PRACA MAGISTERSKA

Skoro Wordem klepie się tak dobrze,
to po co się starać?

Czym będziemy się zajmowali?

PRACA
= FORMA + TREŚĆ

FORMA
= JĘZYK + TYPOGRAFIA + SZATA GRAFICZNA

JĘZYK
= ORTOGRAFIA + INTERPUNKCJA

TREŚĆ
= KAWA + PRACA + STRES…

Język

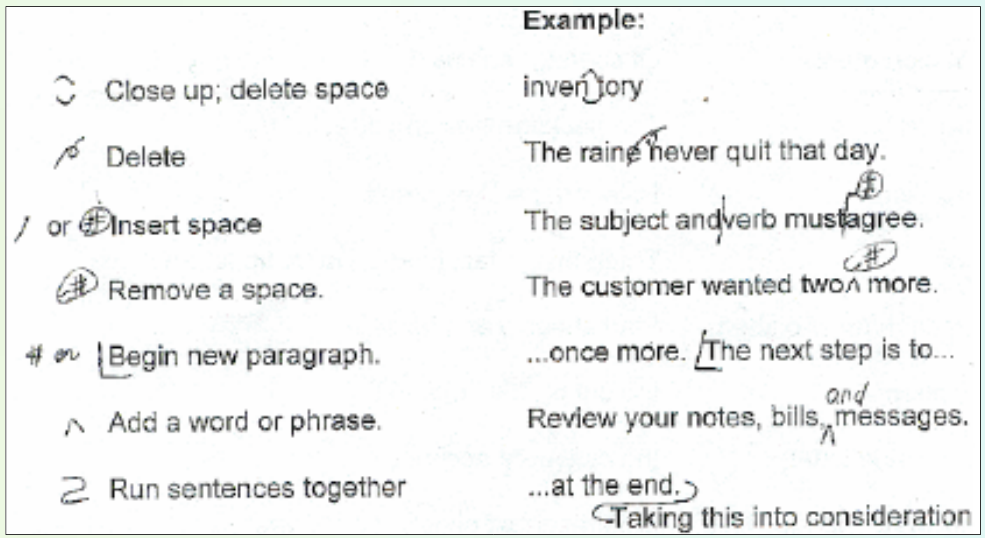
- Praca ze słownikami
  - słowniki wbudowane
  - słowniki zewnętrzne (ispell, aspell, WordNet)
- Powtórne czytanie tekstu
- Styl techniczny
  - Mary McCaskill: *A Handbook for Technical Writers and Editors*
  - Słownictwo obcojęzyczne w języku polskim
  - Wyd. HELION
    - http://helion.pl/autor/6slown.htm

Typografia

- To, na co nie zwraca się uwagi jest ważne!
- Właściwe użycie znaków specjalnych
  - problem „kreski”
    - en dash (–), em dash (—), hyphen (-), minus (−): – — - −
  - problem cudzysłowu i apostrofu („” ‘’ `”” »«)
  - problem wielokropka (… vs. ...)
  - problem cytowań (w tekście i bibliografii)

Typografia (2)

- Typografia zależy od języka
  - Mary McCaskill: *A Handbook for Technical Writers and Editors*
    - *http://stipo.larc.nasa.gov/sp7084/*
  - *Chicago Manual of Style (~1000 stron!)*
  - Robert Chwałowski: *Typografia typowej książki*
- Znaki korektorskie



Szata graficzna

- „Swój styl mam, lubię go sam”
- „klej dobrze trzyma, gdy go ni ma”
  – A. Słodowy (?)
- „Proszę nie silić się na "elegancki" styl, gdyż najczęściej wychodzi styl bombastyczny”
  - M. Drozdowski
- Jeśli nie znasz się na składzie, lepiej zaufać gotowym stylom
  - Czcionki, światła, wcięcia, wyróżnienia, akapity, rozmiary…

Żelazne zasady i reguły

- Zawarte np. na stronie M. Drozdowskiego (choć nie ze wszystkimi trzeba się zgodzić :)
  - http://www.cs.put.poznan.pl/mdrozdowski/txt/jak_mgr.html

W czym pisać pracę?

- Jak z bamboszami – w czym jest wygodnie
- Narzędzia jednak różnią się…
  - poziomem stresu dozowanego użytkownikom
  - łatwością obsługi
  - wymaganym stopniem wyobraźni (Well… You See… Imagine What You'll Get!)
  - wsparciem dla wstawiania rysunków, grafiki, wzorów
  - dostępnością gotowych stylów
  - jakością wydruku

Popularne narzędzia (przegląd)

- WYSIWYG (…sometimes)
  - Microsoft Word
  - Open Office
- WYSIWY… typed
  - LaTeX (TeX)
  - Docbook

Microsoft Word

- Pros
  - WYSIWYG, dobry interfejs
  - znany większości użytkowników
  - świetne słowniki (choć nie nieomylne)
  - sporo gotowych stylów o akceptowalnej jakości

Microsoft Word

- Cons
  - trudna obsługa skomplikowanych stylów (i poprawiania istniejących)
  - naganna implementacja list
    - 24 strony tłumaczenia jak unikać błędów w listach…
  - naganna implementacja wklejania i przenoszenia ilustracji
  - problemy z utrzymywaniem spójności pliku
  - słaba obsługa polskiej interpunkcji
    - PITypo: http://www.cs.put.poznan.pl/dweiss/index.php/projects/pltypo/index.xml?lang=pl

Crawlers, robots and spiders

Specialized software agents were written in order to solve both the coherence and Web discovery problems. These programs, usually named **crawlers, robots** or **spiders**, traverse the graph of *hypertext links*, which are part of HTML, and provide new documents for indexing services. A simple algorithm of a Web spider could look like below:

Figure 2.1.1-1 Simple algorithm of a Web-spider, source: [Salberg, 99]

```
SPIDER(URL_pool url_pool, Document index) {
    while (url_pool not empty) {
        url = pick URL from url_pool
        doc = download url
        new_urls = extract URLs from doc
        index (url,doc) in search engine's database
        for each u in new_urls
            if (u not in indexed_urls)
                add u to url_pool
    }
}
```

It is obvious, that crawlers are characterized by their robot-like behavior: traversing from page to page, except for the hyperlinks, they completely ignore the remaining content of the analyzed page. More intelligent systems utilize Artificial Intelligence techniques, where, for instance, crawlers interact and together with the environment create *multiagent cooperative systems*. Amalthea [Moukas and Maes, 98] is an example of such a project, where a user provides feedback to a set of agents. These agents are born, evolve or die according to the usability of information they found, where that usability can be measured without the knowledge of the user (time spent reading an article), or with his explicit feedback – ranking chart or anything alike.

It must be stated that even though very interesting, such complex systems are not in commercial use to date.

2.1.2 STORING THE INFORMATION

Estimated size of the Web

The Internet is without a doubt an immense source of data. Let us assume an average document (HTML page or any other indexable element) is only 4kB. Taking Google's 1.3 billion indexed documents, the Web would be at least the size of 20 terabytes. This figure has to be multiplied if we realize that search engines cover only a slice of the entire Web (around one third according to the study by Lawrence and Giles [Lawrence and

OpenOffice Writer

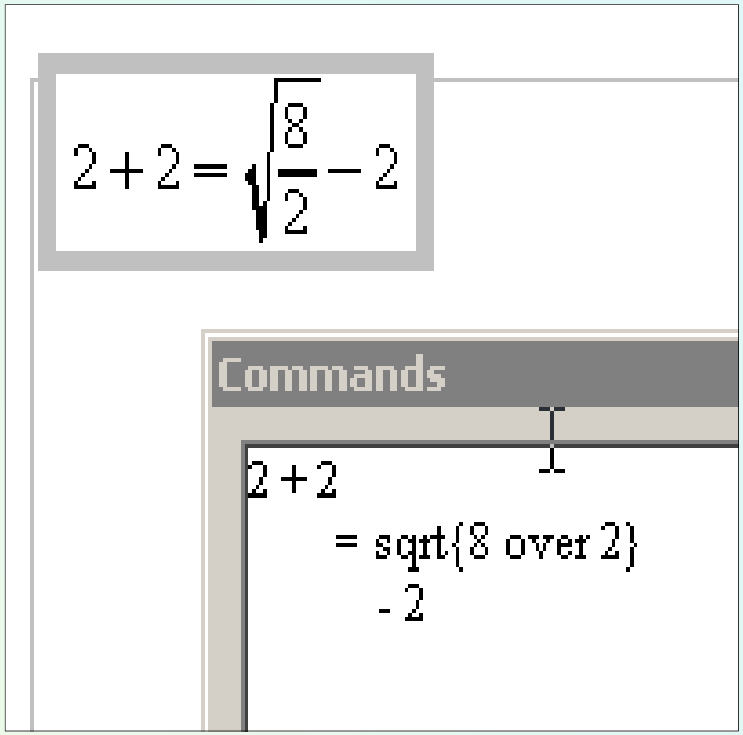
- Pros
  - WYSIWYG
  - dobre słowniki (również polski)
  - darmowy
  - bezproblemowy import/ eksport do MS Word
  - dobra obsługa stylów

OpenOffice Writer

- Cons
  - Inny interfejs niż Worda
  - uboższy w funkcjonalność (ale za to działa)
  - edycja równań tekstowa (ale za to działa)

OpenOffice Writer

- dobry, stabilny edytor
- darmowy



The real thing: TeX i LaTeX

- Pros
  - „Hello world” – jedyny program bez błędów?
  - jeden z najlepszych systemów składu tekstów
  - pięknie składa tekst, wzory, grafikę
  - darmowy
  - *informuje* o błędach lub niemożności składu
  - polska interpunkcja i składnia
  - idealny do dużych publikacji
  - dużo gotowych stylów
  - język *programowania*

The real thing: TeX i LaTeX

- Cons
  - duża praca wejścia
  - wymaga nadzoru (nie wszystko zrobi sam)
  - trudny w czytaniu (szczególnie wzory)
  - sensowny eksport jedynie do PDF/ PS
  - Słabe darmowe edytory ze wsparciem dla LaTeXa (oprócz emacsa ;)

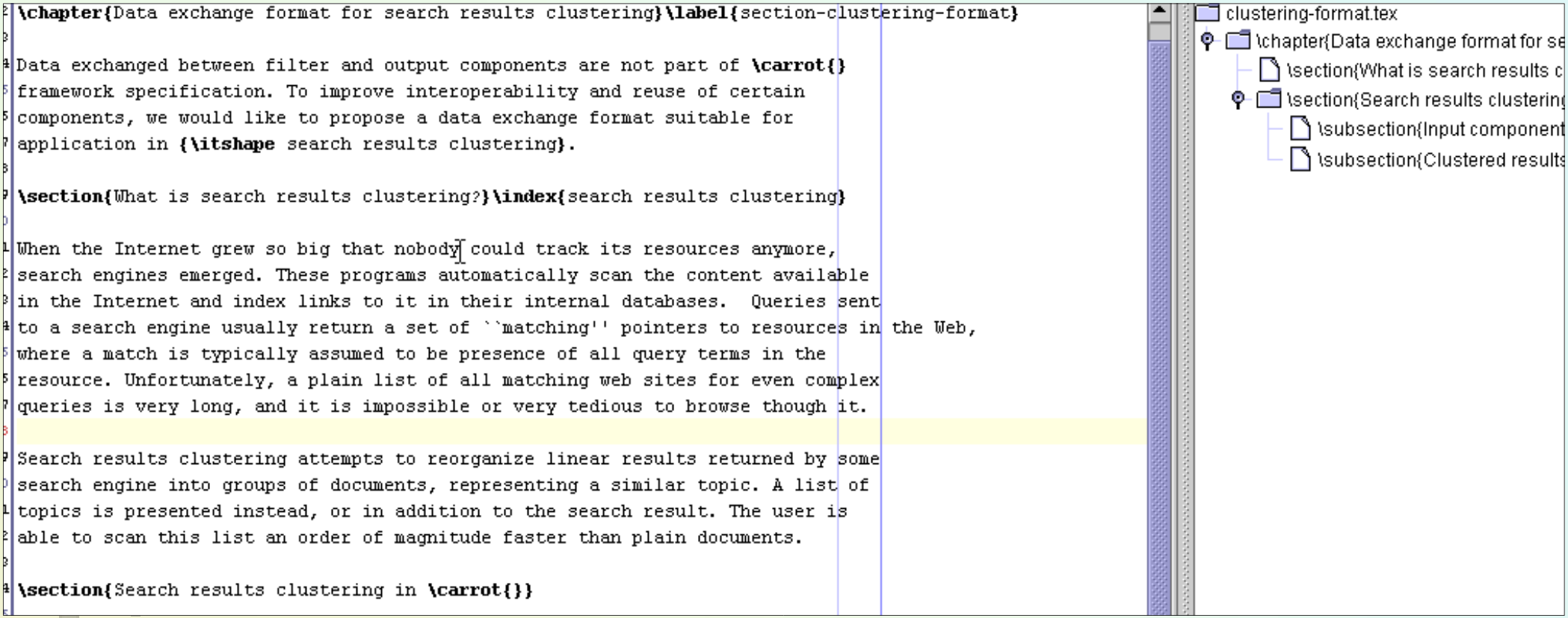
TeX i LaTeX

```
% Carrot2 Guide.

% Load document class and packages.

\newif\ifpdf
\ifx\pdfoutput\undefined
\pdffalse % we are not running PDFLaTeX
\else
\pdfoutput=1 % we are running PDFLaTeX
\pdftrue
\fi

\ifpdf
    \documentclass[pdftex,10pt,a4paper,oneside]{book}
    \usepackage{thumbpdf}
    \usepackage[bookmarks,bookmarksopen,bookmarksnumbered,colorlinks,br
    \usepackage[pdftex]{graphicx}
    \pdfcompresslevel=9
\else
    \documentclass[10pt,a4paper,oneside]{book}
    \usepackage[colorlinks=false,breaklinks=true,pdfborder={0 0 0}]{hyp
    \usepackage{graphicx}
\fi
```

TeX i LaTeX

```
1  %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
2  % Contents: LaTeX style for Carrot documentation
3  %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
4
5
6  \ProvidesPackage{carrot-manual}
7  \RequirePackage[english]{babel}
8  \RequirePackage{fancyvrb}
9  \RequirePackage{fancyhdr}
10 \RequirePackage{amsmath}
11 \RequirePackage{ccaption}
12 \RequirePackage{boxedminipage}
13
14 %
15 % Lets have some nice headings
16 %
17 \pagestyle{fancyplain}
18 \renewcommand{\chaptermark}[1]{\markboth{#1}{}}
19 \renewcommand{\sectionmark}[1]{\markright{\thesectio
20 \lhead[\fancyplain{}{\bfseries\thepage}]
21     {\fancyplain{}{\bfseries\rightmark}}
22 \rhead[\fancyplain{}{\bfseries\leftmark}]
23     {\fancyplain{}{\bfseries\thepage}}
24 \cfoot[]{}
25 \addtolength{\headheight}{1.6pt}
26
```

TeX i LaTeX


\chapter{Data exchange format for search results clustering}\label{section-clustering-format}
Data exchanged between filter and output components are not part of \carrot{}
framework specification. To improve interoperability and reuse of certain
components, we would like to propose a data exchange format suitable for
application in {\itshape search results clustering}.
\section{What is search results clustering?}\index{search results clustering}
When the Internet grew so big that nobody could track its resources anymore,
search engines emerged. These programs automatically scan the content available
in the Internet and index links to it in their internal databases. Queries sent
to a search engine usually return a set of ``matching'' pointers to resources in the Web,
where a match is typically assumed to be presence of all query terms in the
resource. Unfortunately, a plain list of all matching web sites for even complex
queries is very long, and it is impossible or very tedious to browse though it.
Search results clustering attempts to reorganize linear results returned by some
search engine into groups of documents, representing a similar topic. A list of
topics is presented instead, or in addition to the search result. The user is
able to scan this list an order of magnitude faster than plain documents.
\section{Search results clustering in \carrot{}}
clustering-format.tex
\chapter{Data exchange format for se
\section{What is search results c
\section{Search results clustering
\subsection{Input component
\subsection{Clustered results

TeX i LaTeX

```
\subsection{Command-line tools} % {{{

The common library contains a set of useful command-line programs, which can be
used for debugging or testing purposes. In order to use these programs, make
sure you have the shared JAR library in CLASSPATH. This can be for instance
achieved by adding: \texttt{-Djava.ext.dirs=runtime/shared/lib} to the
invokation line of Java (assuming we are in the top directory of the
distribution)\footnote{Backslash character should be used in path on
Windows}. A handful or shell scripts to the most frequently used commands is
also provided in \texttt{/bin} directory of the distribution.

\begin{itemize}
    \item{
        \texttt{QueryInputComponent} utility can be used to send a search reques
        component type. For example, typing the command below (assuming the serv
        at the given URL, should yield 50 snippets from a random snippet
        generator for random seed \textit{anyquery}.

        \footnotesize
        \begin{Verbatim}
java -Djava.ext.dirs=runtime/shared/lib
   com.dawidweiss.carrot.tools.QueryInputComponent
      http://localhost:8080/snippet-generator/service/random 50 anyquery
        \end{Verbatim}
    }

    \item{
        \texttt{QueryFilterComponent} utility can be used to send search result
        to a certain filter or output component.
```

TeX i LaTeX

```
This is pdfTeX, Version 3.14159-1.10a (MiKTeX 2.2)
(developers-guide.tex(pdftex.cfg)
LaTeX2e <2001/06/01>
Babel <v3.7h> and hyphenation patterns for english, french, german, ngerman, du
mylang, nohyphenation, loaded.
(C:\texmf\tex\latex\base\book.cls
Document Class: book 2001/04/21 v1.4e Standard LaTeX document class
(C:\texmf\tex\latex\base\bk10.clo)) (C:\texmf\tex\generic\thumbpdf\thumbpdf.sty

Package thumbpdf Warning: Thumbnail data file `developers-guide.tpt' not found.


) (C:\texmf\tex\latex\hyperref\hyperref.sty
(C:\texmf\tex\latex\graphics\keyval.sty)
(C:\texmf\tex\latex\hyperref\pd1enc.def)
(C:\texmf\tex\latex\00miktex\hyperref.cfg)
Implicit mode ON; LaTeX internals redefined
(C:\texmf\tex\latex\hyperref\backref.sty)
(C:\texmf\tex\latex\latex2html\url.sty))
*hyperref using driver hpdftex*
(C:\texmf\tex\latex\hyperref\hpdftex.def (C:\texmf\tex\latex\psnfss\pifont.sty
(C:\texmf\tex\latex\psnfss\upzd.fd) (C:\texmf\tex\latex\psnfss\upsy.fd)))
(C:\texmf\tex\latex\graphics\graphicx.sty
(C:\texmf\tex\latex\graphics\graphics.sty (C:\texmf\tex\latex\graphics\trig.sty
) (C:\texmf\tex\latex\00miktex\graphics.cfg)
(C:\texmf\tex\latex\graphics\pdftex.def))) (carrot-manual.sty
(C:\texmf\tex\generic\Babel\babel.sty (C:\texmf\tex\generic\Babel\english.ldf
(C:\texmf\tex\generic\Babel\babel.def)))
(C:\texmf\tex\latex\fancyvrb\fancyvrb.sty
Style option: `fancyvrb' v2.6, with DG/SPQR fixes <1998/07/17> (tvz)
No file fancyvrb.cfg.
) (C:\texmf\tex\latex\fancyhdr\fancyhdr.sty)
(C:\texmf\tex\latex\amsmath\amsmath.sty
For additional information on amsmath, use the `?' option.
(C:\texmf\tex\latex\amsmath\amstext.sty (C:\texmf\tex\latex\amsmath\amsgen.sty)
) (C:\texmf\tex\latex\amsmath\amsbsy.sty)
(C:\texmf\tex\latex\amsmath\amsopn.sty))
(C:\texmf\tex\latex\ccaption\ccaption.sty)
(C:\texmf\tex\latex\ltxmisc\boxedminipage.sty)) (developers-guide.aux
(introduction.aux) (architecture.aux) (contributing.aux) (clustering-format.aux
) (component-descriptions.aux)) (C:\texmf\tex\latex\graphics\color.sty
```

TeX i LaTeX

3.3.4 Command-line tools

The common library contains a set of useful command-line programs, which can be used for debugging or testing purposes. In order to use these programs, make sure you have the shared JAR library in CLASSPATH. This can be for instance achieved by adding: `-Djava.ext.dirs=runtime/shared/lib` to the invokation line of Java (assuming we are in the top directory of the distribution)[5]. A handful or shell scripts to the most frequently used commands is also provided in `/bin` directory of the distribution.

- `QueryInputComponent` utility can be used to send a search request to a certain input component type. For example, typing the command below (assuming the service is running at the given URL, should yield 50 snippets from a random snippet generator for random seed *anyquery*.

```
java -Djava.ext.dirs=runtime/shared/lib
   com.dawidweiss.carrot.tools.QueryInputComponent
      http://localhost:8080/snippet-generator/service/random 50 anyquery
```

- `QueryFilterComponent` utility can be used to send search result data stream to a certain filter or output component.

  The program expects an input file, which should conform filter/ output component's contract (for instance – it should represent search results data). The result returned by the component is written to standard output.

LyX

- LyX jest edytorem przeznaczonym do pracy WYSIWYG z TeXem
- Nigdy nie próbowałem…

Konkluzja (co do LaTeXa)

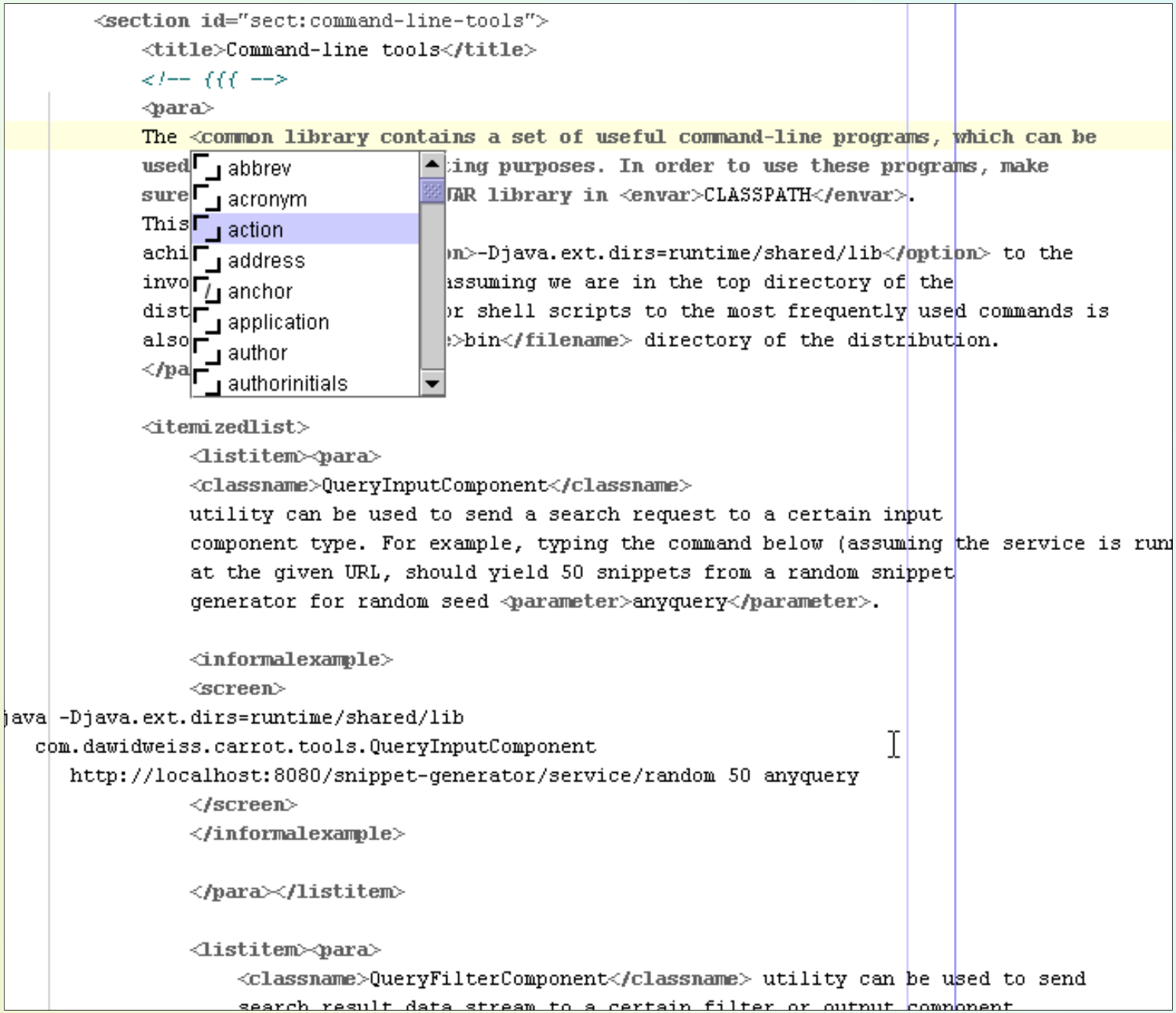
- TeX jest potężnym narzędziem
- jeśli lubisz programować, polubisz TeXa
- jeśli chcesz mieć perfekcyjne formatowanie, użyj TeXa

Docbook

- Pros
  - darmowy
  - definicja pliku w XML
    - poprawność strukturalna zapewniona
    - można używać edytorów XMLowych
  - formatuje ładnie grafikę i tekst
  - eksport do HTML, XSL:FO (PDF)
  - eksport do TeXa i RTFa
  - Instytut ma licencję XEPa ☺

Docbook

- Cons
  - mało gotowych stylów, brzydkie style
  - formatowanie niektórych elementów słabe
    - można użyć eksportu do TeXa
  - problem z wzorami (MathML?)
  - wolny (xsl/ xsl:fo)


<section id="sect:command-line-tools">
<title>Command-line tools</title>
<!-- {{{ -->
<para>
The <common library contains a set of useful command-line programs, which can be
used ing purposes. In order to use these programs, make
sure AR library in <envar>CLASSPATH</envar>.
This
achi n>-Djava.ext.dirs=runtime/shared/lib</option> to the
invo ssuming we are in the top directory of the
dist r shell scripts to the most frequently used commands is
also >bin</filename> directory of the distribution.
</pa
abbrev
acronym
action
address
anchor
application
author
authorinitials
<itemizedlist>
<listitem><para>
<classname>QueryInputComponent</classname>
utility can be used to send a search request to a certain input
component type. For example, typing the command below (assuming the service is run
at the given URL, should yield 50 snippets from a random snippet
generator for random seed <parameter>anyquery</parameter>.
<informalexample>
<screen>
java -Djava.ext.dirs=runtime/shared/lib
com.dawidweiss.carrot.tools.QueryInputComponent
http://localhost:8080/snippet-generator/service/random 50 anyquery
</screen>
</informalexample>
</para></listitem>
<listitem><para>
<classname>QueryFilterComponent</classname> utility can be used to send
search result data stream to a certain filter or output component

3.3.4. Command-line tools

The common library contains a set of useful command-line programs, which can be used for debugging or testing purposes. In order to use these programs, make sure you have the shared JAR library in `CLASSPATH`. This can be for instance achieved by adding: `-Djava.ext.dirs=runtime/shared/lib` to the invokation line of Java (assuming we are in the top directory of the distribution). A handful or shell scripts to the most frequently used commands is also provided in `bin` directory of the distribution.

- `QueryInputComponent` utility can be used to send a search request to a certain input component type. For example, typing the command below (assuming the service is running at the given URL, should yield 50 snippets from a random snippet generator for random seed *anyquery*.

```
java -Djava.ext.dirs=runtime/shared/lib
   com.dawidweiss.carrot.tools.QueryInputComponent
      http://localhost:8080/snippet-generator/service/random 50 anyquery
```

- `QueryFilterComponent` utility can be used to send search result data stream to a certain filter or output component.

3.3.4. Command-line tools

The common library contains a set of useful command-line programs, which can be used for debugging or testing purposes. In order to use these programs, make sure you have the shared JAR library in `CLASSPATH`. This can be for instance achieved by adding: `-Djava.ext.dirs=runtime/shared/lib` to the invokation line of Java (assuming we are in the top directory of the distribution). A handful or shell scripts to the most frequently used commands is also provided in `bin` directory of the distribution.

- `QueryInputComponent` utility can be used to send a search request to a certain input component type. For example, typing the command below (assuming the service is running at the given URL, should yield 50 snippets from a random snippet generator for random seed *`anyquery`*.

```
java -Djava.ext.dirs=runtime/shared/lib
   com.dawidweiss.carrot.tools.QueryInputComponent
      http://localhost:8080/snippet-generator/service/random 50 anyquery
```

- `QueryFilterComponent` utility can be used to send search result data stream to a certain filter or output component.

Docbook

- Docbook w praktyce
  - Edycja w XMLu daje pewność strukturalnej poprawności
  - XML można przetwarzać *przed* Docbookiem przy pomocy dowolnych arkuszy XSLT
  - Docbook i XEP dają ładne wyniki przy wydruku do PDF
  - Stylesheet Staszka Osińskiego
    - http://www.man.poznan.pl/~stachoo/elegant.zip

2 Architecture

2.1 Components and data flow

Carrot^2's architecture is based on certain definitions of *data exchan*
among *components*.

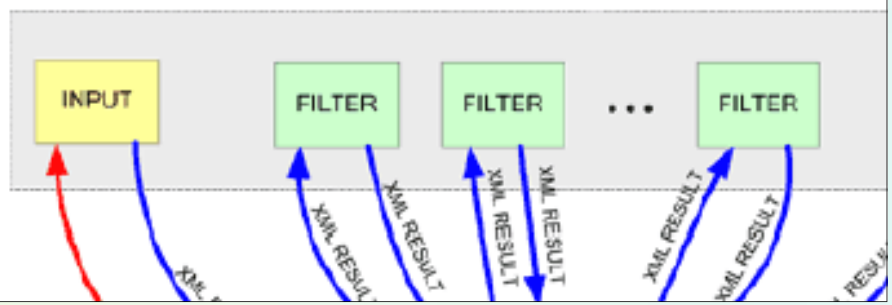


Figure 2.1
Components and data flow in Carrot^2

Przydatne adresy

- OpenOffice
  - http://www.openoffice.org
  - http://www.openoffice.pl
- TeX i LaTeX
  - http://www.tug.org
  - http://www.gust.org.pl
  - Pakiet MikTeX:
- Docbook
  - http://www.docbook.org/
  - http://www.miktex.org/
  - http://www.renderx.com/ (XEP)