

Leksykografia elektroniczna

Praktyki, metody, narzędzia

Krzysztof Nowak

Pracownia Łaciny Średniowiecznej

Instytut Języka Polskiego PAN

Plan wykładu

- **retrodygitalizacja słowników**

- przebieg
- OCR i obróbka obrazu
- narzędzia
- formaty binarne i tekstowe
- modele prezentacji

- **anotacja XML**

- cele
- planowanie

- **XML**

- budowa pliku
- składnia

- **laboratorium 1**

reprezentacja hasła słownikowego (XML)

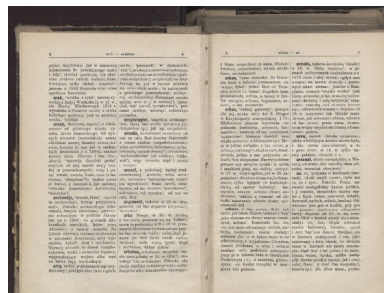
- **standardy**

- LMF
- TEI

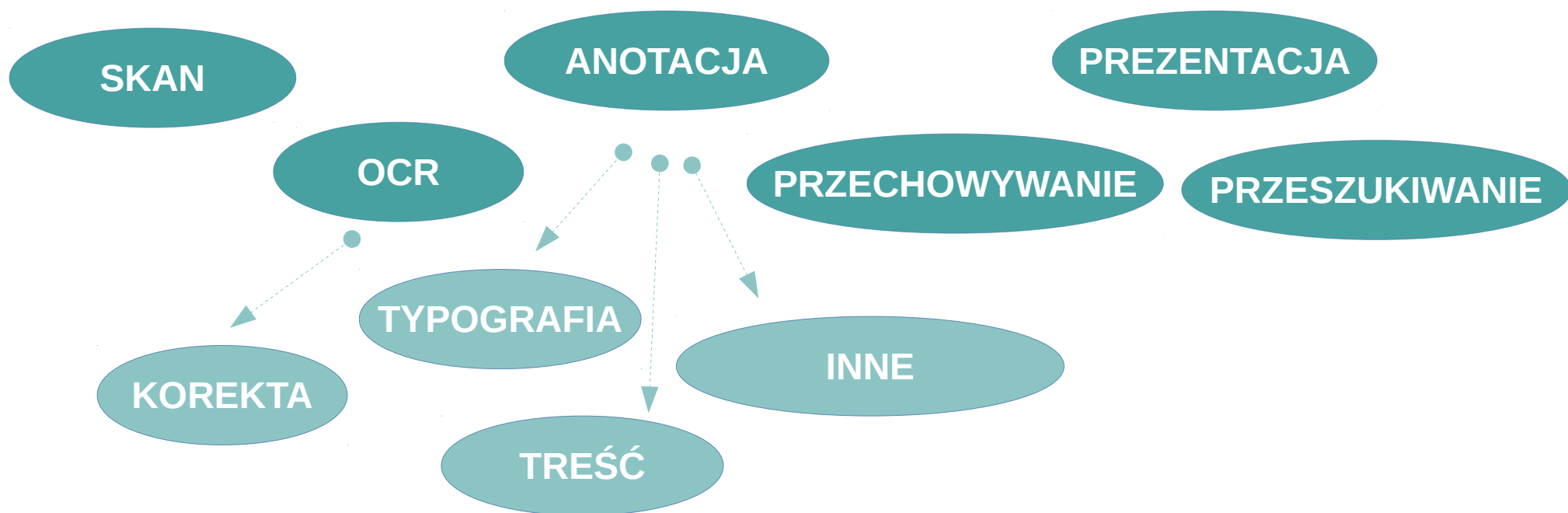
- **laboratorium 2**

- anotacja hasła słownikowego (TEI XML)

Retrodygitalizacja: etapy



arcydziegiel, 'angelica
archangelica', łączy
łac. *archi-* (*arcy-*) z pol.
dziegielem (p.), jak np.
arcydzieło



Pozyskiwanie obrazu: repozytoria i biblioteki cyfrowe



**Federacja Bibliotek
Cyfrowych**

<http://fbc.pionier.net.pl>



Europeana

<http://www.europeana.eu/portal/>



Internet Archive

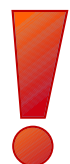
<https://archive.org/>



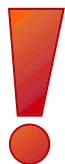
Google Books itp.

<https://books.google.pl/>

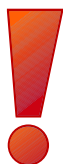
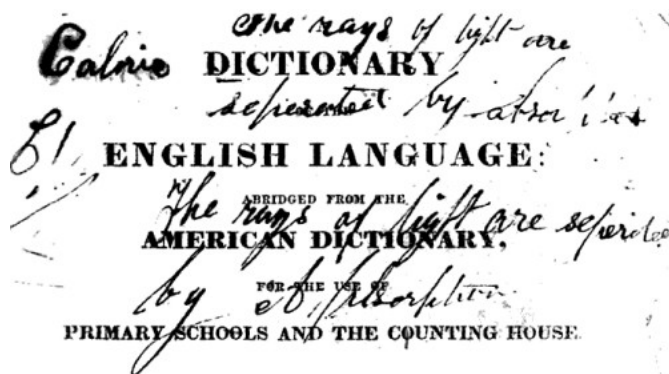
Pozyskiwanie obrazu z repozytoriów: problemy



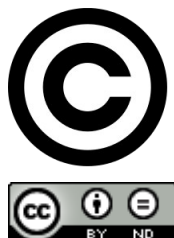
artefakty



niejasna
licencja



wady
egzemplarza



błędy skanowania

- opuszczone strony
- niska rozdzielczość
- zagięcia itp.



Pozyskiwanie obrazu: skanowanie

PARAMETRY

- rozdzielczość: ok. 300-600 DPI
- typ: rodzaj dokumentu
- łatwość obsługi
- szybkość skanowania
- ilość skanów
- dołączone oprogramowanie

SKANER



działowy
(eng. *planetary*)



do dokumentów

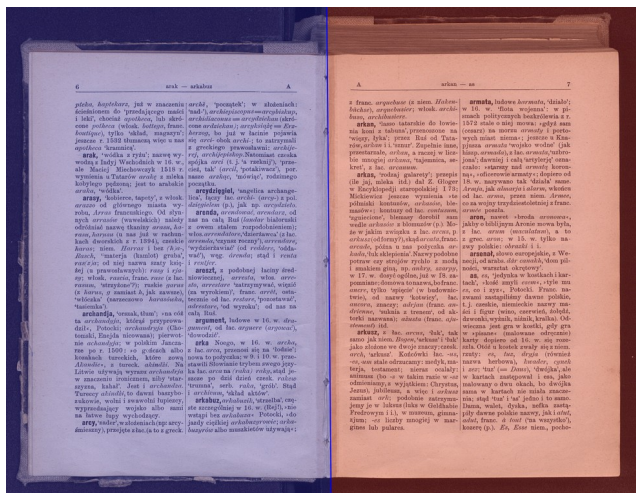


płaski

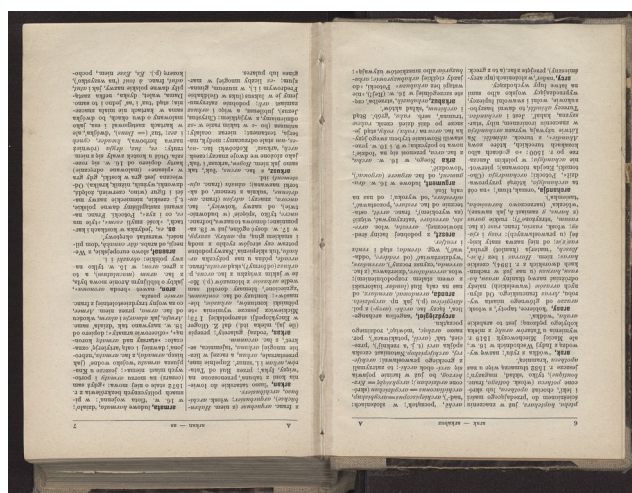
nasze arszak, 'mówią', rodzimego
początku.
arcydzięgiel, 'angelica archange-
lica', łączy łac. archi- (arcy-) z pol.
dzięgielom (p.), jak np. nasze arszak, 'mówią', rodzimego
arrenda, arrendować, arrendarz, od
nas na całą Ruś (landar białoruski
z owym stałym rozpodobnieniem);
włos. arrendatore, 'dzierżawca' (z łac.
arrenda, 'czynsz roczny'), arrendare,
'wyzdzierżawiać' (od reddere, 'odda-
wać'), węg. arrenda; stąd i renta
i rentjer.

Optymalizacja obrazu

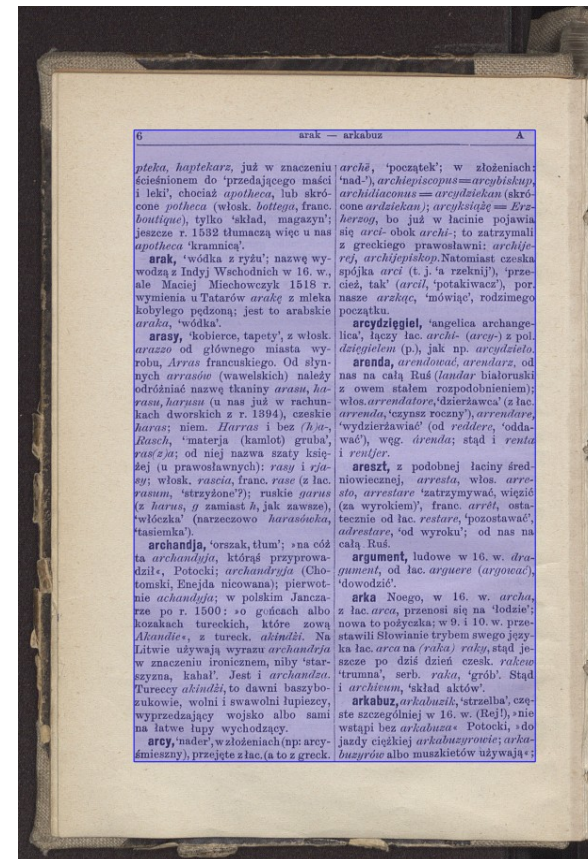
podział strony



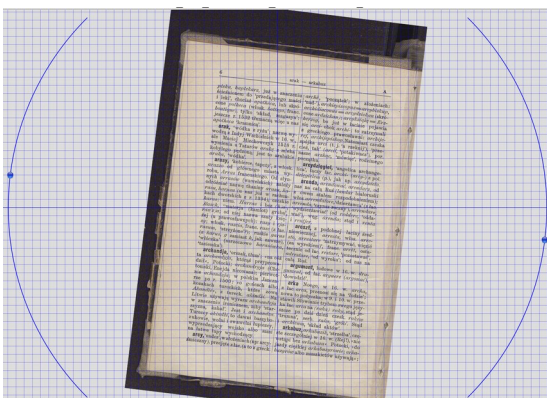
orientacja



wycinanie



wyrównywanie tekstu



binaryzacja, ostrość, kontrast, balans bieli

row, *armata* i t. *szaur*. *Zupełnie inne, przestarzałe, arkan*, a raczej w liczbie mnogiej *arkana*, 'tajemnica, sekret', z łac. *arcanum*.

arkas, 'rodzaj galarety'; przepis (ile jaj, mleka itd.) dał Z. Gloger w Encyklopedii staropolskiej I 78; Mickiewicz jeszcze wymienia »te półmiski kontuzów, *arkasów*, *blernasów*«; kontuzy od łac. *contusum*, 'zgnieciony', *blernasy* - dorobił sam wedle *arkasów* z blonuzów (p.). Może w jakim związku z łac. *arcus*, p. *arkusz* (odformy?), skąd *arcata*, franc. *arcade*, późna u nas pożyczka *arkada*, 'łuk sklepienia'. Nazwy podobne potraw czy strojów rychło z modą i smakiem giną, np. *ankry*, *szarypy*, w 17. w. dosyć ogólne, już w 18. za-nomniane: domowa to nazwa ho franc.

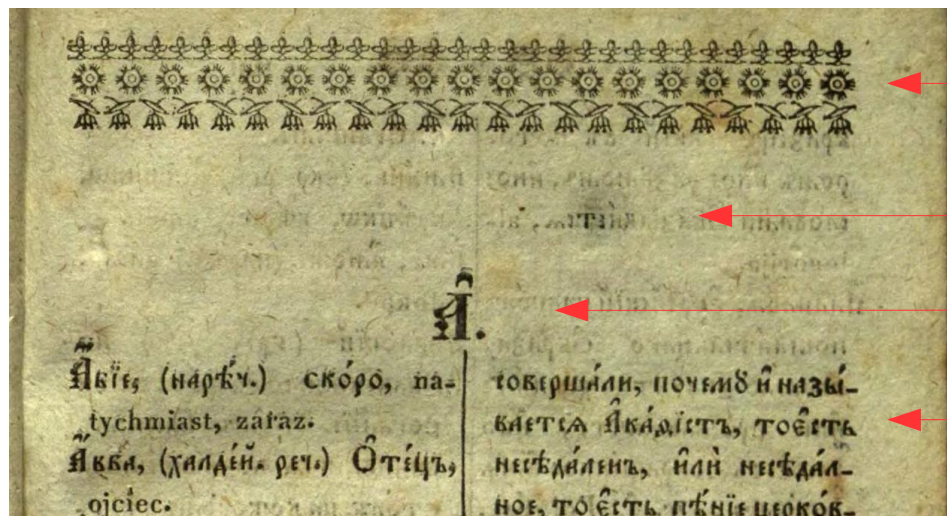
pusza *armusu* wojsko wotnie (jak hiszp. *armada*, z łac. *armata*, 'uzbrojona'; dawniej i całą 'artylerję' oznaczało: »starszy nad *armatą* koronna«, »oficerowie armaty«; dopiero od 18. w. nazywano tak 'działa' same. *Armja*, jak *almarja* i *alarm*, wkońcu od łac. *arma*, przez niem. *Armee*, co za wojny trzydziestoletniej z franc. *armée* poszła.

aron, nawet »broda *aronowa*«, jakby o biblijnym Aronie mowa była, z łac. *arum* (*maculatum*), a to z grec. *aron*; w 15. w. tylko nazwy polskie: *obrazki* i i.

arsenal, słowo europejskie, z Wene-cji, od arabs. *dār canakk*, 'dom pil-ności, warsztat okrętowy'.

as, *es*, 'jedynka w kostkach i kar-tach'. »kość zmvlia *esem*«. »tyle mu

Obraz: problemy z nadestłanych plików

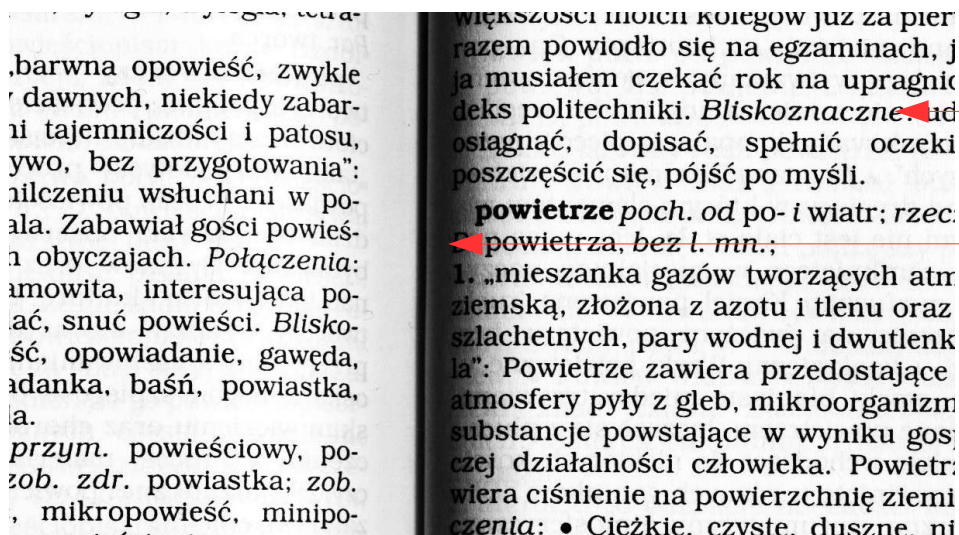


Elementy typograficzne
(linie, zdobienia itp.)

Przebiecia

Niewspierane alfabety,
historyczna czcionka

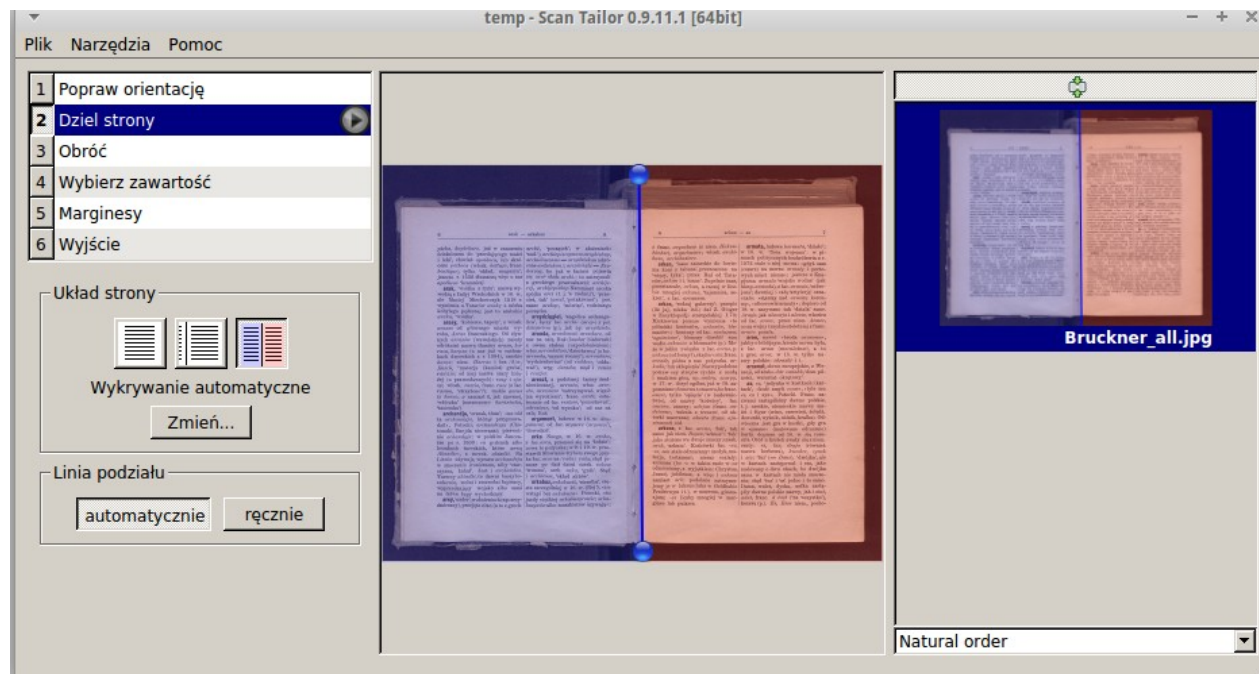
Słaby kontrast tło - tekst



Licencja

Zagięcia

Optymalizacja obrazu: narzędzia



Okno programu ScanTailor

- programy graficzne ogólnego zastosowania (np. Gimp, Photoshop itd.)
- wbudowane w programy do skanowania (np. gscan2pdf, Abbyy FineReader itd.)
- programy do obróbki skanów (np. ScanTailor)

OCR: metody i narzędzia

1. Algorytmy rozpoznawania

- dopasowanie do zdefiniowanego wzorca
- ekstrakcja cech (*feature extraction*) - klasyfikacja

2. Korekta automatyczna

- słowniki
- listy kolokacji

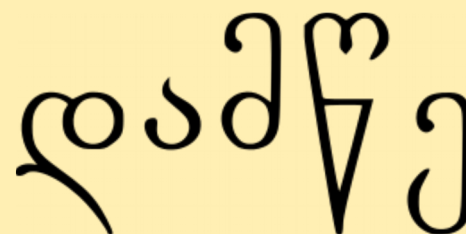
3. Korekta *ad-hoc* i ręczna

T
R
E
N
I
N
G

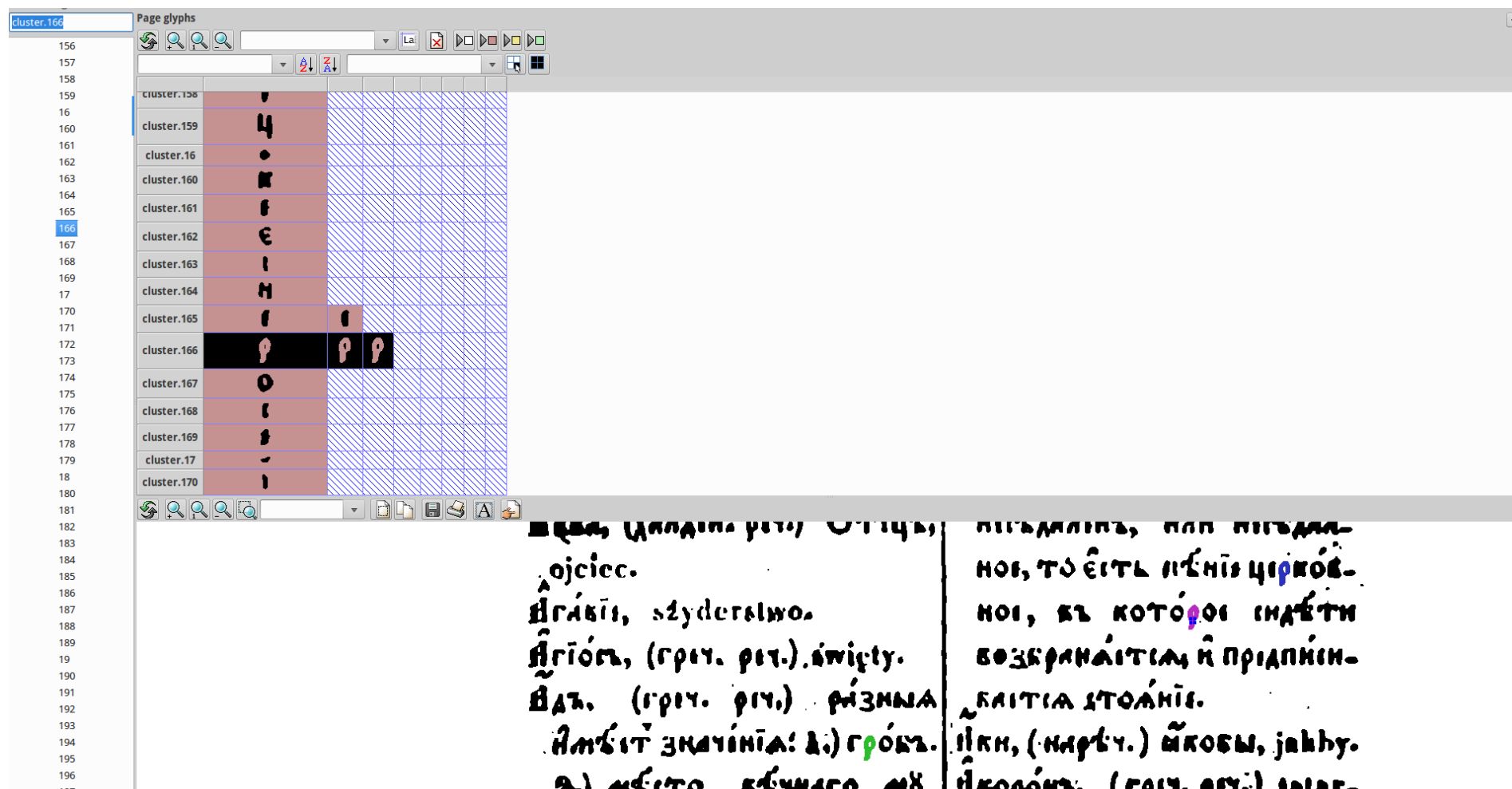
pisma historyczne

	Textur	Rotunda	Schwa- bacher	Fraktur
a	Ꝑ	Ꝓ	Ꝕ	Ꝗ
d	Ꝙ	Ꝛ	Ꝝ	Ꝟ
g	Ꝡ	Ꝣ	ꝣ	ꝥ
n	Ꝩ	Ꝫ	Ꝭ	Ꝯ
o	Ꝯ	ꝰ	ꝲ	ꝴ
Δ	ꝺ	ꝼ	Ꝿ	ꝿ

nieobsługiwane alfabety



Gamera: przykład sesji wykrywania glifów



OCR: narzędzia

Komercyjne / zamknięte

- Abbyy FineReader (Professional / Corporate)
- urządzenia mobilne: CamScanner

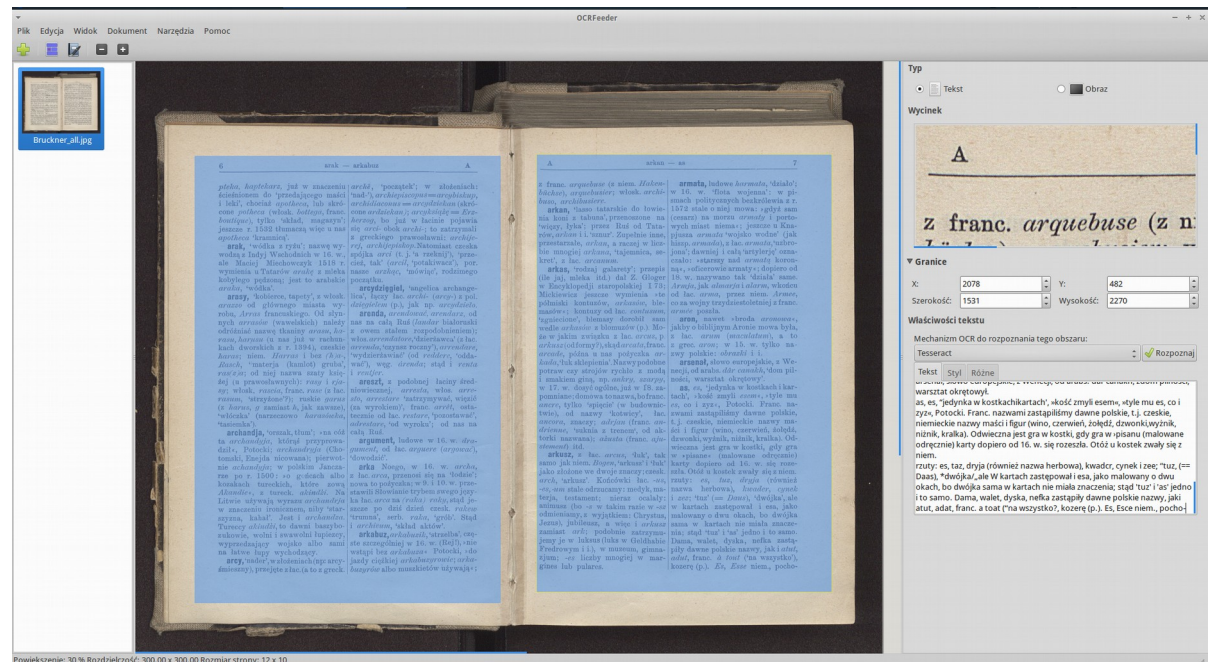
Otwartoźródłowe / darmowe

- GUI: gscan2pdf, OCRFeeder, Gamera
- silniki: Tesseract, CuneiForm, OCRopus

Wbudowane

- Google Dysk
- „Białkowe”
- reCAPTCHA

following
finding



OCR: pliki wynikowe

OBRAZ

arkusz, arkusze i. arkusze - arkusze arkuszy, przestarzałe, *arkana*, a raczej w liczbie mnogiej *arkana*, 'tajemnice, sekrety', z łac. *arcana*.

arkas, 'rodzaj galarety'; przepis (ile jaj, mleka itd.) dał Z. Gloger w Encyklopedji staropolskiej 178; Mickiewicz jeszcze wspomina »te półmiski kontuszów, *arkasów*, białasów«; kontusz od łac. *contusum*, 'zgniecione', białaszy dorobił sam wedle *arkasów* z blonuzów (p). Może w jakim związku z łac. *arcus*, p. *arkusz*(odformy?), skąd *arcata*, franc. *arcade*, późna u nas pożyczka *arkada*, 'łuk sklepienia'. Naszy podobne potraw czy strojów rychło z modą i smakiem giną, np. *ankry*, *szarpy*, w 17. w. dosyć ogólne, już w 18. zapomniane: domowa tonaswa, bo franc.

ryszak, *armata* wyrazu *armata* (łac. hiszp. *armada*), z łac. *armata*, 'uzbrojona'; dawniej i całą 'artylerię' oznaczało: »stareży nad *armatą* koronną«, »oficerowie *armaty*«; dopiero od 18. w. nazywano tak 'działa' same. *Arania*, jak *aimaria* i *alarm*, w końcu od łac. *arma*, przez niem. *Armes*, co za wojny trzydziestoletniej z franc. *armée* poszło.

aran, nawet »broda *aronowa*«, jakby o biblijnym Aronie mowa była, z łac. *arum* (*maculatum*), a to z grec. *aron*; w 15. w. tylko naszy polskie: *obrazki* i i.

aranat, słowo europejskie, z Węgrzech, od arabs. *dār canaah*, 'dom pilności, warsztat okrętowy'.

as, es, 'jedynka w kostkach i kartach'. »kość zmiły *esem*«, »style mu

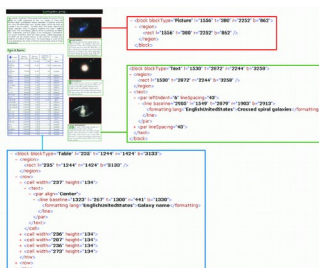


TIFF



PNG, JPG

OBRAZ + TEKST



PDF, DJVu,
natywne pliki
OCR

PDF, DJVu

TEKST + ANOTACJA

```
<entry>
  <sense n="1"/>
  <sense n="2"/>
</entry>
```

pliki tekstowe

pliki tekstowe,
bazy danych,
SPARQL
endpoints

Formaty

binarne (*binary files*)

- **DOC, XLS**
- **PDF**

historyczne

- **ChiWriter**

tekstowe (*text files*)

- **TXT**
- **CSV, TSV**
- **XML (por. ODT, DOCX; ODS, XLSX), HTML**

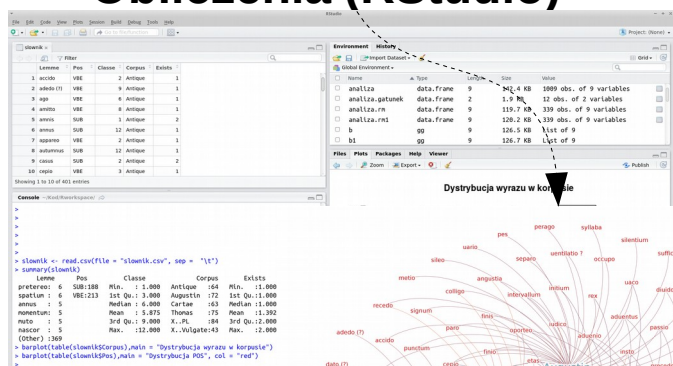
-  vendor lock-in (np. trudna, o ile możliwa, konwersja)
-  odczyt możliwy tylko z pomocą określonego oprogramowania
-  często nieudokumentowane (do niedawna DOC, XLS)

-  dokumentacja, otwarty format: łatwa konwersja
-  odczyt możliwy za pomocą prostego edytora tekstowego
-  łatwa kontrola wersji (git, CSV itd.)
-  łatwo użyć w mnogich kontekstach

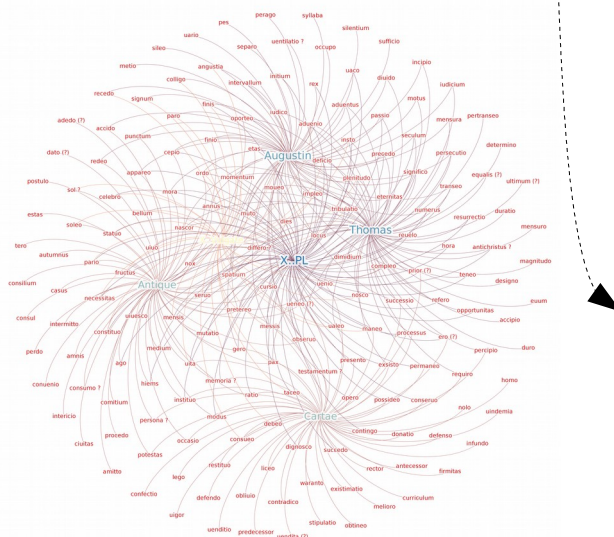
Jeden format, różne środowiska

Lemmat	POS	Klasa	Korpus	Istnieje
accido	VBE	2	Antique	1
adedo	VBE	9	Antique	1
ago	VBE	6	Antique	1

Obliczenia (RStudio)



Wizualizacja (Gephi)



Wyświetlanie (geany)

	Lemmat	POS	Klasa	Korpus	Istnieje
1	accido	VBE	2	Antique	1
2	adedo	(?)	VBE	9	Antique 1
3	ago	VBE	6	Antique	1
4	amitto	VBE	8	Antique	1
5	amnis	SUB	1	Antique	2
6	annus	SUB	12	Antique	1

Edycja (LibreOffice Calc)

	A	B	C	D	E	F
1	Lemmat	POS	Klasa	Korpus	Istnieje	
2	accido	VBE	2	Antique	1	
3	adedo	VBE	9	Antique	1	
4	ago	VBE	6	Antique	1	
5	amitto	VBE	8	Antique	1	
6	amnis	SUB	1	Antique	2	
7	annus	SUB	12	Antique	1	
8						
9						
10						
11						

Obliczenia: przykład sesji w RStudio

File

Edit

Code

View

Plots

Session

Build

Debug

Tools

Help

Go to file/function

Project: (None)

slownik

Filter

	Lemme	Pos	Classe	Corpus	Exists
1	accido	VBE	2	Antique	1
2	adedo (?)	VBE	9	Antique	1
3	ago	VBE	6	Antique	1
4	amitto	VBE	8	Antique	1
5	amnis	SUB	1	Antique	2
6	annus	SUB	12	Antique	1
7	appareo	VBE	2	Antique	1
8	autumnus	SUB	12	Antique	1
9	casus	SUB	2	Antique	2
10	cepio	VBE	3	Antique	1

Showing 1 to 10 of 401 entries

Console

```

>
>
>
>
>
>
> slownik <- read.csv(file = "slownik.csv", sep = "\t")
> summary(slownik)
  Lemme      Pos      Classe      Corpus      Exists
pretereo: 6 SUB:188 Min.   : 1.000 Antique:64 Min.   :1.000
spatium  : 6 VBE:213 1st Qu.: 3.000 Augustin:72 1st Qu.:1.000
annus    : 5          Median : 6.000 Cartae :63  Median :1.000
momentum: 5          Mean   : 5.875 Thomas :75  Mean   :1.392
muto     : 5          3rd Qu.: 9.000 X..PL  :84  3rd Qu.:2.000
nascor   : 5          Max.   :12.000 X..Vulgate:43 Max.   :2.000
(Other)  :369
> barplot(table(slownik$Corpus),main = "Dystrybucja wyrazu w korpusie")
> barplot(table(slownik$Pos),main = "Dystrybucja POS", col = "red")
>

```

Environment

History

Import Dataset

Global Environment

	Name	Type	Length	Size	Value
	analiza	data.frame	9	142.4 KB	1009 obs. of 9 variables
	analiza.gatunek	data.frame	2	1.9 KB	12 obs. of 2 variables
	analiza.rm	data.frame	9	119.7 KB	339 obs. of 9 variables
	analiza.rm1	data.frame	9	120.2 KB	339 obs. of 9 variables
	b	gg	9	126.5 KB	List of 9
	b1	gg	9	126.7 KB	List of 9

Files Plots Packages Help Viewer

Zoom Export Publish

Dystrybucja wyrazu w korpusie

Dystrybucja POS

Eksploracja i prezentacja: biblioteki cyfrowe, programy natywne

wyszukiwanie przyrostowe

wyrażenia regularne

The screenshot shows a web application interface for the University of Warsaw's digital library. The title bar reads "e-biblioteka Uniwersytetu Warszawskiego - Słownik języka polskiego. T....". The main content area displays search results for the term "hajdamacha". The left sidebar shows a list of pages with hits: Page 18 (1 hit), Page 25 (1 hit), Page 42 (1 hit), and Page 52 (1 hit). The main text area shows the definition of "hajdamacha" and "hajduker". The definition of "hajdamacha" includes the following text: "Hajdamacha, y, lm. y] p. Hajdamak: Wy, to nie goście, ale hajdamachy. Hajdamacki i. a. X Hajdamaczny przym. od Hajdamak: Ja ś. nie znałem, nawet nie widziałem z Jackiem, ledwie słyszał o jego życiu hajdamackim. Mick. (= hultajskim, łotrowskim). Mina hajdamačka (= tęga, imponująca, zawadłacza). 2. [Maś hajdamačka] p. Maś. Po hajdamacku przys. jak hajdamak, w sposób łotrowski, niegodziwy: Postępować po hajdamacku. Hajdamactwo, a, blm. 1. życie rozbójnicze, rozbójnictwo, łotrostwo: Miał wyryte hajdamactwa herby na gębie. Boh. (= szramy, blizny). 2. zb. p. Hajdamaczyna: H. ukraińskie. 3. wog. rozbójnicy, zbójce, rabusie, zbójcy. Hajdamaczek, czka, lm. czki p. Hajdamak: Mójże ty hajdamaczku najmiłszy! — rzekł rozrzucony Zagłoba. Sienk. Hajdamaczenie, a, blm., czynność cz. Hajdamaczyć." The definition of "hajduker" includes the following text: "hajduka: Że on taniec (t. j. H.) w niesmak poszedł w narodzie, dowodzi XVI i XVII w. przysłowie: na brzydki, niezgrabny i niełuby taniec mówiono, że to H. taniec lub skoki. Wóje. 4. [Maś hajduka] p. Maś. Po hajduku przys. jak hajduk: Co po hajduku, to nie po ludzku. Prz. <p. Hajduk> Hajductwo, a, blm. 1. życie i stan hajducki. 2. zb. hajducy, kupa hajduków. Hajduczek, czka, lm. czkowie i. p. Hajduk. Przen. (o osobie zuchowatej): Pryncypalnie mi ś. ów H. udał. Sienk. Dajże spokój, hajduczku najśłodszy. Sienk. 2. lokajczyk ubrany po węgiersku. Hajduczenie, a, blm., czynność cz. Hajduczyć. Hajduczka, i, lm. i i. forma ż. od Hajduk: Rejencja utaskawiła hajduczkę Miłę. 2. żona hajduka. 3. chatka, chałupka: Zawsze śmierć za pasem, byś też dobrze nigdy na wojnie nie był, tylko w licej gdzie hajduczce leżał. Jan. Ona ubogusiałka, co H. była, miała ś. w...

Eksploracja i prezentacja: SOLR, XTF itp.



Eksploracja i prezentacja: Poliqarp dla DJVu

Zapytanie:

Szukaj

Wyniki

Dotychczas znaleziono 317 wyników

Wyświetlanie wyników 1-25

Następnych 25

1.	,t]]ltsy,	zapaśnictwo	. Cf BIEDZIĆ SIĘ,	Zakładka
2.	porządne po[tępowanie Albo	namieftnictwo	) oc	Zakładka
3.	: wiedząc i[z	veżeHnictwo	z H	
4.	Byczek: Wezmie[na	ojiarownictwo	] cie	

zapaśnictwo
Sł stp; Cn, Linde brak.
ENIEC
Sł stp; Cn brak, Linde XVIII w.: *biodrowy*,
OWY
CA Cn; Linde XVIII w.
UNKA
IEC (23) sb m
nacza é; prawdopodobnie pierwsze i ostat-
Cn).
miec (9). ◇ *A* bierzeńcu (10). ◇ *A* bie-
I bierzeńc(e)m (2).
otuje, Linde XVIII w.
ella saxifraga L. (Rost); roślina lecznicza
zkowatych; pimpinella Mymer', Cn; hir-
agoselinum sirenii, saxifraga hircina Cn

namieftnictwo
Mącz, Calep; pugilatio Mącz (11): A v Grekow. Czterzi
rzeczy były ktozych [!] oni swoich dzieci zwykli byli
veżyć. Pierw[za/ Piłmo. Wtora/ Biedzenie. KwiatKsiąż
H4; Potrzeba ich [młodzieńcykow] telfz czwicz[yc] w bie-
ganiu/ tkakaniu/ biedzeniu/ y w[czyrmowaniu KwiatKsiąż
Ov, H4v; Conflictatio verbale. Częfte biedzenie Mordo-
wanie/ fztotowanie albo potyk[anie]. Mącz 130c, 198b
[2 r.], 330a; Biedzenie Iakob[z Aniolem SkarŻyw 348
marg; Calep 615a.
Szereg: »biedzenie a starcie«: yż w biedzeniu a wftar-
ciu zrownymi takow[ą] cierpliwość miewali/ yże chociay
też moc[ą] byli o ziemię vderzeni/ albo z przigody w bie-
dzeniu vpadli/ tedyby woleli być zabitemi y krom dufze/
n[by]by zwycięstwa pozwolić mieli. KwiatKsiąż N2.
Synonimy: mordowanie, pasowanie, starcie, zapaso-
wanie, zapasy, **zapaśnictwo**.
Cf BIEDZIĆ SIĘ, BIEDZENIE SIĘ
ZZie
BIEDZENIE SIĘ (1) sb n

<http://poliqarp.wbl.klf.uw.edu.pl/pl/>

- Karłowicz, Kryński, Niedźwiedzki, Słownik języka polskiego
- Słownik polszczyzny XVI wieku
- Linde, Słownik języka polskiego (2. wydanie)

Eksploracja i prezentacja: Doroszewski



Słownik języka polskiego
pod redakcją Witolda Doroszewskiego

<http://doroszewski.pwn.pl/>

wyszukiwanie wg wyrazu hasłowego

poprzednie hasła

- słowniczek
- słowniczy
- słownie
- słownik**
- słownikarnia
- słownikarski
- słownikarstwo
- słownikarz
- słownikowo
- słownikowy
- słowność
- słowny
- słowo

słownik *m III* 1. «zbiór wyrazów ułożonych według pewnej zasady (najczęściej alfabetycznej), objaśnianych pod względem znaczeniowym, często zilustrowanych przykładami użycia»: Samo stwierdzenie, że słownik musi być normatywny, dotyczy wszystkiego w słowniku, każdego hasła i sposobu jego opracowania. *DOR. Leks. 61*. Solidny słownik, dzieło wymagające rzetelnej roboty lingwistycznej, połączonej z gruntowną wiedzą o rzeczach (co zresztą stanowi niezbędny warunek językoznawstwa w ogóle), jest dobrem społecznym o ogromnej doniosłości i dalekim zasięgu czasowym. *Por. Jęz. 2, 1949, s. 11*. W 1814 [roku] ukazał się ostatni,

Wyszukaj w słowniku

Lista haseł

A • Ą • B • C • Ć • D • E • Ę •
F • G • H • I • J • K • L • Ł •
M • N • O • Ó • P • R • S •
Ś • T • U • V • W • X • Y •
Z • Ż • Ź

prezentacja – plik graficzny

Eksploracja i prezentacja: DuCange (ENC)

<http://ducange.enc.sorbonne.fr/>

równoległa prezentacja wersji tekstowej i obrazu

DU CANGE et al., *Glossarium mediæ et infimæ latinitatis*. Niort : L. Favre, 1883-1887.

« 1 **VOCABULUM** » (par les Bénédictins de St. Maur, 1733–1736), dans DU CANGE, et al., *Glossarium mediæ et infimæ latinitatis*, éd. augm., Niort : L. Favre, 1883-1887, t. 8, col. 371c. <http://ducange.enc.sorbonne.fr/VOCABULUM1>

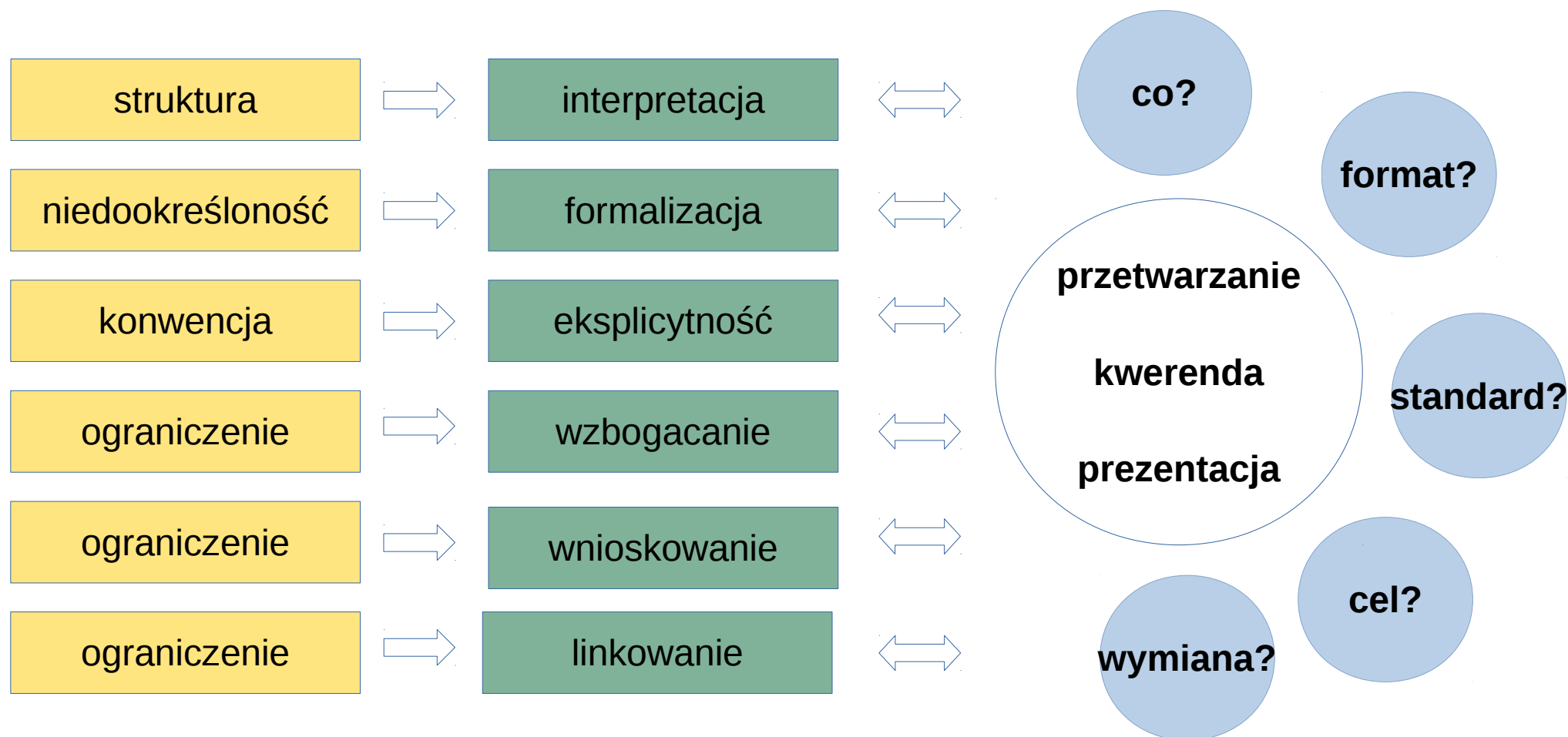
Présentation Publications Documentation Téléchargements Aide

☒ consulter un article ☐ recherche plein texte ☐ formes exactes

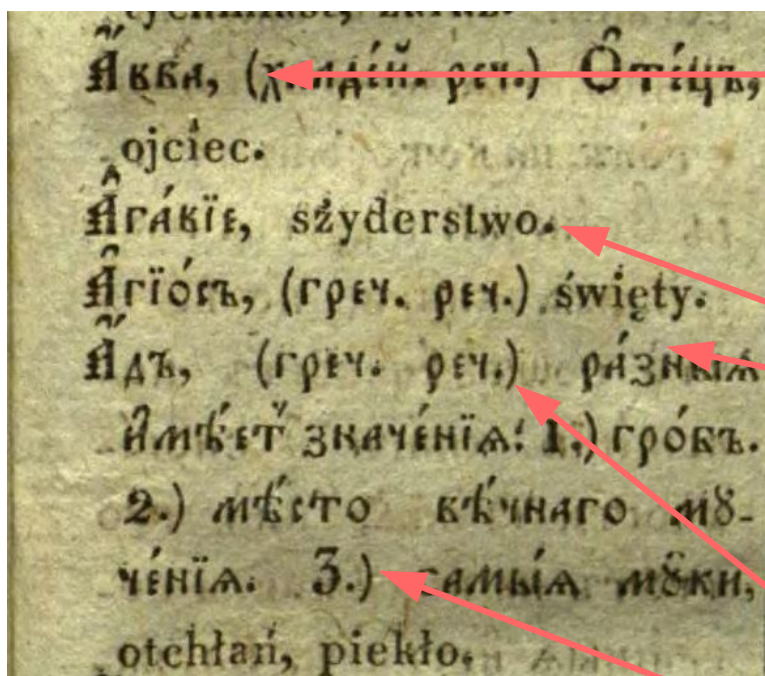
☐ citations latines ☐ citations françaises ☐ citations grecques

1. VOCABULUM, Villa, prædium. Charta ann. 1064. apud Ughellum tom. 1. Ital. sacræ edit. ann. 1717. col. 534 :
Vendo et trado vobis a die presentis quoque modo de terra posita in Vocabulo, quod nuncupatur Caratioli cum ejus fine et mensura. Alia ann. 857. ibid. col. 753 :
Concedimus..... totam jam dictam civitatem,.... ac Vocabula circumquaque posita. Placitum ann. 1105. apud Le Blanc in Dissert. de Monetis pag. 62 :
Non omnis Italia, sed quædam in ea loca sive Vocabula parti B. Petri Ecclesiæ patrimonio fuerunt largita. Chron. Farfense apud Murator. tom. 2. part. 2. col. 535. Item

Anotacja zasobów leksykograficznych: po co?



Anotacja zasobów leksykograficznych: po co?



wyraz
hasłowy

język

etymologia

zagnieżdżenie

definicja

synonimia i in. relacje

wyszukiwanie semazjologiczne

otchłań, piekło

lista lematów

wyszukiwanie zawężone

AKKA

odsyłacze

ISO 639-1:chu

lemmatyzacja

wyszukiwanie derywatów

ukrywanie

mapowanie



derywaty

listy wyrazów

linkowanie słowników

polisemia, hiperonimia

rozwijana lista

menu dostępu

Anotacja zasobów leksykograficznych: po co?

powieść się poch. od powieść₂; ~~czas do-~~
konany; powiedzie się, powiodą się, powiódł
się, powiodła się, powiodło się, powiódłby
się [powiódłby się], powiodłaby się [powiod-
łaby się], powiodłoby się [powiodłoby się],
powiodłyby się [powiodłyby się]; rzecz. po-
wiedzenie się;
„zrealizować czyjeś oczekiwania, marzenia,
plany, stać się takim, jakiego oczekiwano,
zostać osiągniętym, pomyślnie załatwio-
nym”: Brawurowy plan obrabowania banku
i wywiezienia pieniędzy za granicę nie po-
wiódł się i wszyscy trzej trafili do więzienia.
Większości moich kolegów już za pierwszym
razem powiodło się na egzaminach, jedynie
ja musiałem czekać rok na upragniony in-
deks politechniki. *Bliskożnaczne*: udać się,
osiągnąć, dopisać, spełnić oczekiwania,
poszczęścić się, pójść po myśli.

etymologia

formy
zależne i
derywaty

frazeologizmy

synonimy

ekstrakcja relacji

wyszukiwanie synonimów

linki do haseł

formalizacja typów

wyszukiwanie wg typów

odsyłacze

mise-en-page

lemmatyzator

fonetyka

wyszukiwanie wg paradygmatu



korpus



frekwencja

typologia

wyszukiwanie w korpusie

wyszukiwanie wyrażeń

linki do komponentów

mise-en-page

Anotacja zasobów leksykograficznych: po co?

skróty

dymki

rozwiniecie

cytaty

słownik jako korpus

szermować *ndk* 'sprawnie posługiwać się jakąś bronią': Mości książę – ośmielił się ozwać Zaćwilichowski – niechże oni szermują językiem, my mieczem będziemy dalej szermowali. *Sienk. Ogn. I/382*

- ⊕ część mowy
- ⊕ nazwy własne
- ⊕ etymologia

sigla

wyszukiwanie wg miejsca

wyszukiwanie wg czasu

mapy

linie czasu

- ⊕ czas
- ⊕ miejsce
- ⊕ gatunek
- ▶ korpus
- ▶ biblioteki

Format: XML

```
<słownik>
  <hasło>
    <wyraz_hasłowy>
      powieść się
    </wyraz_hasłowy>
    <formy>
      <forma>
        powiedzie się
      </forma>
    </formy>
    <synonim>
      udać się
    </synonim>
  </hasło>
</słownik>
```

- plik tekstowy
- zrozumiały dla człowieka (*human readable*)
- „samo-się-opisujący” (*self-describing*)
- oparty o model drzewa (*tree-like structure*)
- niezależny od platformy (*platform-independent*)
- szeroko wspierany
 - dokumentacja
 - edytory
 - przeszukiwanie i prezentacja
- liczne standardy dla danych lingwistycznych
- „rozwlekły” (*verbose*)
- ciężki (redundancja)
- problem z wyrażaniem relacji niehierarchicznych

(Bardzo) szybkie wprowadzenie do XML

Extensible Markup Language

(pol. *Rozszerzalny Język Znaczników*)

wybór znaczników
nie jest ograniczony

pozwała na łatwe wyodrębnienie
tekstu i anotacji

inne języki znacznikowe

- HTML
- TeX

Kilka zastosowań

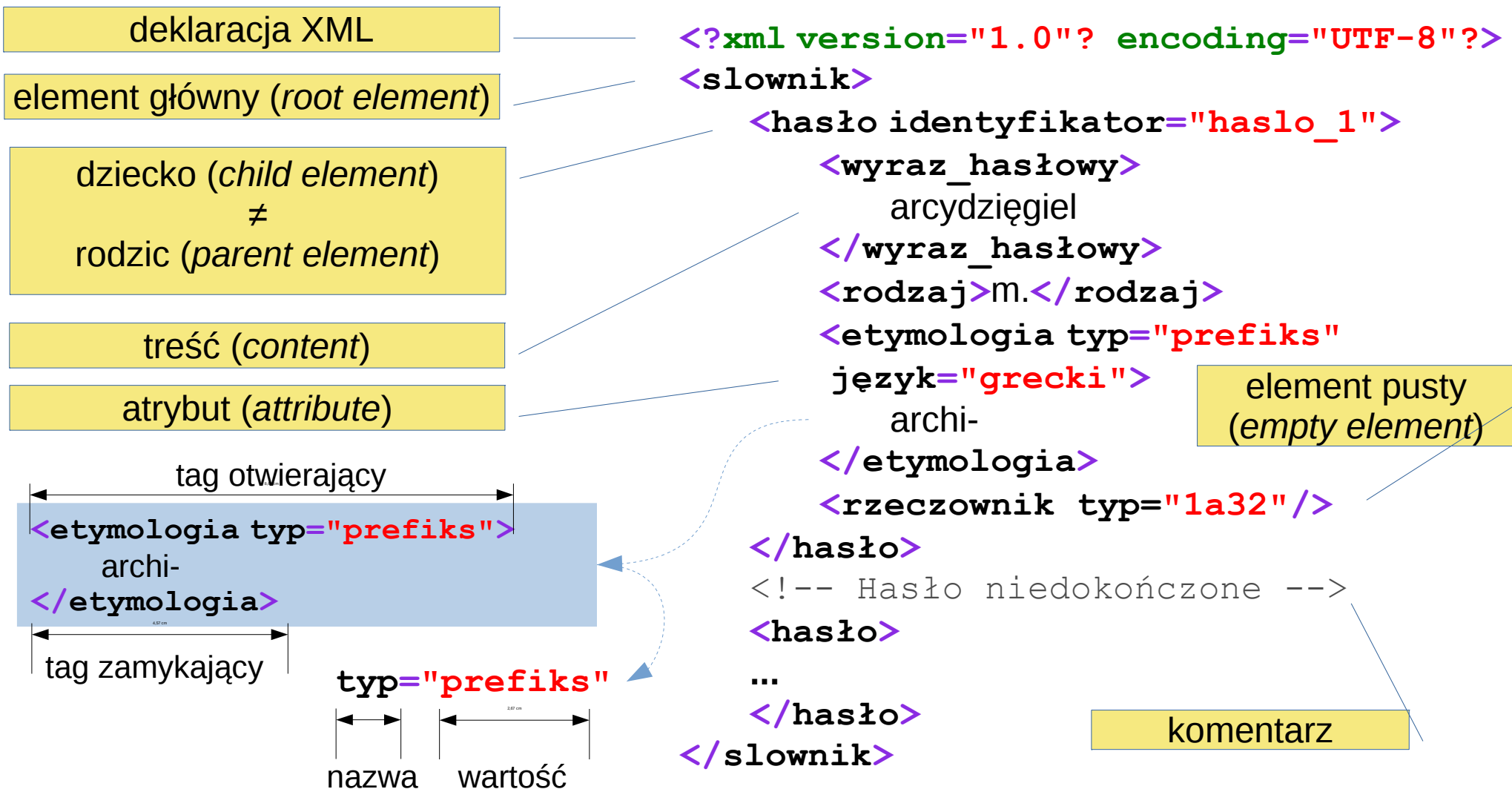
- RSS
- ODT, DOCX
- pliki konfiguracyjne na serwerze
- Systems Biology Markup Language, Chemical Markup Language, bibteXML: XML for BibTeX,

```
<co_chce_o_niej_powiedziec>  
    sekwencja znaków  
</co_chce_o_niej_powiedziec>
```

Przykład pliku HTML

```
<html lang="pl">  
<head>  
    <title>Mój słownik</title>  
</head>  
<body>  
    <h1>Przedmowa</h1>  
    <p>Mój słownik jest bardzo  
dobry.<br /> Ma bardzo dużo  
haseł.</p>  
</body>  
</html>
```

Plik XML: komponenty



Plik XML: poprawność składniowa (*well-formed document*)

musi zawierać
element główny

```
<?xml version="1.0"?>
<slownik>
  <haslo> ... </haslo>
</slownik>
```

```
<?xml version="1.0"?>
<haslo> ... </haslo>
<haslo> ... </haslo>
<haslo> ... </haslo>
```

tagi muszą być
domknięte

```
<czesc>rzecz.</czesc>

<rzeczownik/>
```

```
<czesc>rzecz.
```

prawidłowe
zagnieżdżanie

```
<sens>
  <definicja>
    arcydzięgiel
  </definicja>
</sens>
```

```
<sens>
  <definicja>
    arcydzięgiel
  </sens>
</definicja>
```

wielkość liter
ma znaczenie

```
<sens></sens>
<Sens></Sens>
```

```
<sens></Sens>
<Sens></sens>
```

Plik XML: edytory

Jaki edytor?

- **tekstowy** (Lin: geany, gedit, emacs, vim; Win: notepad, notepad++) **z wtyczkami ułatwiającymi edycję**
- **środowisko programistyczne**
- **specjalistyczny** (darmowe: XML Copy Editor itd. komercyjne: Oxygen XML Editor)

https://en.wikipedia.org/wiki/Comparison_of_XML_editors

Widok drzewa

Podświetlanie składni i indentacja

Edytor schematu, dokumentacji

Automatyczna walidacja

Autouzupełnianie

Edytor graficzny

```
<wyraz_hasłowy>arcydzi  
ęgiel</wyraz_hasłowy><r  
odzaj>m.</rodzaj>
```



```
<wyraz_hasłowy>  
    arcydzięgiel  
</wyraz_hasłowy>  
<rodzaj>  
    m.  
</rodzaj>
```

Laboratorium 1: plik XML

1. W edytorze tekstu utworzyć plik XML: „nazwisko.xml”

2. Tekst hasła

a) skopiować z pliku OCR tekst hasła

b) wkleić tekst do utworzonego pliku XML

c) poprawić

3. Planowanie anotacji

a) wybrać elementy do kodowania (max. 4-5)

b) wybrać nazwy dla elementów i atrybutów

4. Plik XML

a) stworzyć deklarację

b) stworzyć element główny

c) anotować hasło, bacząc na poprawność składniową

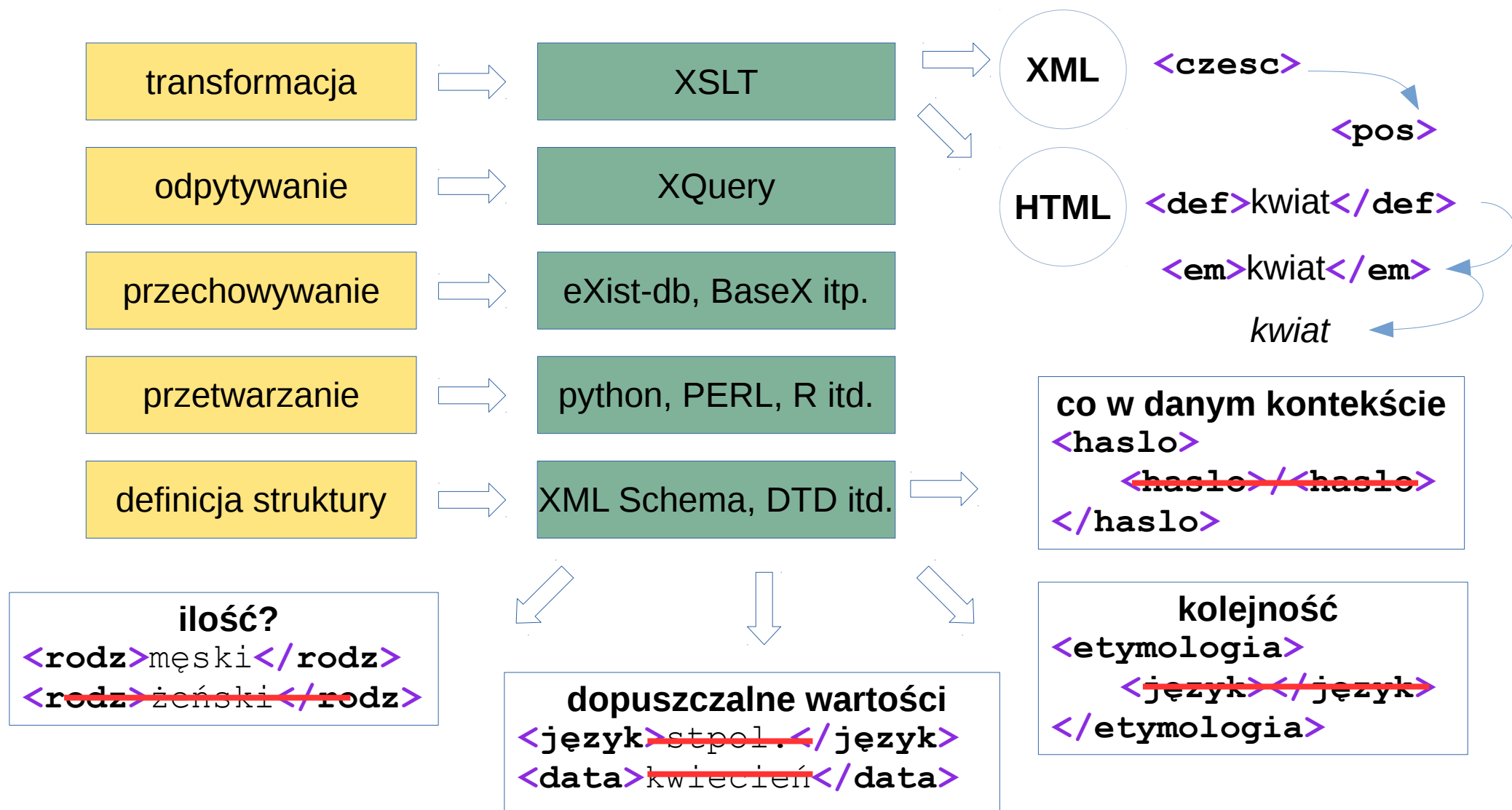
5. Walidacja

a) otworzyć plik XML za pomocą przeglądarki

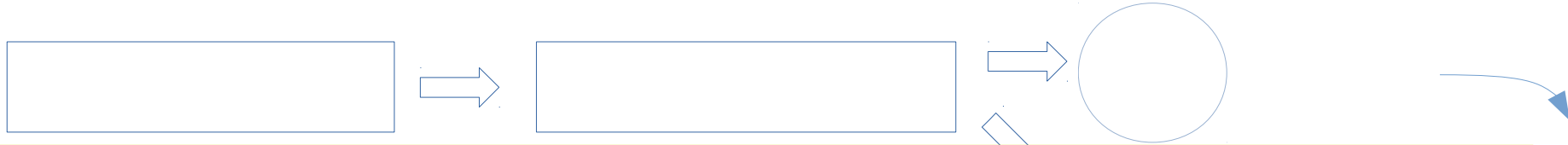
b) sprawdzić poprawność składniową (*well-formed*) za pomocą serwisu on-line

<http://validator.w3.org/check>

XML: co dalej?

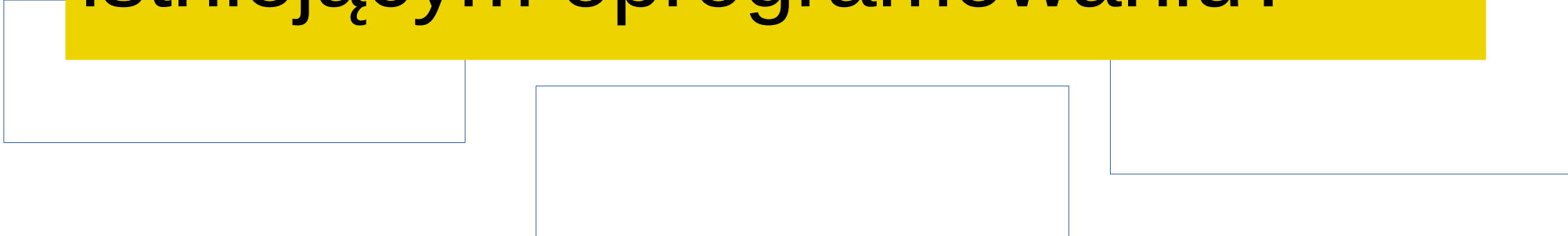


XML: co dalej?

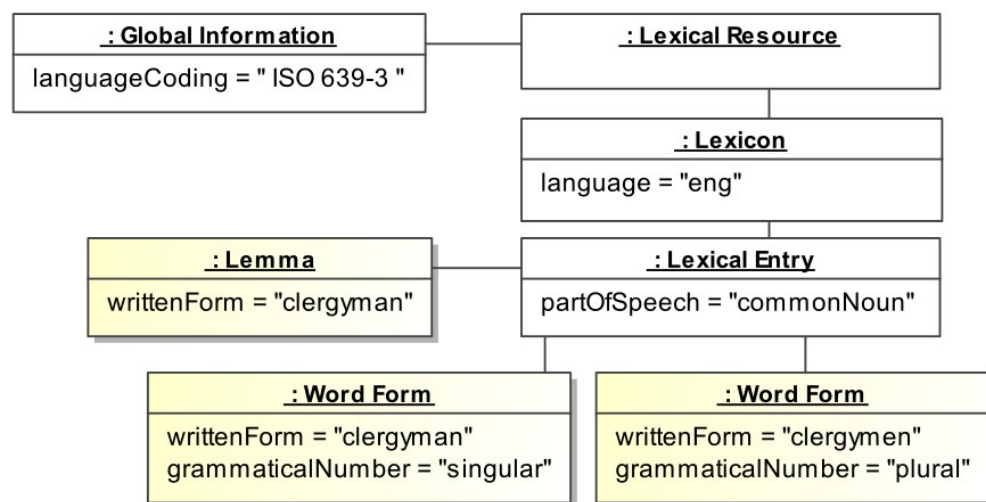


Jakie problemy napotkamy:

- 1) chcąc podzielić się naszymi danymi z innymi badaczami?
- 2) chcąc wykorzystać dane w istniejącym oprogramowaniu?



Standardy: Lexical Markup Format



- od 2008
- standard ISO 24613
- główne zastosowanie: leksykony NLP
- nie ogranicza typów słownika

```
<?xml version="1.0" encoding="UTF-8"?>
<LexicalResource dtdVersion="16">
  <GlobalInformation>
    <feat att="label" val="Simple English LMF test suites"/>
    <feat att="languageCoding" val="ISO 639-3"/>
  </GlobalInformation>
  <Lexicon>
    <feat att="language" val="eng"/>
    <LexicalEntry>
      <feat att="partOfSpeech" val="commonNoun"/>
      <Lemma>
        <feat att="writtenForm" val="clergyman"/>
      </Lemma><WordForm>
        <feat att="writtenForm" val="clergyman"/>
        <feat att="grammaticalNumber" val="singular"/>
      </WordForm>
      <WordForm>
        <feat att="writtenForm" val="clergymen"/>
        <feat att="grammaticalNumber" val="plural"/>
      </WordForm>
    </LexicalEntry>
  </Lexicon>
</LexicalResource>
```

Standardy: TEI XML

Historia TEI

- 1987: P1
- 1994: P3
- 2002: P4 (TEI XML)
- 2007: P5



Co daje TEI?

- ok. 500 tagów dla anotacji niemal wszystkich typów tekstów spotykanych w humanistyce (korpusy, rękopisy, poezja, dramat)

Dlaczego TEI XML?

- długa historia
- liczni użytkownicy
- doskonała dokumentacja
- bogate materiały edukacyjne
- wspierany przez oprogramowanie i narzędzia
- popularny wśród humanistów
- możliwość użycia w innych kontekstach badawczych (korpusy, rękopisy, bibliografie)

TEI XML: Guidelines

<http://www.tei-c.org/release/doc/tei-p5-doc/en/html/>

Front Matter

- i. [Releases of the TEI Guidelines](#)
- ii. [Dedication](#)
- iii. [Preface and Acknowledgments](#)
- iv. [About These Guidelines](#)
- v. [A Gentle Introduction to XML](#)
- vi. [Languages and Character Sets](#)

Back Matter

- Appendix A [Model Classes](#)
- Appendix B [Attribute Classes](#)
- Appendix C
- Appendix C
- Appendix E
- Appendix F
- Appendix G
- Appendix H

Text Body

- 1 [The TEI Infrastructure](#)
- 2 [The TEI Header](#)
- 3 [Elements Available in All TEI Documents](#)
- 4 [Default Text Structure](#)
- 5 [Characters, Glyphs, and Writing Modes](#)
- 6 [Verse](#)
- 7 [Performance Texts](#)
- 8 [Transcriptions of Speech](#)
- 9 [Dictionaries](#)
- 10 [Manuscript Description](#)
- 11 [Representation of Primary Sources](#)
- 12 [Critical Apparatus](#)

TEI sourcecode

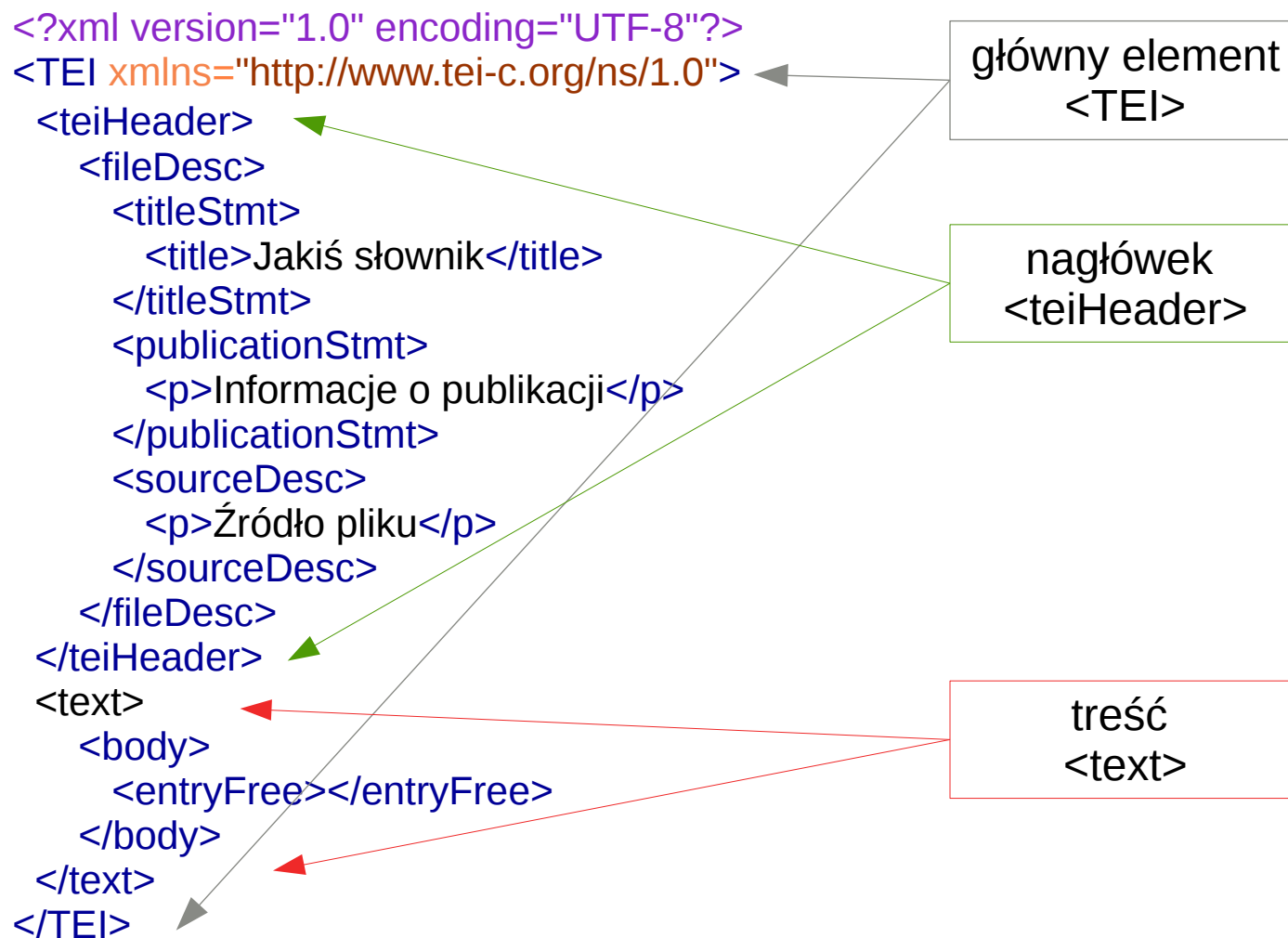
- [Github Repository](#)
- [Bug Reports, Feature Requests, etc.](#)

<entryFree> (unstructured entry) contains a single unstructured entry in any kind of dictionary body. See [Dictionary Body and Overall Structure 9.2 The Structure of Dictionary Entries](#)

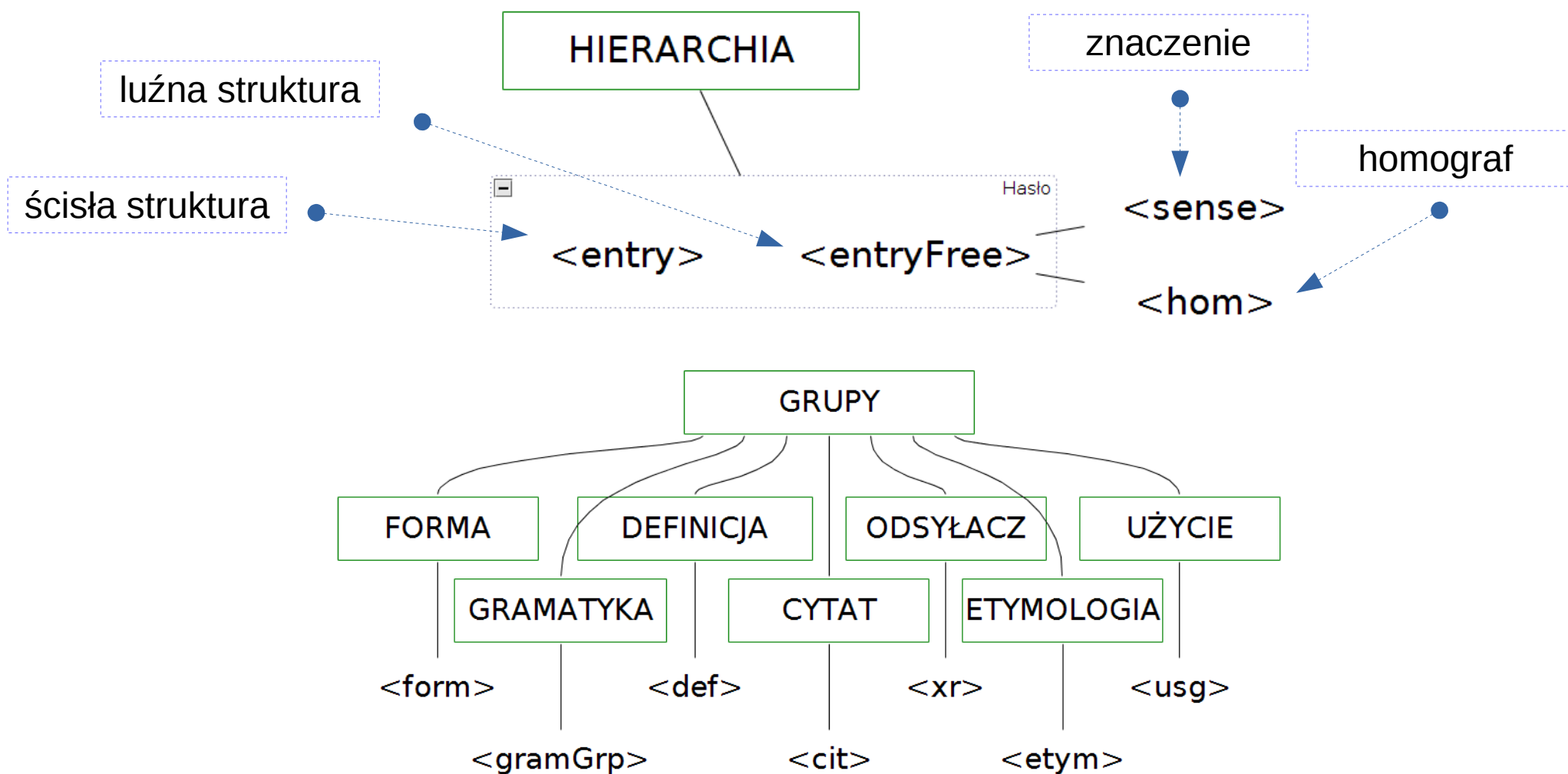
Module	dictionaries — Dictionaries	 13 Names, Dates, People
Attributes	att.global (@xml:id, @n, @xml:lang, @xml:base, @xml:sp (att.global.linking (@corresp, @synch, @sameAs, @copyOf (att.global.facs (@facs)) (att.global.change (@change)) (att.lexicographic (@expand, @norm, @split, @value, @orig @valueDatcat)) att.sortable (@sortKey)	 14 Tables, Formulæ, Graphs  15 Language Corpora
Member of	model.entryLike	
Contained by	drama: castList epilogue performance prologue set figures: figure textstructure: argument body div div1 div2 div3 div4 div5 div6 div7 epigraph postscript	
May contain	analysis: c cl interp interpGrp m pc phr s span spanGrp w certainty: certainty precision respon core: abbr add address bibl biblStruct binaryObject cb choice cit corr date del desc distinct email emph expan foreign gloss graphic hi index label lb list listBibl measure measureGrp media mentioned milestone name note num orig pb quote ref reg sic said sic soCalled stage term time title unclear	

- 9 [Dictionaries](#)
- 10 [Manuscript Description](#)
- 11 [Representation of Primary Sources](#)
- 12 [Critical Apparatus](#)
- 13 [Names, Dates, People, and Places](#)
- 14 [Tables, Formulæ, Graphics and Notated Music](#)
- 15 [Language Corpora](#)

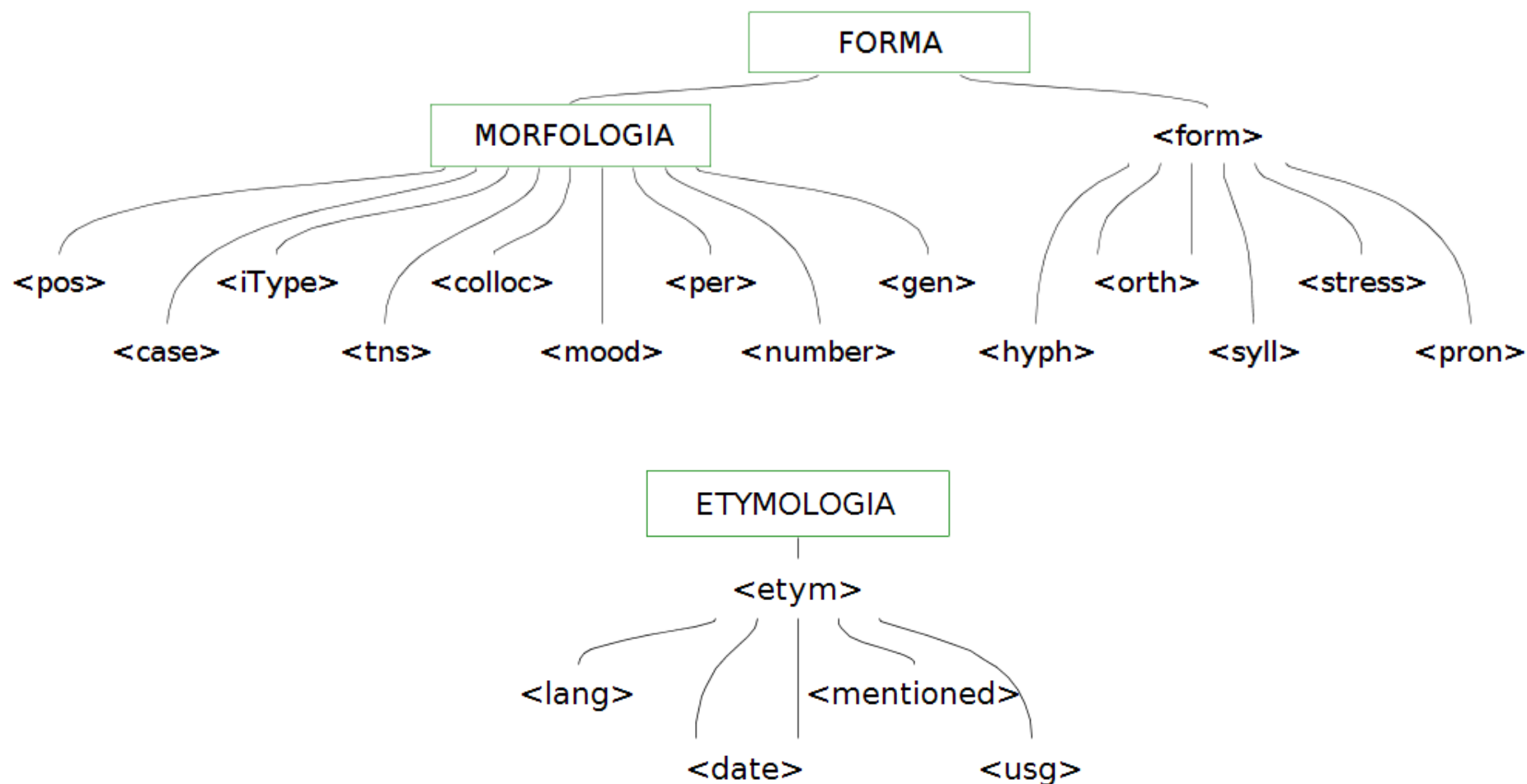
TEI XML: fundamenty



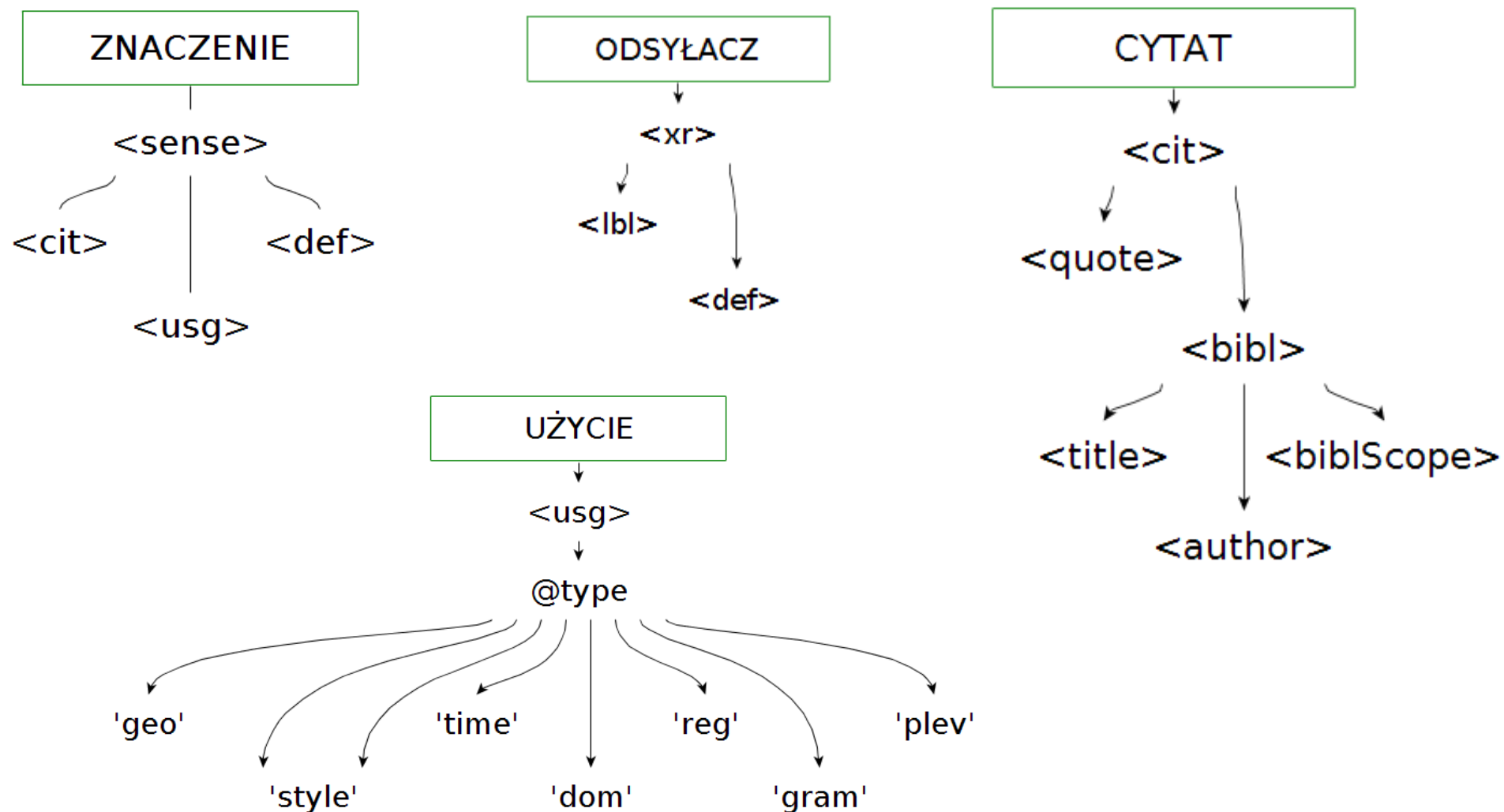
TEI XML: makro- i mikrostruktura



TEI XML: forma, gramatyka



TEI XML: znaczenie, użycie



TEI XML: kilka przydatnych atrybutów

@xml:id



identyfikator

<entryFree xml:id="arcy_191n18">

@n



numer

<sense n="1">
<sense n="a.">

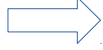
@xml:lang



język

<def xml:lang="pl">

@norm



normalizacja

<usg norm="muzyka">muz.</usg>

@type



typ

<form type="inflected">życia</form>
<entry type="hom">
<cit type="translation">
<cit type="example">

Laboratorium 2: dokument zgodny z TEI

1. Przedstawienie pliku XML w postaci zgodnej z TEI

a) skopiowanie wcześniej utworzonego pliku XML

b) wybór z podanych elementów i atrybutów

c) odszukanie brakujących elementów przy użyciu dokumentacji

2. Dodanie atrybutów

a) @xml:id, @xml:lang, @n, @type

b) skomentowanie wybranych partii

3. Konwersja do innych formatów za pomocą

<http://www.tei-c.org/oxgarage/>

Następne zajęcia

- **data**

- piątek, 4 III 2016, godz. 18

- **temat**

- *Ab ovo*. Metody cyfrowe w pracy leksykografa

- **wymagania**

- lektura (lista na stronie internetowej od 15 stycznia)
<http://scriptores.pl/pracownia/knowak/leksykografia-elektroniczna/>
- pobranie oprogramowania (lub maszyna wirtualna)
 - na stronie internetowej od 15 lutego
- zapisanie do serwisów on-line
 - na stronie internetowej od 15 lutego

Bibliografia

<http://www.danvk.org/2015/01/09/extracting-text-from-an-image-using-ocropus.html>