Koordinované získavanie a extrakcia dát z webových portálov cez spolupracujúce rozšírenia webových prehliadačov

Ústav informatiky PF UPJŠ

5. decembra 2018

Autor: Bc. Matej Perejda

Vedúci práce: RNDr. Peter Gurský, PhD.

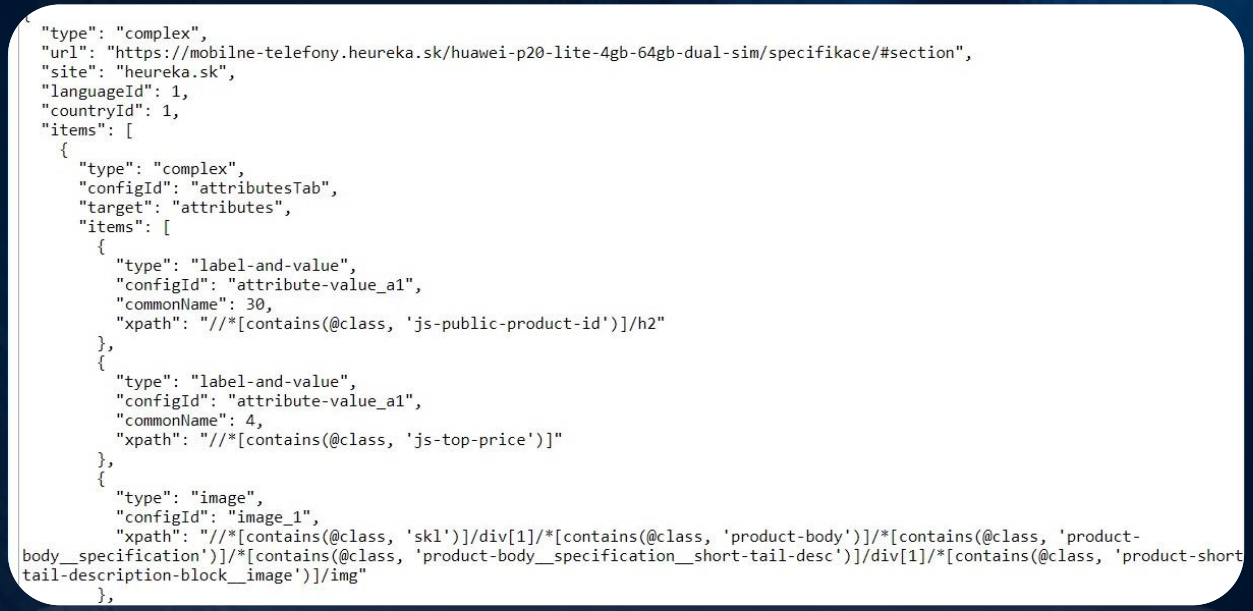
Kapsa

- katalóg produktov s anotáciou
- vznik na ÚINF UPJŠ KE
- vytváranie katalógu produktov ponúkaných e-shopmi
- porovnávanie produktov podľa rôznych vlastností, recenzií používateľov a ponuky predajcov

Exago

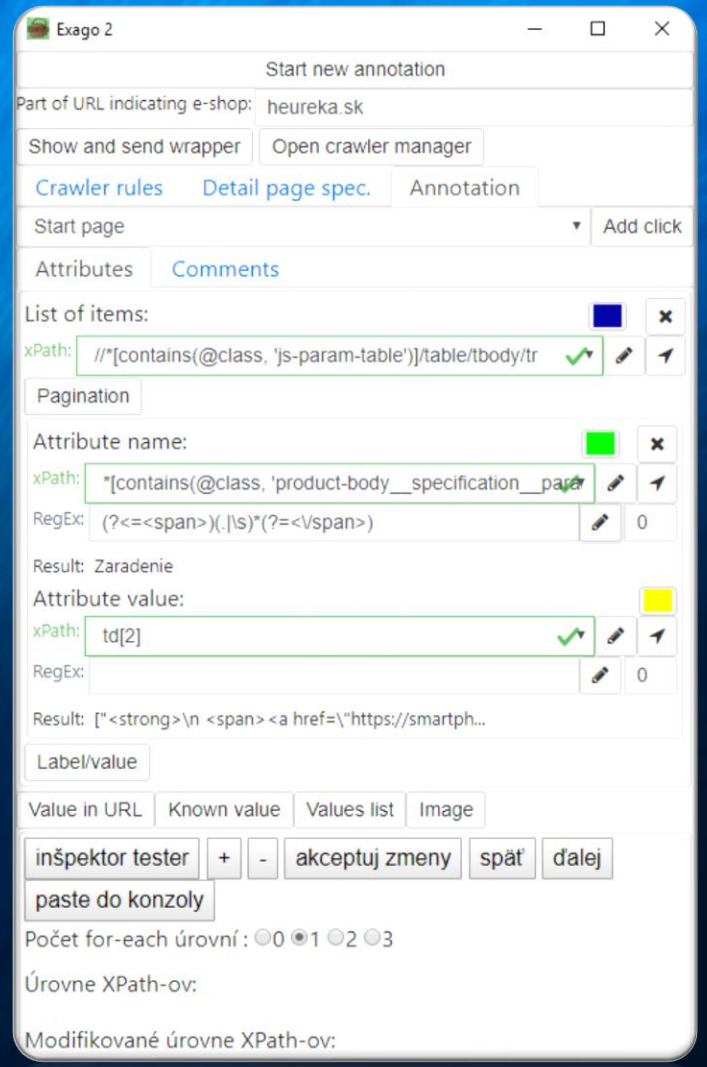
- zásuvný modul, rozšírenie do webových prehliadačov,
- interaktívna anotácia webových produktových katalógov,
- automatické extrahovanie atribútov z popisov produktov,
- zaslanie výslednej anotácie na server (wrapper),
- editovanie existujúcej anotácie zo servera,
- generovanie XPath a regex **(nové)**,
- výnimočnosť: označovanie viacerých atribútov naraz.

Exago – interaktívna anotácia



```
{
  "type": "complex",
  "url": "https://mobilne-telefony.heureka.sk/huawei-p20-lite-4gb-64gb-dual-sim/specifikace/#section",
  "site": "heureka.sk",
  "languageId": 1,
  "countryId": 1,
  "items": [
    {
      "type": "complex",
      "configId": "attributesTab",
      "target": "attributes",
      "items": [
        {
          "type": "label-and-value",
          "configId": "attribute-value_a1",
          "commonName": 30,
          "xpath": "//*[contains(@class, 'js-public-product-id')]/h2"
        },
        {
          "type": "label-and-value",
          "configId": "attribute-value_a1",
          "commonName": 4,
          "xpath": "//*[contains(@class, 'js-top-price')]"
        },
        {
          "type": "image",
          "configId": "image_1",
          "xpath": "//*[contains(@class, 'skl')]/div[1]/*[contains(@class, 'product-body')]/*[contains(@class, 'product-
body__specification')]/*[contains(@class, 'product-body__specification__short-tail-desc')]/div[1]/*[contains(@class, 'product-short
tail-description-block__image')]/img"
        },
```

Exago – wrapper



Exago - GUI

Profesijná motivácia

- distribúcia úloh medzi viaceré stroje (odľahčenie servera),
- využitie JavaScript-u (Java nie je dobrá voľba),
- „viac strojov, viac IP adries“ (nepôsobiť ako zlodej dát).

Ciele diplomovej práce

1. **Porovnanie** súčasných spôsobov extrakcie dát z webových portálov najmä z hľadiska schopnosti extrahovať dáta z dynamicky vytváraných webových stránok cez AJAX volania a schopnosti distribúcie procesu prehľadávania a extrakcie.
2. **Obohatenie** existujúceho rozšírenia webového prehliadača na anotáciu webových stránok o schopnosť prehľadávania a extrakcie dát z webu aj pre dynamické webové stránky simuláciou správania používateľa.
3. **Návrh a vytvorenie** škálovateľného servera koordinujúceho spoluprácu viacerých inštancií vytvoreného rozšírenia webového prehliadača z cieľa 2.
4. **Otestovanie** korektnosti a škálovateľnosti vytvoreného riešenia extrakciou reálnych webových portálov.

Postup práce

- vytvorenie prehľadu webových scraperov,
- pochopenie Exaga,
- článok - ITAT 2018 - Framework for Distributed Computing on the Web (Šiller, Kuchař),
- vytvorenie extraktora dát z webových stránok v Exagu,
- rozšírenie Exaga o prechádzanie webovým portálom a hľadanie stránok na extrakciu,
- návrh škálovateľného servera na koordináciu úloh extrakcie,
- implementácia a nasadenie servera,
- koordinácia viacerých klientov prostredníctvom servera,
- testovanie.

Vytvorenie prehľadu webových scraperov

- typ (webová/desktopová aplikácia, rozšírenie, service, framework, ...)
- licencia
- open-source
- interaktívna anotácia elementov webstránky
- automatická extrakcia
- relevantnosť dát
- export dát (API, .CSV, .JSON, .XLSX, ...)
- podpora dynamicky načítavaných dát
  - AJAX, JavaScript
  - infinite scrolling
  - iné

Vytvorenie prehľadu webových scraperov (2)

Č.	Názov	Typ	Licencia	Open-source	Interaktívna anotácia	Automatická extrakcia	Export	Dynamicky načítavané dáta	API
1	Agenty	chrome rozšírenie, web app (cloud)	platená (14 dní trial)	×	✓	×	CSV, FTP, JSON, TSV, XML	✓	✓
2	Apify	web app	bezplatná, platená	×	×	×	CSV, HTML, JSON, JSONL, RSS, XML	✓	✓
3	Connotate	desktop app, web app	platená	×	✓	×	CSV, databáza, email, HTML, XLS, XML	✓	?
4	Content Grabber	desktop app	platená (30 dní trial) na žiadosť	×	✓	×	CSV, DOCX, FTP, JSON, MySQL, Oracle, PDF, SQL Server, XLS, XML	✓	✓
5	CrawlMonster	web app	bezplatná, platená	×	×	×	email	?	?
6	Data Miner	chrome rozšírenie	bezplatná, platená	×	✓	×	CSV, TSV, XLS, XLSX	✓	×
7	DataScrapper	firefox, chrome rozšírenie	bezplatná	✓	×	×	CSV	×	×
8	Data Toolbar	IE, firefox, chrome rozšírenie	bezplatná, platená	×	✓	×	CSV, HTML, SQL, XLS, XML	✓	×
9	Dexi.io	web app (SaaS)	platená (60 minút trial)	×	✓	×	Amazon S3, CSV, FTP, Google Drive, Google Docs, JSON, SCSV, SFTP, XLS, XLSX, XML	✓	✓
10	Diffbot	web app, knižnice	platená (14 dní trial)	×	×	✓	CSV, JSON	✓	✓
11	Diggernaut (Excavator - visual extractor)	desktop app, chrome rozšírenie, web app (cloud)	bezplatná, platená	×	✓	×	CSV, JSON, XLS	×	✓
12	Easy Web Extract	desktop app	platená (14 dní trial)	×	✓	×	CSV, HTML, SQL server, XML	✓	×
13	Embed.ly	web app	platená (30 dní trial)	×	×	✓	JSON	-	✓
14	Expired Domain Scraper	chrome rozšírenie	bezplatná	×	×	✓	-	podpora iba Youtube, Google, Bing, Yandex	×
15	Fminer	desktop app	platená (15 dní trial)	×	✓	×	CSV, FTP, HTML, JSON, MS SQL, MySQL, Oracle, SQL, SQLite, XLS, XML	✓	×
16	GetData.IO	chrome rozšírenie	bezplatná, platená	×	✓	×	CSV, JSON	×(iba pomocou API)	✓
17	Grepsr	chrome rozšírenie, web app	bezplatná, platená	×	✓	×	CSV, Dropbox, FTP, Google Docs, JSON, RSS, XLSX, Web hooks	✓	✓
18	Handy Web Extractor	desktop app	bezplatná	×	×	×	×	×	×
19	Helium Scraper	desktop app	platená (10 dní trial)	×	✓	×	CSV, MySQL, XML	✓	✓
20	iMacros	IE, firefox, chrome rozšírenie, desktop app	platená (30 dní trial)	×	✓	×	CSV, databáza, TXT, XML	✓	✓

Tabuľka 1.1: Prehľad vlastností existujúcich nástrojov

Vytvorenie prehľadu webových scraperov (3)

Č.	Názov	Typ	Licencia	Open-source	Interaktívna anotácia	Automatická extrakcia	Export	Dynamicky načítavané dáta	API
21	Import.io	web app (cloud)	platená (7 dní trial)	×	✓	×	CSV, JSON, XLSX	✓	✓
22	Instant Data Scraper	chrome rozšírenie	bezplatná	×	×	✓	CSV, XLSX	✓	×
23	KantuX	desktop app	bezplatná, platená	×	✓(OCR)	×	CSV	✓	✓
24	Kido Scraper Generator	chrome rozšírenie	bezplatná	?	✓	×	-	-	×
25	Morph.io	scrapovacia platforma	platená	✓	-	-	CSV	-	✓
26	Mozenda	web app (cloud), desktop app (agent builder)	platená (30 dní trial)	×	✓	×	CSV, JSON, TSV, XLSX, XML	✓	✓
27	myTrama	chrome rozšírenie, web app	platená (trial)	×	✓	×	CSV, HTML, JSON, PDF, XML	-	✓
28	Octoparse	desktop app, cloud	bezplatná, platená	×	✓	×	CVS, HTML, MySQL, Oracle, SQL, TXT, XLS	✓	✓
29	OutWit Hub	desktop app, firefox rozšírenie	bezplatná, platená	×	×	✓	CSV, FTP, HTML, JSON, SQL, TXT, XLS	✓	×
30	ParseHub	desktop app, web app (cloud)	bezplatná, platená	×	✓	×	CSV, Dropbox, Google Sheets, JSON	✓	✓
31	PhantomJS	WebKit, cloud	bezplatná, platená	✓	-	-	-	-	✓
32	QuickCode (ScraperWiki)	web app (IDE)	?	✓	-	-	-	-	✓
33	Rank Scraper	chrome rozšírenie	bezplatná	×	?	?	?	?	×
34	Regex Scraper	chrome rozšírenie	bezplatná	?	×	×	HTML	?	×
35	RegexSearch	firefox rozšírenie	bezplatná	✓	×	×	?	?	×
36	ScrapBook	firefox rozšírenie	bezplatná	×	×	×	-	-	×
37	Scrape.it	chrome rozšírenie, web app (cloud)	platená (7 dní trial)	?	✓	×	?	×	✓
38	ScrapeHero	service	platená	×	?	?	Amazon S3, Cassandra, CSV, databáza, Dropbox, DynamoDB, FTP, Hadoop, Hbase, JSON, MongoDB, MySQL, Oracle, XML	✓	✓
39	Scraper	chrome rozšírenie	bezplatná	?	×(iba XPath referencie)	×	Google Docs	?	×
40	Scraper Crawler	chrome rozšírenie	bezplatná, platená	×	×	✓	?	?	×

Tabuľka 1.2: Prehľad vlastností existujúcich nástrojov

Vytvorenie prehľadu webových scraperov (4)

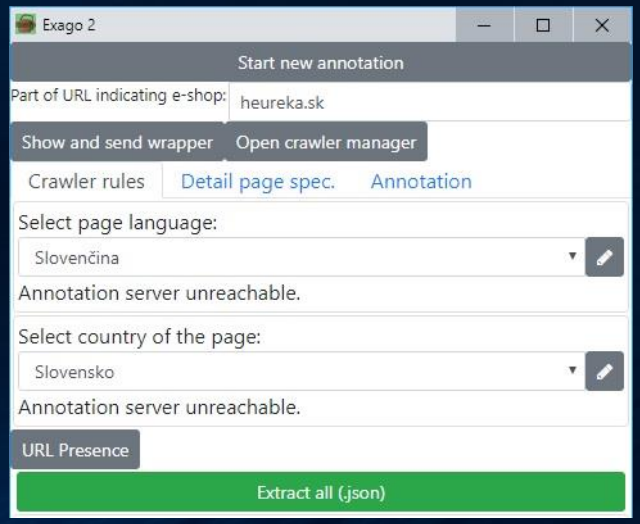
Č.	Názov	Typ	Licencia	Open-source	Interaktívna anotácia	Automatická extrakcia	Export	Dynamicky načítavané dáta	API
41	Scrapinghub (Portia)	web app	bezplatná	✓	✓	×	CSV, JSON, TXT, XML	✓	×
42	Scrapy	framework	bezplatná	✓	-	-	CSV, FTP, JSON, XML	-	✓
43	Screen Scraper	desktop app, chrome rozšírenie	bezplatná, platená (30 dní trial - vyžaduje kreditnú kartu)	×	×	×	CSV, databáza, HTML, JSON, MySQL, SQL, TXT, XML	✓	✓
44	UIPath	desktop app (Studio), chrome, firefox rozšírenie, web app (Orchestrator)	platená (60 dní trial) na žiadosť	×	✓	×	CSV, email, XLS	✓	✓
45	uScraper	web app	bezplatná, platená	-	×	✓	CSV	-	×
46	Visual Web Ripper	desktop app	platená (14 dní trial)	×	✓	×	CSV, JSON, MySQL, OleDB, Oracle, PDF, SQL, SQLite, XLS, XML	✓	✓
47	Web Content Extractor	desktop app	platená (14 dní trial)	×	✓	×	CSV, FTP, HTML, HTTP, MS Access, MySQL, ODBC, SQL, TXT, XLS, XML	✓	×
48	Web Data Extractor (Pro)	desktop app	platená (15 dní trial)	×	✓	×	CSV, TXT, XLSX	×	×
49	WebHarvy Web Scraper	desktop app	platená (15 dní trial)	×	✓	×	CSV, JSON, MySQL, MS SQL, Oracle, TSV, XLS, XML	✓	×
50	Webhose.io	web app	bezplatná, platená	×	×	×	JSON, RSS, XLS, XML	-	✓
51	Web Robots Scraper	chrome rozšírenie (scraping IDE)	bezplatná	×	×	×	CSV, database, server, XLSX	-	×
52	Web Scraper	chrome rozšírenie, web app (cloud)	bezplatná (rozšírenie), platená (cloud)	✓	✓	×	CouchDB, CSV, Dropbox	✓	×
53	WebSundew	desktop app	platená (15 dní trial)	×	✓	×	CSV, email, FTP, MySQL, Oracle, RSS, SQL Server, XLS, XML	✓	✓
54	WinAutomation	desktop app	platená (30 dní trial)	×	✓	×	CSV, FTP, XLS	✓	✓
55	80legs	web app	bezplatná, platená	×	×	×	CSV, JSON	-	✓

Tabuľka 1.3: Prehľad vlastností existujúcich nástrojov

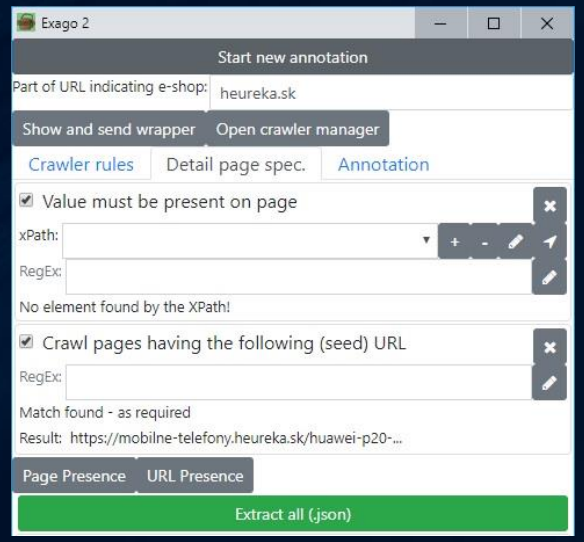
Postup práce

- vytvorenie prehľadu webových scraperov,
- **pochopenie Exaga**,
- **článok - ITAT 2018 - Framework for Distributed Computing on the Web (Šiller, Kuchař)**,
- **vytvorenie extraktora dát z webových stránok v Exagu**,
- rozšírenie Exaga o prechádzanie webovým portálom a hľadanie stránok na extrakciu,
- návrh škálovateľného servera na koordináciu úloh extrakcie,
- implementácia a nasadenie servera,
- koordinácia viacerých klientov prostredníctvom servera,
- testovanie.

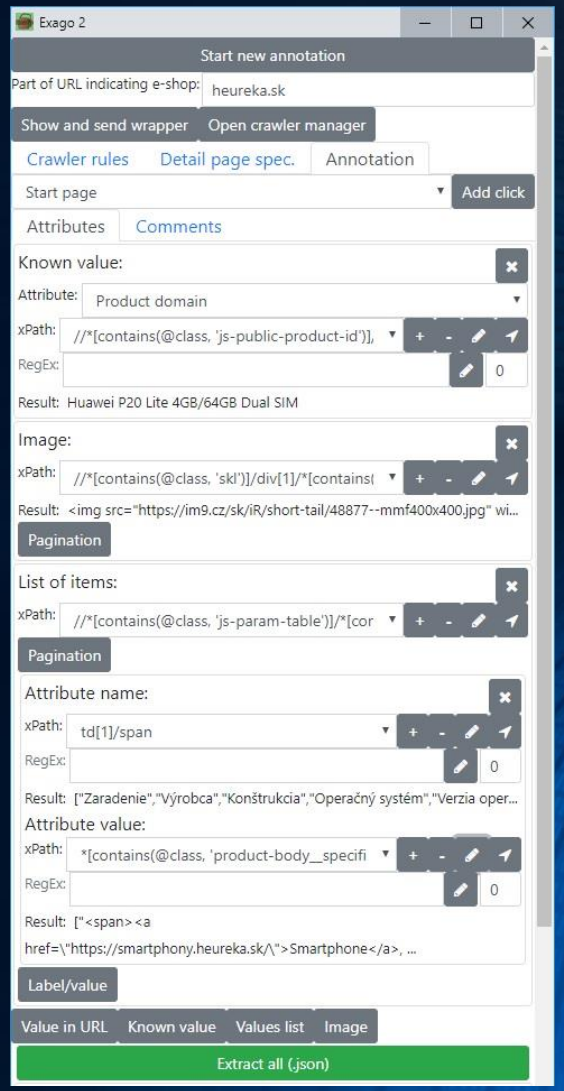
Exago - tvorba extraktora dát



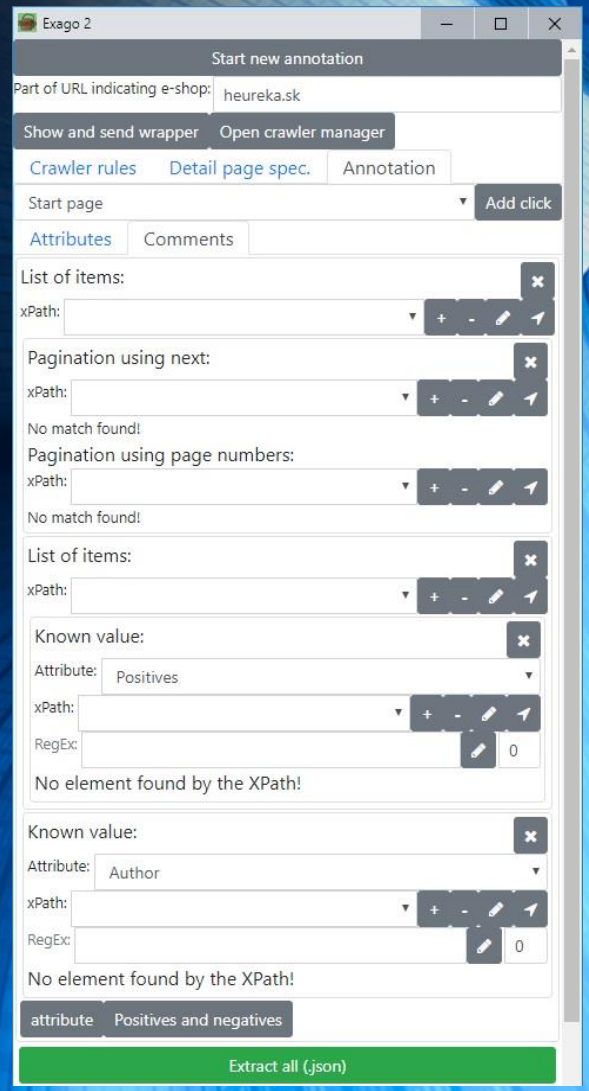
Crawler rules



Detail page specification

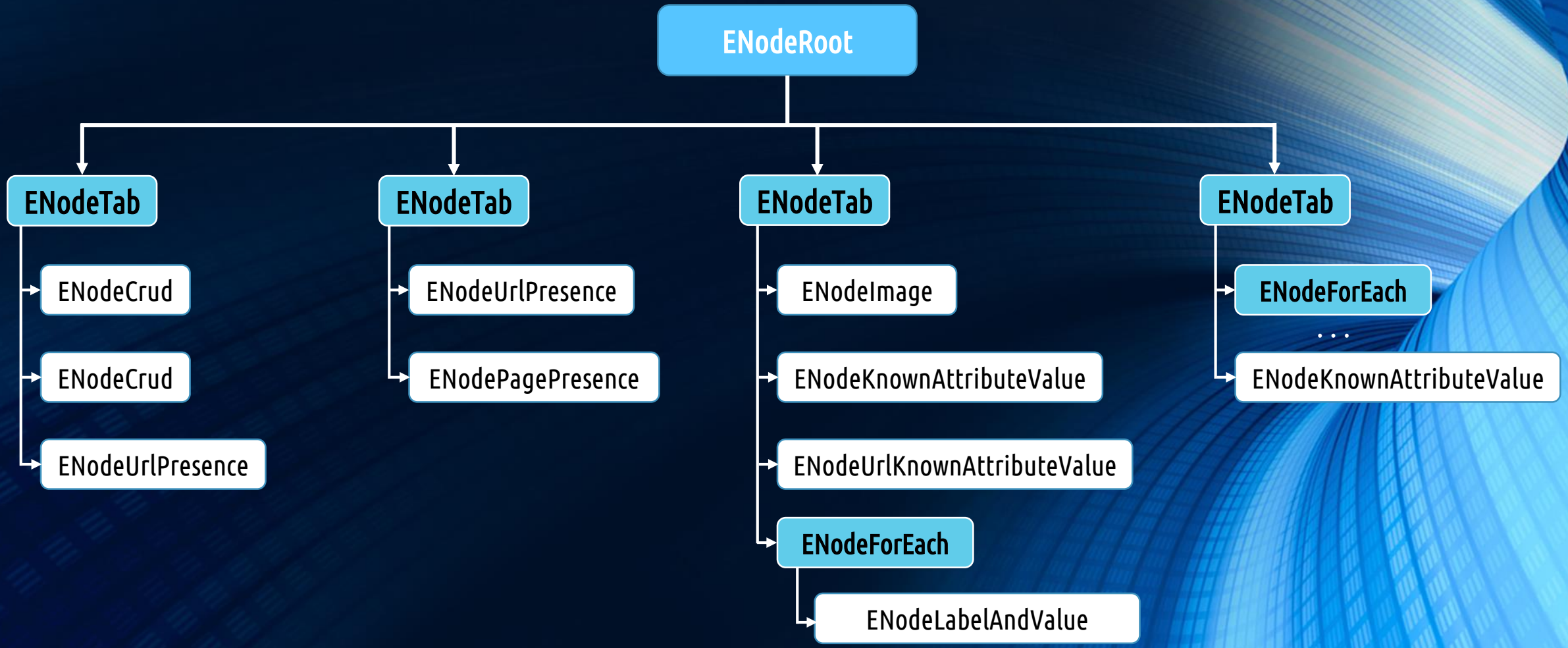


Annotation - attributes

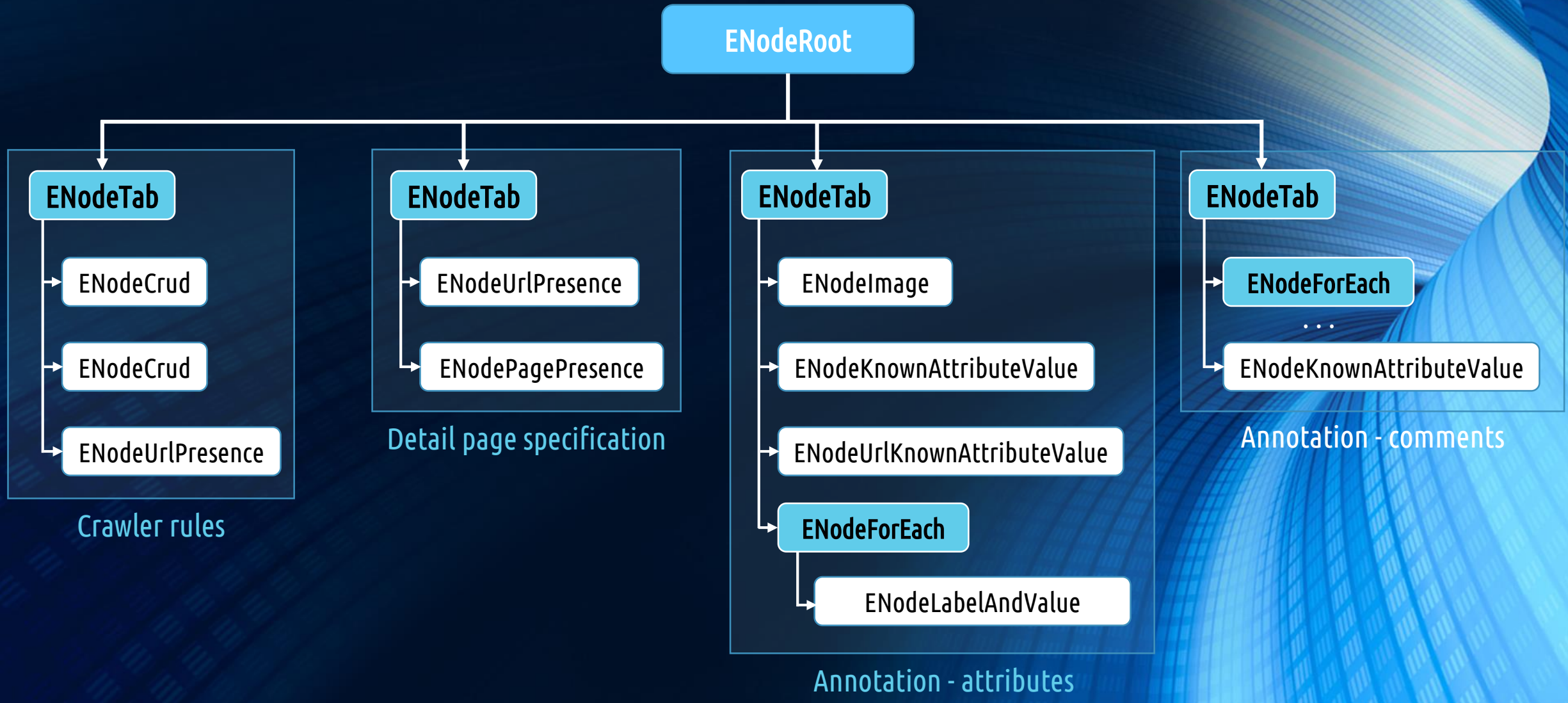


Annotation - comments

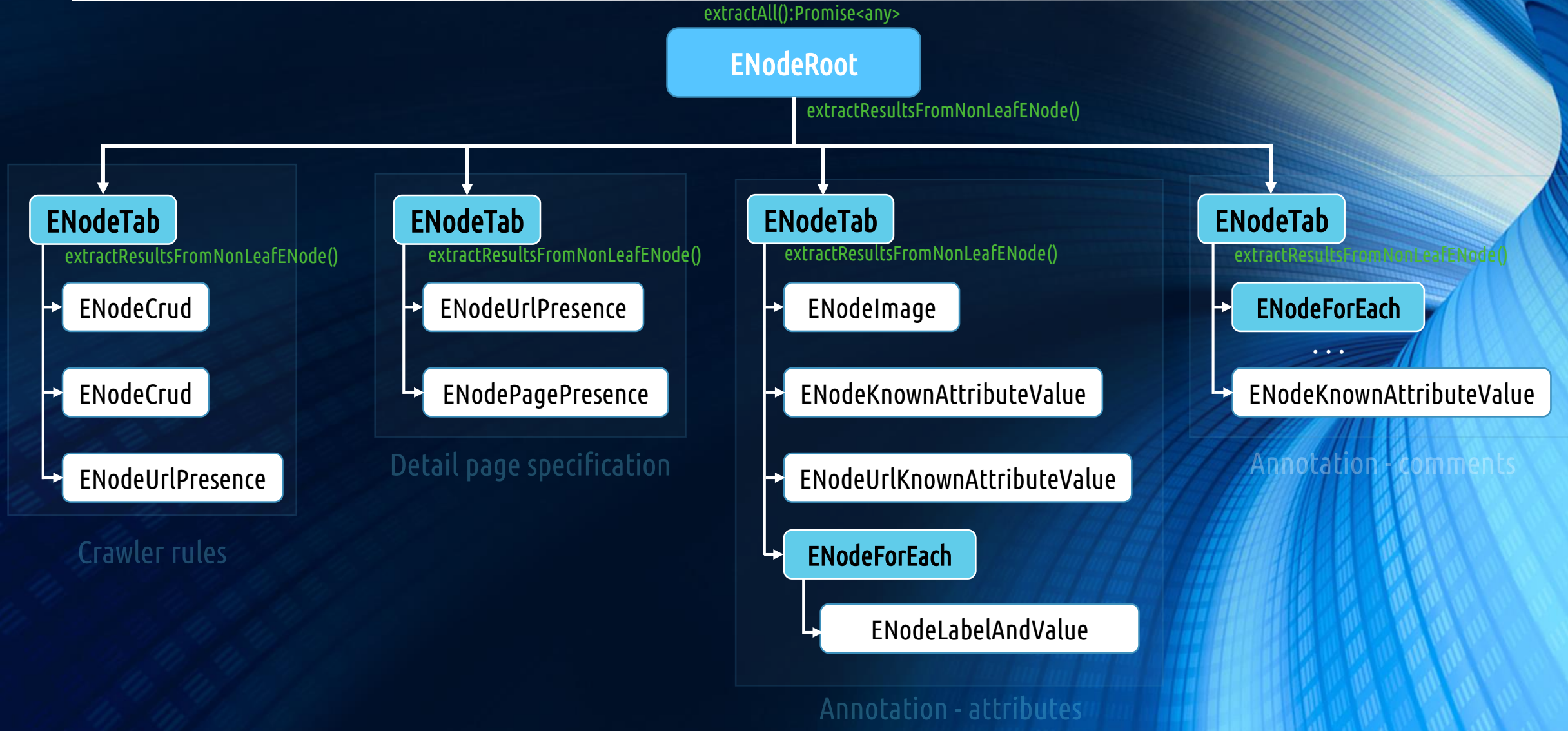
Exago - tvorba extraktora dát



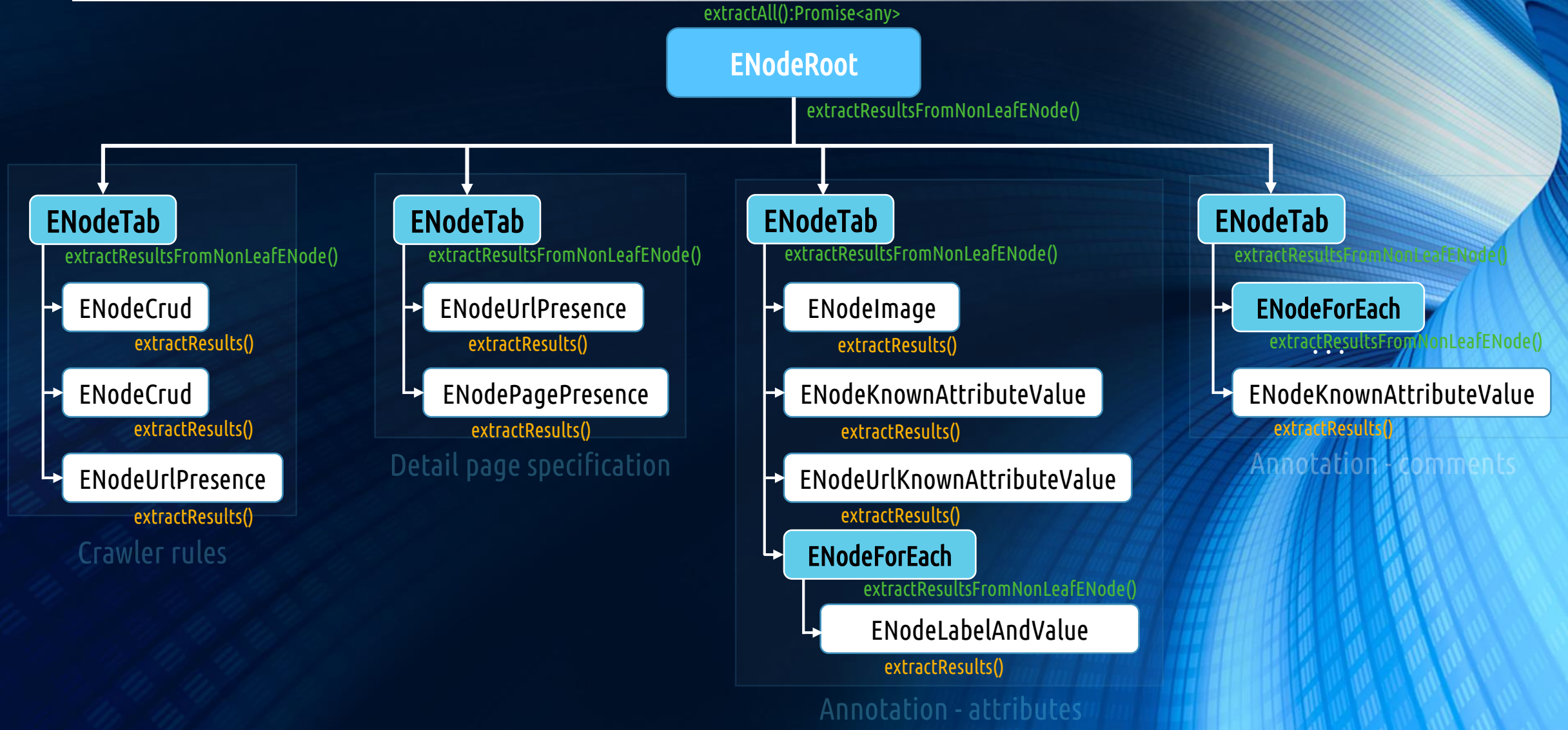
Exago - tvorba extraktora dát



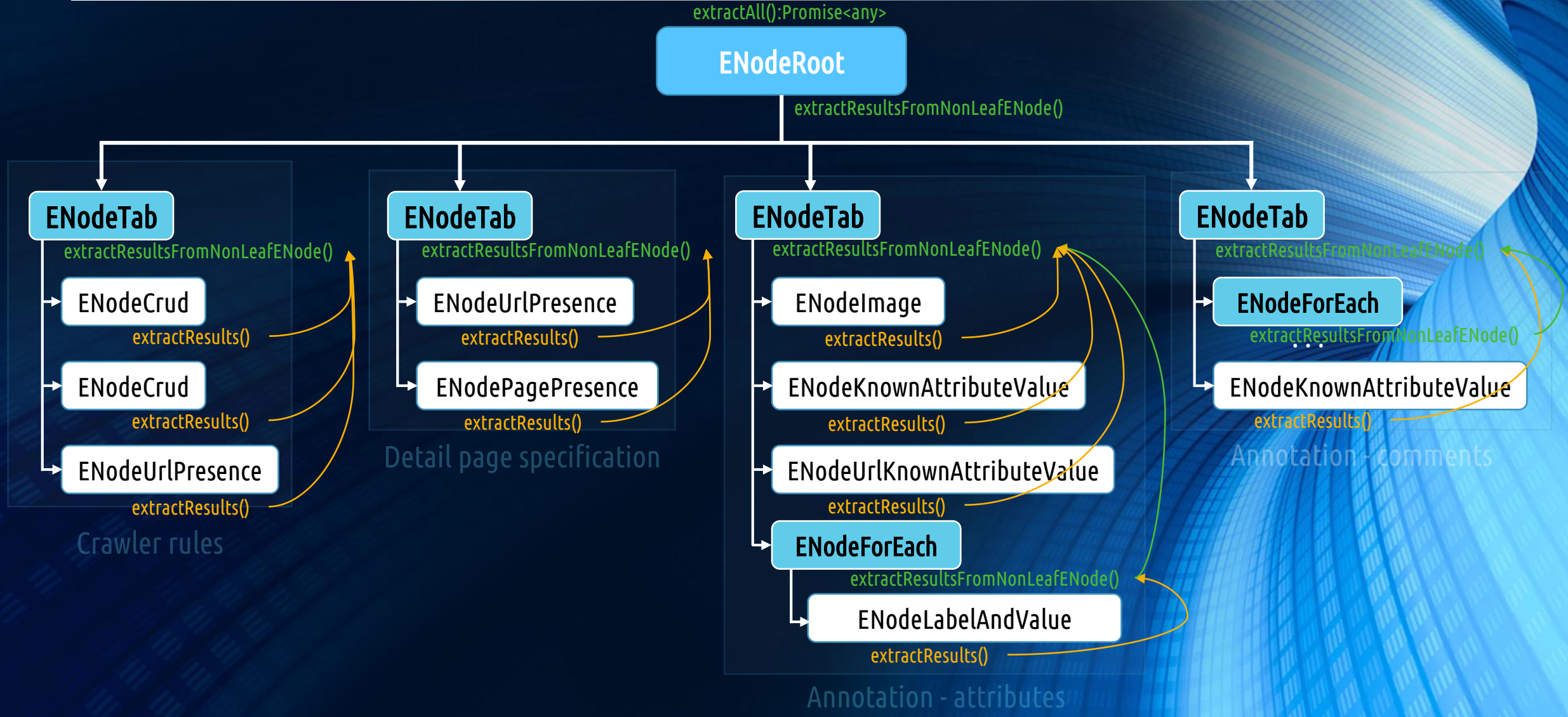
Exago - tvorba extraktora dát



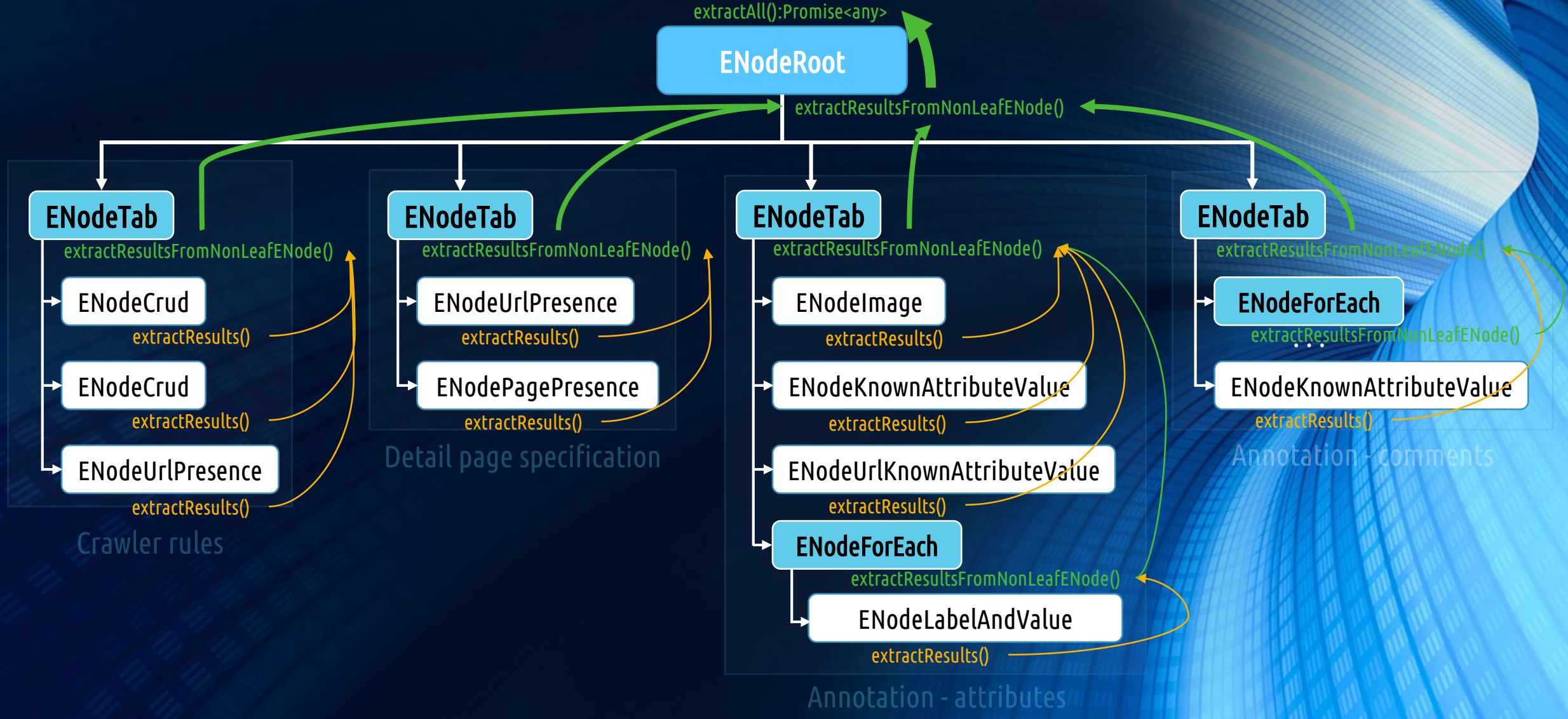
Exago - tvorba extraktora dát



Exago - tvorba extraktora dát



Exago - tvorba extraktora dát



Exago - tvorba extraktora dát

Postup práce

- vytvorenie prehľadu webových scraperov,
- pochopenie Exaga,
- článok - ITAT 2018 - Framework for Distributed Computing on the Web (Šiller, Kuchař),
- vytvorenie extraktora dát z webových stránok v Exagu,
- rozšírenie Exaga o prechádzanie webovým portálom a hľadanie stránok na extrakciu,
- návrh škálovateľného servera na koordináciu úloh extrakcie,
- implementácia a nasadenie servera,
- koordinácia viacerých klientov prostredníctvom servera,
- testovanie.

Zdroje a literatúra

Prehľad webových scraperov: google.com, github.com, obchod Google Chrome, obchod Firefox Add-ons, oficiálne stránky nástrojov, scraping.pro, hongkiat.com/blog/web-scraping-tools/

[1] Liu, Bing: *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data.* Second Edition, ISBN 978-3-642-19459-7, Springer, 2011

[2] Kushmerick, N.*: Wrapper induction: efficiency and expressiveness*. Artificial Intelligence, 118:15-68, 2000.

[3] Muslea, I., Minton, S. and Knoblock, C.: *A hierarchical approach to wrapper induction*. Agents-99, 1999.

[4] Cohen, W., Hurst, M., and Jensen, L.: *A flexible learning system for wrapping tables and lists in HTML documents*. WWW-2002, 2002.

[5] Hsu, C.N., Dung, M.T.: *Generating finite-state transducers for semistructured data extraction from the Web.* Information Systems. 23(8): 521-538, 1998.

[6] Chabaľ, V: *Poloautomatická extrakcia komentárov z produktových katalógov*. Diplomová práca. Košice 2014

[7] Crescenzi, V., Mecca, G., Merialdo,P.: Roadrunner*: Towards automatic data extraction from large web sites.* In Proceedings of VLDB 2001, pp. 109-118.

Ďakujem za pozornosť!

Otázky?