



Metody matematyczno-ekonomiczne oraz informatyka w biznesie  
Studia podyplomowe

Blok I6  
Nowoczesne postaci dokumentów - tworzenie  
i wymiana dokumentów komputerowych

semestr letni 2006/2007

Szymon Ziolo  
szioło@mimuw.edu.pl

# Dzień dobry!

- O mnie:
  - absolwent Wydziału Matematyki, Informatyki i Mechaniki UW,
  - starszy analityk w firmie Bull Polska:
    - kierowanie pracami analitycznymi w projektach,
    - kursy, szkolenia,
    - merytoryczne wsparcie sprzedaży,
    - zainteresowania: systemy obiegu dokumentów i zarządzania treścią;
  - autor wykładu „XML i nowoczesne techniki zarządzania treścią” na Wydziale Matematyki, Informatyki i Mechaniki UW,
  - inicjator powstania grupy newsowej pl.comp.xml,
  - redaktor prowadzący wydań 6'2001, 6'2003 i 6'2004 czasopisma Software 2.0 poświęconych XML-owi,
  - autor kursów komercyjnych „Podstawy XML-a” i „Modelowanie informacji w XML-u”.
- A kim są Państwo?

# Program bloku I6

- Wykłady:
  - systemy zarządzania dokumentami,
  - poczta elektroniczna, załączniki, typy MIME, szyfrowanie, podpis elektroniczny,
  - HTML, CSS,
  - TeX,
  - XML i standardy pokrewne.
- Dodatkowo na pracowni:
  - eleganckie formatowanie dokumentów Ms Word,
  - kodowanie znaków narodowych,
  - formaty graficzne i kompresja,
  - XML Schema,
  - XSLT.

## Sprawy organizacyjne

- Strona internetowa wykładu:
  - <http://www.mimuw.edu.pl/~sziolo/podyplomowe>
- Zaliczenie przedmiotu:
  - laboratorium:
    - 1 punkt za każdą godzinę – maksymalnie 20 punktów,
    - kryteria: aktywność, wykonanie zadań;
  - wykład:
    - egzamin złożony z 20 pytań testowych wielokrotnego wyboru,
    - 1 punkt za każde pytanie – maksymalnie 20 pytań;
  - aby zaliczyć, trzeba zdobyć łącznie co najmniej 25 punktów.

# Dokumenty

## Statystyka

• **90%** zasobów informacyjnych firm  
jest przechowywanych w dokumentach  
a nie w bazach danych (Deloitte & Touche)

• **92 miliardy** dokumentów  
tworzonych co roku (AIM)

# Dokument

- Słownik Języka Polskiego PWN:
  - pismo urzędowe,
  - materiał w postaci tekstu, fotografii lub jakiegokolwiek przedmiot, mający wartość dowodową lub informacyjną,
  - plik komputerowy zawierający informacje zapisane w odpowiednim formacie.
- Wikipedia:
  - rzeczowe świadectwo jakiegoś fenomenu sporządzone we właściwej dla danego czasu i miejsca formie,
  - w bibliotekoznawstwie: utrwalony z przeznaczeniem do rozpowszechniania wyraz myśli ludzkiej.

Pojęcie *dokument* ma bardzo szerokie znaczenie, nie ograniczające się jedynie do materiałów pisanych. W życiu codziennym praktycznie na każdym kroku mamy do czynienia z dokumentami.

# Klasyfikacja dokumentów

- Wg sposobu sporządzenia:
  - wizualne:
    - piśmiennicze (np. rękopisy, druki, książki, czasopisma),
    - niepiśmiennicze (np. fotografie, plany, mapy, dzieła sztuki);
  - audialne (np. taśmy i płyty dźwiękowe);
  - audiowizualne (np. filmy).
- Wg wzajemnych odniesień:
  - pierwotne (w formie jaką nadał im twórca);
  - wtórne, dokładnie odwzorowane (duplikat, odpis, kopia);
  - pochodne, zawierające informacje o dokumencie pierwotnym i jego zawartości.

Źródło: Wikipedia, <http://pl.wikipedia.org/wiki/Dokument>

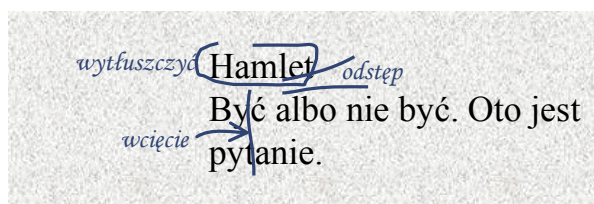
W naszych dalszych rozważaniach będziemy się skupiać przede wszystkim na dokumentach piśmienniczych.

## Znakowanie tekstu

Markup:

*the process of marking manuscript copy for typesetting with directions for use of type fonts and sizes, spacing, indentation, etc.*

The Chicago Manual Of Style



Dla dokumentów komputerowych kluczowym pojęciem jest znakowanie tekstu, zwane także adjustacją techniczną. Angielski termin *markup*, utrwalony w tradycji amerykańskiego edytorstwa, oznacza nanoszenie na rękopisie wskazówek odnośnie typów i rozmiarów czcionek, odstępów, wcięć, itp. Wskazówki takie pozwalają na przygotowanie dokumentu do druku.

# Znakowanie tekstu w epoce komputerów

Treść

Hamlet Być albo nie być. Oto jest pytanie

+

Formatowanie, adjustacja

{nowy\_wiersz} {bold} {wyłącz\_bold} {wcięcie}

=

Dokument

**Hamlet**

Być albo nie być. Oto jest pytanie.



2006-10-28

Dokumenty

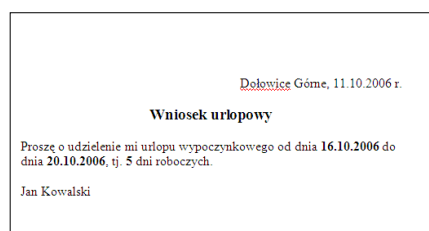
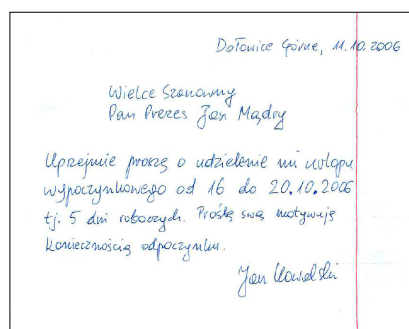
10

Z bardzo podobnym znakowaniem tekstu mamy do czynienia w dokumentach komputerowych. Tekst oznakowany zawiera – oprócz samej treści – także pewne znaczniki opisujące formatowanie tekstu. Tak oznakowany tekst może być przetworzony np. przez system składu, który wyprodukuje czytelny dla człowieka, sformatowany dokument.

## Baza danych vs. dokument

### URLOPY

Nazwisko	Od	Do	Status
Kowalski Jan	2006-10-16	2006-10-20	Do akceptacji
Zawadzki Zenon	2006-08-10	2006-08-20	Zaakceptowany
Nowak Janina	2006-07-02	2006-07-15	Odrzucony



2006-10-28

Dokumenty

11

W systemach komputerowych bardzo istotne jest rozróżnienie między danymi bazodanowymi a dokumentami. Dane bazodanowe mają regularną, tabelaryczną postać ze ściśle określonymi kolumnami. Dzięki temu można je łatwo przetwarzać automatycznie. Większość współczesnych systemów informatycznych opiera się na bazach danych.

Z drugiej strony dokumenty zawierają treść zapisaną w sposób swobodny. Treść taka jest oczywiście czytelna dla człowieka, ale trudna do przetworzenia dla systemu komputerowego. Jednak to właśnie z dokumentami mamy najczęściej do czynienia. Coraz częściej są to dokumenty w formie elektronicznej.

## Dokument $\neq$ plik

- Plik – ciąg danych zapisany w systemie plików, stanowiący całość dla systemu operacyjnego.
- Możliwe fizyczne reprezentacje dokumentu:
  - kartka papieru,
  - plik,
  - rekord w bazie danych,
  - strona internetowa wyświetlona w przeglądarce internetowej,
  - wiadomość SMS,
  - zapis na taśmie VHS,
  - ...

W systemach komputerowych posługujemy się często pojęciem *pliku* jako jednostki informacji. System plików jest jednym z najważniejszych składników systemu operacyjnego każdego komputera.

Nie powinniśmy jednak utożsamiać dokumentu z plikiem. Najczęściej dokumenty zapisujemy w plikach na dysku, ale w ogólności nie jest to jedyna możliwość. Dokumentem jest też przecież też strona internetowa wyświetlana w przeglądarce. Nowoczesne systemy zarządzania dokumentami najczęściej przechowują dokumenty w bazie danych. Z drugiej strony nie każdy plik jest dokumentem – wiele plików na dysku ma znaczenie czysto techniczne (np. programy).

## Formaty dokumentów wizualnych

- Formaty tekstowe:
  - czysty tekst (ang. *plain text*),
  - Hypertext Markup Language (HTML),
  - Standard Generalized Markup Language (SGML), Extensible Markup Language (XML),
  - Rich Text Format (RTF),
  - TeX, LaTeX.
- Formaty binarne:
  - Portable Document Format (PDF),
  - Postscript (PS),
  - Microsoft Office (doc, xls, ppt),
  - OASIS Open Document Format for Office Applications (OpenDocument).

Formaty tekstowe są opracowane na bazie czystego tekstu. Dlatego dokumenty w tych formatach można w ostateczności obejrzeć w najprostszym zwykłym edytorze tekstu. Natomiast w formatach binarnych zawartość dokumentu jest zakodowana w taki sposób, że można ją zobaczyć tylko przy pomocy oprogramowania wspierającego dany format. W zwykłym edytorze tekstowym widzimy bowiem „krzaczkę”.

Czysty tekst jest najprostszym z możliwych formatów. Dokument zapisany czystym tekstem nie zawiera formatowania (no może z wyjątkiem znaku łamania wiersza). Dzięki temu można go obejrzeć i edytować praktycznie w każdym systemie operacyjnym, bez konieczności posiadania specjalistycznego oprogramowania. Jednak już tu pojawiają się problemy – w różnych systemach operacyjnych inaczej koduje się znak łamania wiersza, nie mówiąc już o znakach narodowych...

HTML jest formatem używanym w Internecie do zapisywania treści stron internetowych. Programami, które potrafią wyświetlać dokumenty HTML są oczywiście przeglądarki internetowe.

SGML i XML są bardzo ciekawymi formatami dokumentów strukturalnych. Będzie im poświęcona spora część naszych zajęć.

RTF jest opracowanym przez Microsoft popularnym formatem zapisu dokumentów wraz z bogatym formatowaniem tekstu, możliwością dołączania grafiki, itp. Dokumenty RTF można tworzyć, otwierać i edytować w wielu popularnych edytorach tekstu (np. Ms Word, Open Office Writer).

TeX jest komputerowym systemem profesjonalnego składu tekstu, obejmującym język formatowania tekstu, jak i narzędzia do składu i przetwarzania dokumentów. TeX (i jego następca – LaTeX) jest popularny w środowisku naukowym. Dokumenty złożone przy użyciu TeX-a wyglądają dużo bardziej profesjonalnie niż wydruki z programu Word.

PDF jest opracowanym przez firmę Adobe Systems formatem przeznaczonym do prezentacji i wymiany dokumentów tekstowych z bogatym formatowaniem i elementami graficznymi. Pełna specyfikacja formatu PDF jest dostępna bezpłatnie. Program do przeglądania dokumentów PDF (Acrobat Reader) także jest dostępny za darmo, płatne jest jedynie oprogramowanie firmy Adobe służące do tworzenia i modyfikowania dokumentów PDF. Na szczęście istnieje sporo darmowych programów pozwalających na generowanie dokumentów PDF. Często działają one jako sterowniki wirtualnych drukarek. Zaletą formatu PDF jest to, że dokument w nim zapisany wygląda jednakowo na każdej platformie sprzętowej i w każdym systemie operacyjnym. W dodatku przeciętny czytelnik (nie posiadający specjalistycznego oprogramowania firmy Adobe) go nie zmodyfikuje. Dlatego format ten jest popularny także w Internecie.

Postscript to język opisu strony opracowany przez firmę Adobe Systems, używany głównie w zastosowaniach poligraficznych.

Microsoft Office to pakiet oprogramowania biurowego oraz związana z nim rodzina formatów dokumentów biurowych (dokumentów tekstowych, arkuszy kalkulacyjnych, prezentacji) bardzo popularna w zastosowaniach biurowych i domowych – mimo że oprogramowanie jest płatne. Mimo że Microsoft nie publikuje specyfikacji formatów dokumentów pakietu Office, zostały one całkiem niezłe rozgryzione (szczególnie format Worda) i dzięki temu sporo darmowych programów potrafi lepiej lub gorzej czytać dokumenty pakietu Office.

OpenDocument to standard zapisu dokumentów biurowych, stanowiący alternatywę dla zamkniętych formatów takich jak np. Ms Word. OpenDocument jest rozwijany przez niezależną organizację OASIS (*Organization for the Advancement of Structured Information Standards*). Dokumenty w formatach OpenDocument można tworzyć i edytować w darmowych pakietach biurowych, takich jak np. OpenOffice, StarOffice. OpenDocument jest standardem ISO (ISO/IEC 26300). Jest także wymieniony w rozporządzeniu Rady Ministrów z dnia 11 października 2005 r. w sprawie minimalnych wymagań dla systemów teleinformatycznych jako jeden z podstawowych formatów dla administracji publicznej.

## Formaty dokumentów graficznych

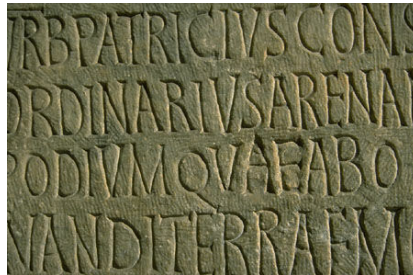
- Grafika rastrowa:
  - JPEG (Joint Photographic Experts Group),
  - Tagged Image File Format (TIFF);
  - Graphics Interchange Format (GIF);
  - Windows Bitmap (BMP).
- Grafika wektorowa:
  - Scalable Vector Graphics (SVG),
  - Adobe Flash,
  - Encapsulated Postscript (EPS).

Grafika rastrowa jest zapisywana w postaci dużej liczby punkcików – tzw. pikseli. Formaty rastrowe są znakomite do przechowywania zdjęć. Ze względu na duże rozmiary grafik rastrowych, w formatach graficznych najczęściej stosuje się kompresję bezstratną lub stratną (tzn. mogącą powodować pogorszenie jakości dokumentu).

Natomiast formaty wektorowe traktują rysunek jako zbiór figur geometrycznych – dokument w formacie wektorowym zawiera opis tych figur. Formaty wektorowe stosuje się najczęściej w odniesieniu do rysunków technicznych, diagramów, itp.

# Technologia komputerowa

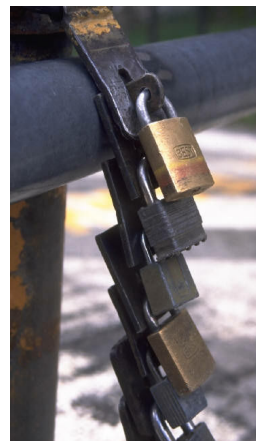
- Miejsce:
  - rylca,
  - dłuta,
  - pióra,
  - maszyny do pisaniazajął komputer.
- Ale czy w istocie wiele się zmieniło?
  - kalka → *copy*
  - gumka → *back-space*



Mimo ogromnych możliwości, jakie daje technologia komputerowa, najczęściej używamy komputera jako nieco lepszej maszyny do pisania.

## Dostępność

- Większość z istniejących dokumentów, mimo iż dostępnych elektronicznie, jest nadal nieużyteczna.
- Powody:
  - zamknięte i niekompatybilne formaty (n.p. DTP, Ms Word),
  - niedostępne/nieznane miejsce przechowywania,
  - rozproszone, rozbieżne i nieaktualne wersje,
  - nieużywane już aplikacje.



Problem ten nie jest krytyczny np. w przypadku małego biura prowadzącego korespondencję, zapisującego dokumenty w formacie Ms Word w systemie plików, ponieważ właściwe dokumenty można odnaleźć korzystając po prostu z wiedzy pracowników.

Jednak w instytucjach operujących dużymi zbiorami informacji o skomplikowanej strukturze, gdzie nad każdym dokumentem pracuje wiele osób (np. w wydawnictwach), problem dostępności staje się realny.

## Programy i ich formaty

- Prawie każda aplikacja wprowadza swój wewnętrzny format.
- Nowe wersje tej samej aplikacji wprowadzają zmiany do używanego formatu:
  - wsteczna kompatybilność,
  - brak możliwości zapisu do formatu poprzednich wersji.
- Aplikacje dostarczają konwerterów:
  - tylko do najpopularniejszych formatów,
  - możliwość utraty danych podczas konwersji.



2006-10-28

Dokumenty

18

Jakieś 12 lat temu w świecie pecetów królował jeszcze DOS, poprzednik systemu Windows. Komputery w polskich biurach służyły za maszyny do pisania, arkusze kalkulacyjne dopiero zaczynały zdobywać użytkowników. W wielu firmach nie było sieci wewnętrznej (o Internecie też nie było co marzyć), więc jeśli pani Krysia z kadr chciała wydrukować swój dokument, nagrywała go na dyskietkę i niosła do księgowości. Tam pan Mietek wkładał dyskietkę do stacji, uruchamiał TAG-a, otwierał jej dokument i drukował.

Szef z kolei przychodził do pana Mietka po raporty księgowe. Pan Mietek nagrywał mu na dyskietce odpowiedni arkusz Lotusa 1-2-3, a szef otwierał go u siebie, wprowadzał poprawki i oddawał księgowemu.

Nikt nie myślał o „otwartych formatach plików” — przecież wszyscy mieli TAG-a i Lotusa, prawda? Nikt nie wyobrażał sobie, że kiedyś mogą zniknąć z rynku — o wiele bardziej realne było zagrożenie związane z „pluską roku 2000”.

Paweł Tkaczyk, *Otwarte formaty plików*, <http://paweltkaczyk.midea.pl/25>

## Formaty zamknięte i otwarte

- Format zamknięty:
  - binarny,
  - opracowany o rozwijany przez producenta oprogramowania obsługującego ten format,
  - niedostępna specyfikacja formatu.
- Format otwarty:
  - pełna specyfikacja formatu dostępna publicznie bez opłat,
  - nie związany z konkretną platformą sprzętową, systemem operacyjnym, pakietem biurowym,
  - nie obwarowany żadnymi patentami ani licencjami,
  - uznany przez międzynarodową organizację normalizacyjną,
  - tekstowy.

Aby format plików mógł być nazwany „otwartym standardem” powinien spełniać kilka kryteriów:

- nie powinien być obwarowany żadnymi patentami ani licencjami — dzięki temu nie jesteśmy zależni od konkretnej firmy, nikt nie może nagle zabronić nam używać danego formatu ani żądać opłat licencyjnych za korzystanie z niego;
- jego pełna dokumentacja powinna być publicznie dostępna — dzięki temu każdy będzie w stanie sprawdzić, czy format nie zawiera rażących błędów lub luk w zabezpieczeniach (bardzo ważne w przypadku przechowywania ważnych informacji);
- powinien być uznany przez międzynarodowy organ normalizacyjny (np. ISO) — aby różni producenci nie próbowali „ciągnąć” formatu w swoją stronę rozbijając go na dziesiątki niekompatybilnych ze sobą wersji;
- powinien być „human readable” (czyli tekstowy, a nie binarny) — aby w ostateczności można było „wyciągnąć” informacje za pomocą zwykłego edytora tekstowego. [...]

Jeśli zapisujesz swoje dane w formacie, którego dokumentacja nie jest publicznie dostępna, prosisz się o kłopoty. Wiem, że dziś mówisz sobie, że MS Word nie zniknie z dnia na dzień i — jakby co — to się przekonwertuje. Ale zastanów się, ile takich dokumentów dziennie produkujesz? Ile potrwa konwersja? Będzie Ci się chciało? Będziesz pamiętać o wszystkich plikach?

Paweł Tkaczyk, *Otwarte formaty plików*, <http://paweltkaczyk.midea.pl/25>

## Najczęściej stosowane „systemy” zarządzania dokumentami

- Tradycyjny system obiegu dokumentów papierowych (szafy, segregatory, asystentka, goniec).
- Poczta elektroniczna, wymiana przy pomocy dyskietek, pen-drive’ów, itp.
- Współdzielony system plików (dysk sieciowy).

Najczęściej stosowanym „systemem” obiegu dokumentów elektronicznych jest poczta elektroniczna. Dokumenty przesyłamy sobie jako załączniki do listów. Jednak taka praktyka powoduje problemy w sytuacji, gdy kilka osób pracuje nad tym samym dokumentem – każdy ma swoją kopię dokumentu i coś z nią robi, więc po jakimś czasie nikt nie wie, kto ma najnowszą wersję.

Alternatywnym rozwiązaniem jest tzw. dysk sieciowy. Jeśli otwieramy i edytujemy dokumenty bezpośrednio na dysku sieciowym, to opisanego powyżej problemu udaje się często uniknąć. Ale wciąż łatwo o kłopoty – ewentualne nieświadome nawet zniszczenie dokumentu czy części jego zawartości powoduje niemożność jego odzyskania (chyba że z systemowych kopii zapasowych). Nie można też śledzić historii zmian w dokumencie, co jest istotne, gdy pracuje nad nim wiele osób.

## Kiedy przestaje wystarczać system plików

- Zasoby informacyjne:
  - o dużej objętości,
  - o skomplikowanej strukturze i powiązaniach,
  - o dużej wartości,
  - o długim cyklu życia informacji,
  - o dużej częstotliwości aktualizacji informacji.
- Organizacja:
  - wieloosobowe zespoły,
  - wysoka specjalizacja członków zespołu,
  - rozproszenie geograficzne.

W wymienionych przypadkach, korzystając wyłącznie z systemu plików oraz biurowych aplikacji, szybko natkniemy się na problemy redundancji (ta sama informacja powtórzona w wielu miejscach), nieaktualnych informacji oraz kłopoty ze znalezieniem właściwej informacji czy koordynacją prac zespołu autorów/redaktorów.

## Kiedy przestaje wystarczać system plików

- Przykłady:
  - wydawnictwo encyklopedyczne,
  - wydawnictwo prawnicze,
  - wydawca czasopism,
  - koncern przemysłowy, producent zaawansowanych technicznie urządzeń,
  - operator rozległej sieci telekomunikacyjnej, energetycznej, ...,
  - organizacja oparta na wiedzy,
  - administracja państwowa.

Organizacje wymienione na slajdzie można podzielić na dwie grupy:

- wydawnictwa, dla których treść dokumentów jest głównym produktem i podstawą egzystencji ekonomicznej,
- organizacje oferujące na tyle skomplikowane produkty i usługi, że ich istnienie jest niemożliwe bez dokumentacji, instrukcji, itp.

## Proste rozwiązania

- Centralne repozytoria, np.:
  - CVS (Concurrent Versions System),
  - SVN (Subversion).
- Typowe funkcje:
  - centralne składowanie dokumentów,
  - lokalne kopie, synchronizowane z repozytorium,
  - blokowanie dokumentów do edycji i zwalnianie blokady po edycji,
  - wersjonowanie dokumentów,
  - możliwość równoległej edycji dokumentów przez wiele osób i scalanie dokumentów.
- Rozwiązania typu wiki, np. MediaWiki, MoinMoin.

Repozytoria typu CVS i SVN są popularne wśród informatyków i powszechnie używane przy programowaniu zespołowym. Korzystanie z takiego repozytorium polega na pobraniu jego zawartości na własny dysk i korzystaniu z niej jak z lokalnych plików. Zaletą takich repozytoriów jest możliwość jednoczesnej edycji jednego pliku przez kilka osób. Zmiany wprowadzone przez poszczególne osoby są bowiem scalane z głównym dokumentem.

Popularność zdobywają też systemy typu wiki, czyli strony internetowe, które można tworzyć i modyfikować bez ograniczeń z poziomu przeglądarki internetowej. Wikipedia – wolna encyklopedia, oparta na tym pomysle, jest fenomenem socjologicznym. Także wiele firm używa własnych serwerów wiki do organizowania i udostępniania wewnętrznych zasobów informacyjnych, np. instrukcji, procedur.

## Rozwiązania z wyższej półki

- Rodzaje i odmiany systemów zarządzania dokumentami:
  - *Web Content Management Systems* – zarządzanie zawartością witryny internetowej,
  - *Enterprise Content Management Systems* – zarządzanie dokumentami biznesowymi organizacji,
  - systemy obiegu dokumentów i spraw,
  - systemy publikacyjne,
  - portale korporacyjne,
  - system do pracy grupowej,
  - elektroniczne archiwa.

Systemy informatyczne, w których mamy do czynienia z dokumentami, mogą mieć bardzo różne przeznaczenie, funkcjonalność i budowę. Łączy je w zasadzie jedynie to, że posługują się dokumentami. Są wśród nich systemy publikacyjne – zarówno profesjonalne systemy do publikacji papierowych, zaawansowane rozwiązania do tworzenia portali korporacyjnych, jak i proste systemy do tworzenia witryn internetowych. Są też systemy wspierające obieg dokumentów wewnątrz organizacji, w tym systemy obiegu dokumentów i spraw, coraz częściej spotykane w urzędach. Są wreszcie systemy wspierające pracę grupową zespołów współpracowników, czy rozwiązania obsługujące elektroniczne archiwa akt.

