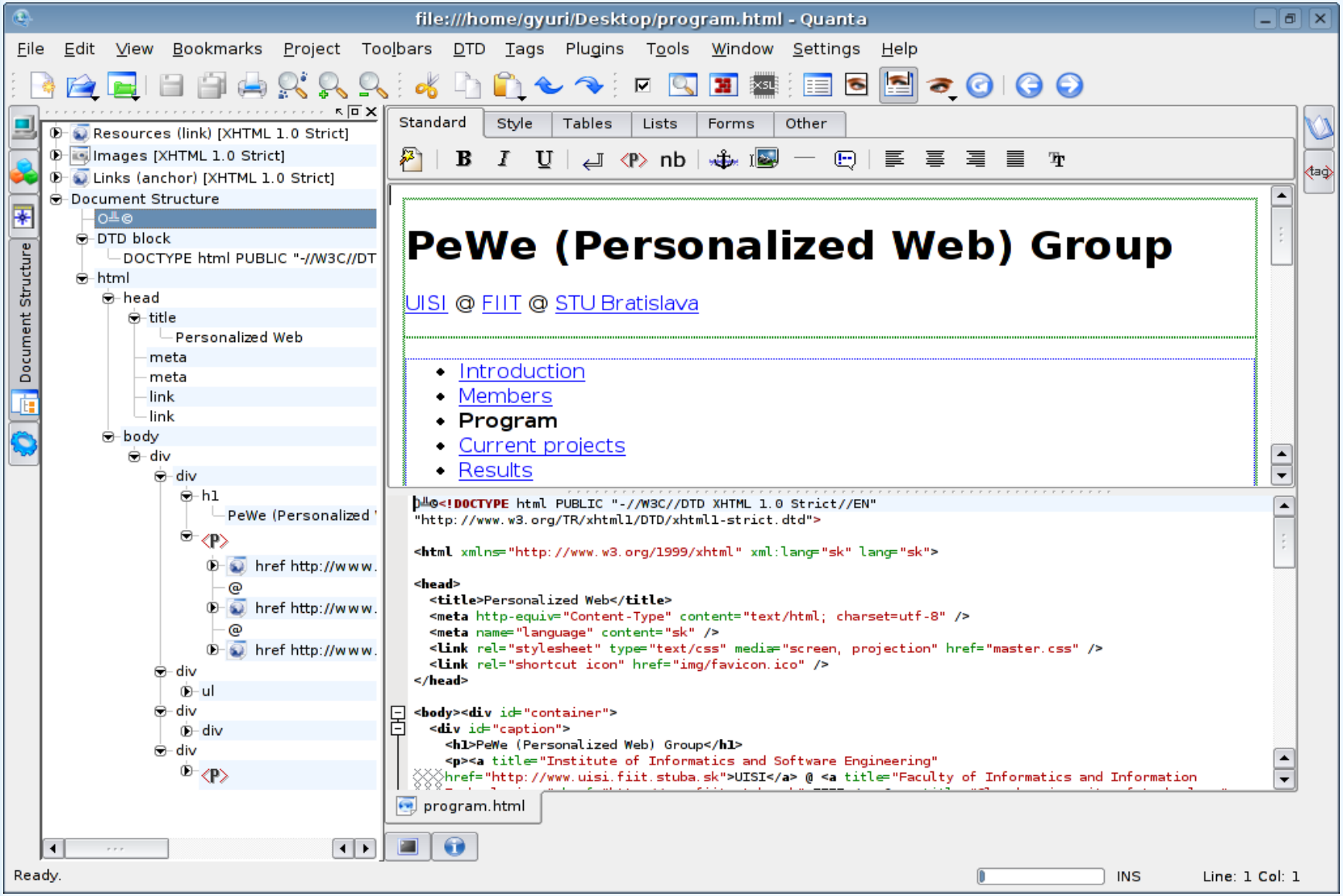
Obaľovač (Wrapper) web stránok

- Čo sú a na čo slúžia obaľovače?
- Problémy
  - parsovanie HTML dokumentov
  - verifikácia korektnosti obaľovača
  - čitateľnosť a rozšíriteľnosť obaľovača
  - navigácia
- Existujúce prostriedky

Prečo potrebujeme obaľovače - Aplikácie

- Porovnávanie ponúk/cien
- Vyhľadávanie
- Adaptácia starých alebo nevhodných rozhraní
- Integrácia údajov (napr. údaje z pobočiek)
- Sledovanie konkurencie

Vstup obaľovačov



Prístupy wrappovania

- Práca s HTML zdrojovým kódom ako so stringom (regulárne výrazy)
- Využitie stromovej štruktúry HTML kódu
- Vytváranie wrappera vymenovaním príkladov
- Adaptívne prístupy

HTML dokumenty

- HTML dokumenty sú pološtrukturované dáta, kde stromová štruktúra tagov môže prezradiť niečo o význame dát
- Často však HTML dokumenty nespĺňajú špecifikáciu
  - využitie robustných parserov (mozilla, beautifulsoap a iné)
  - "očistenie" HTML dokumentov, ich konverzia do XHTML

Vývoj a udržiavanie obaľovača

- Vývojový nástroj má zadať jednoduchý rámec na tvorbu obaľovača
- Nástroj má poskytovať spôsob na verifikáciu: ak obaľovač zlyhá kvôli zmene stránky, vývojár má o tom dostať správu
- Identifikácia prípadov použitia, model používateľa – aké znalosti sa od používateľa očakávajú

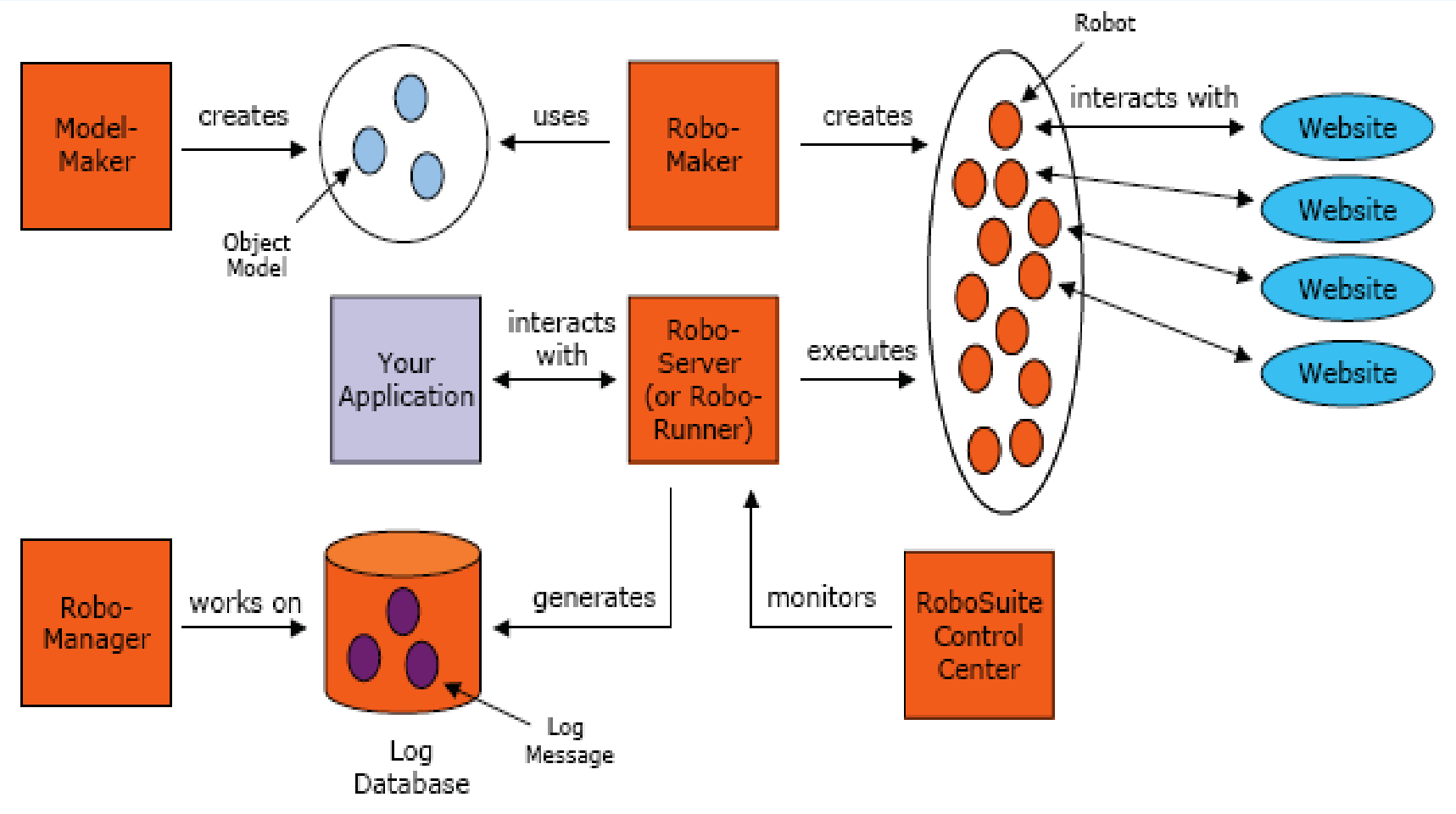
Navigácia

- Aby sme sa dostali k údajom, potrebujeme sa prebiť cez linky, cookies, heslá, formuláre
- Je potrebné riešiť autentifikáciu na webové sídla
- Zachytenie a nakódenie navigačných sekvencií môže viesť neprehľadnému kódu
- Spôsoby na zachytenie navigácie:
  - Proxy
  - Zmena liniek v HTML kóde
  - Modifikovaný prehliadač (resp. plug-in)

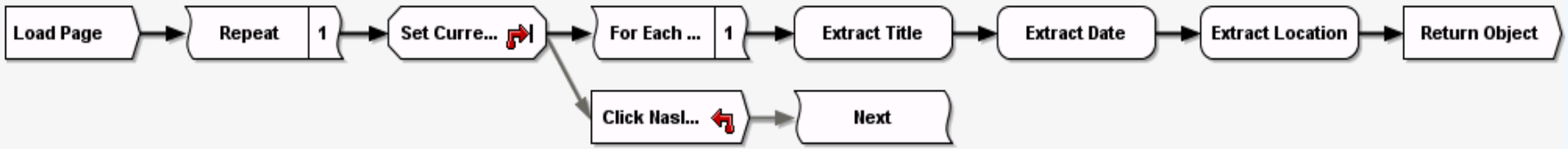
Existujúce prostriedky a služby

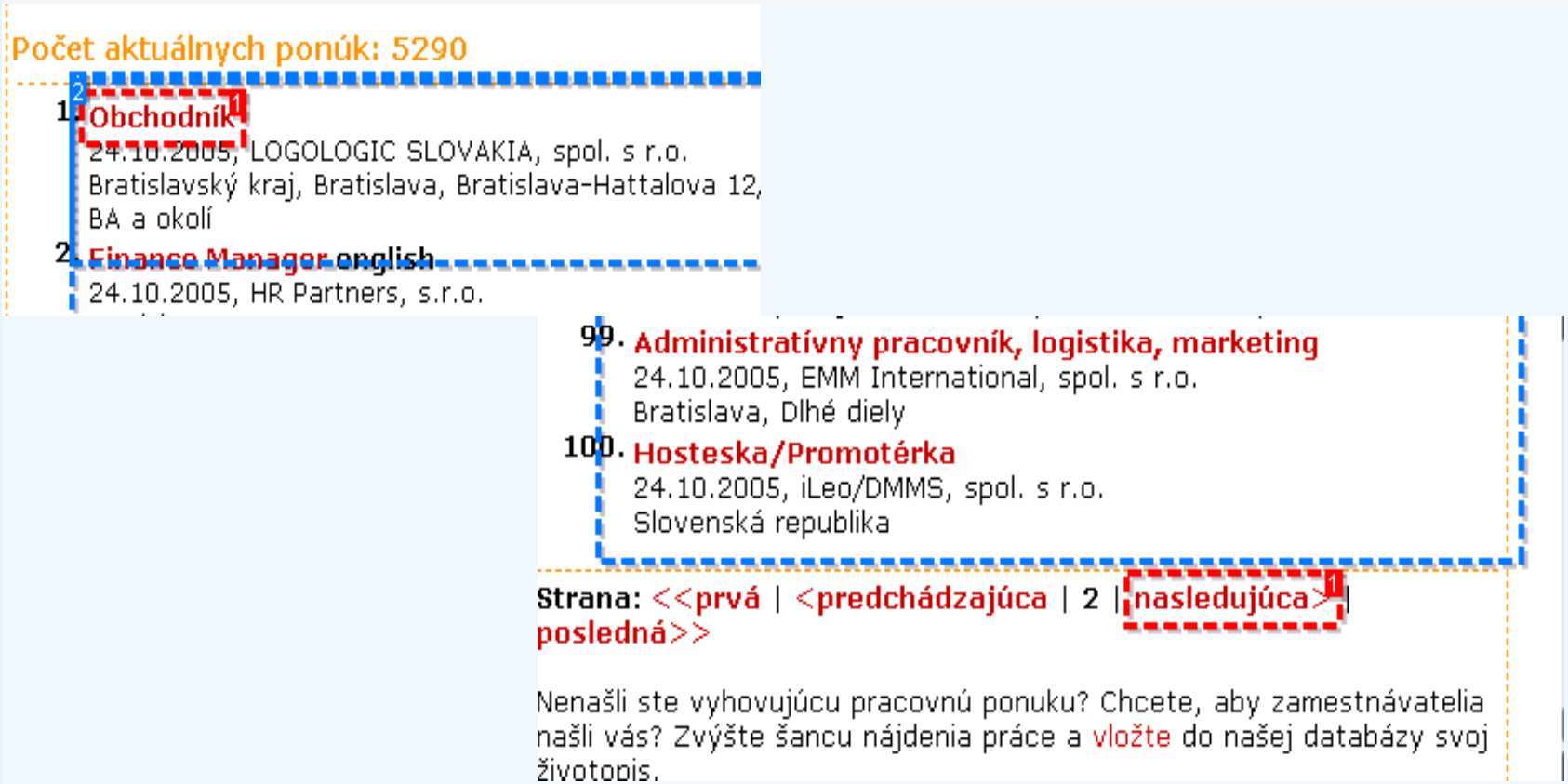
- I-point
  - ponúkajú konkrétne wrapovanie konkrétnych sídiel
  - neponúkajú nástroj na tvorbu
- Kapowtech
  - integrovaný nástroj umožňujúci jednoduchú tvorbu obaľovačov
  - Wrapper je automat generujúci XML
- LiXto
  - dva nástroje: Visual Wrapper a Transformation Server
  - Wrapper je sada pravidiel – rule-based prístup

Kapowtech


Model-Maker
creates
Object Model
uses
Robo-Maker
creates
Robot
interacts with
Website
Website
Website
Website
Your Application
interacts with
Robo-Server (or Robo-Runner)
executes
Robo-Manager
works on
Log Database
generates
Log Message
monitors
RoboSuite Control Center

Kapowtech


Load Page
Repeat
1
Set Curre...
For Each ...
1
Extract Title
Extract Date
Extract Location
Return Object
Click Nasl...
Next


Počet aktuálnych ponúk: 5290
1. Obchodník
24.10.2005, LOGOLOGIC SLOVAKIA, spol. s r.o.
Bratislavský kraj, Bratislava, Bratislava-Hattalova 12,
BA a okolí
2. Finance Manager english
24.10.2005, HR Partners, s.r.o.
99. Administratívny pracovník, logistika, marketing
24.10.2005, EMM International, spol. s r.o.
Bratislava, Dlhé diely
100. Hosteska/Promotérka
24.10.2005, iLeo/DMMS, spol. s r.o.
Slovenská republika
Strana: <<prvá | <predchádzajúca | 2 | nasledujúca> | posledná>>
Nenašli ste vyhovujúcu pracovnú ponuku? Chcete, aby zamestnávatelia našli vás? Zvýšte šancu nájdenia práce a vložte do našej databázy svoj životopis.

LiXto

- Visual Wrapper
  - Vizuálny nástroj na trvorbu obaľovača
- Trasformation Server
  - Nástroj na transformáciu, prezentáciu a doručenie výsledkov získaných z obaľovačov

Otázky

- Aký je vhodný model používateľa obaľovača?
- Aké sú bežné a najčastejšie problémy pri tvorbe obaľovača?
- Aký je vhodný jazyk na zápis obaľovača?
- ...