報告の内容を正しく把握できていない可能性も残される．10 年のデジタルドキュメント研究の流れをより深く把握するために，さらに，研究報告本文に出現する文章や語句についても解析を行う必要があろう．

参考文献

[1]“情報処理学会電子図書館”．情報処理学会．
<http://www.ipsj.or.jp/05system/digital_library/index.html>, (参照 2005-4-27)
[2] “TRUE TELLER”．野村総合研究所．
<http://www.trueteller.net/>, (参照 2005-5-2)
[3]三津濱元一：Java を利用したアプリケーションについて，情処研報 DD-01, pp.1-5(1996)