Sémantické hřiště (SemPg)

(návrh SW projektu)

Vedoucí:	Jan Dědek (jan.dedek@mff.cuni.cz) Peter Vojtáš (peter.vojtas@mff.cuni.cz)
Velikost týmu:	4-7 řešitelů
Jazyk, OS:	libovolný
Termín dokončení:	9 měsíců od zahájení projektu

Motivace

V souvislosti s výzkumem Sémantického webu a především myšlenkou jeho postupné *sémantizace* [1] potřebujeme experimentální platformu pro:

1. vývoj nástrojů pro extrakci informací z webu a
2. aplikaci těchto nástrojů v získávání sémantických dat z webu.

Cíl projektu

Cílem projektu je vytvoření experimentální SW platformy zmíněné v motivaci. Projekt by měl poskytnout zázemí pro vývoj a aplikaci extrakčních nástrojů, konkrétně poskytovat následující služby:

- Stahování webových stránek a dokumentů
- Archivace a správa dat:
  - Stažených stránek a dokumentů
  - Extrahovaných dat
  - Sémantických dat a ontologií
  - Vazeb mezi zdroji dat a extrahovanými daty
  - Konfigurace komponent systému, logování
- Řetězení (orchestraci) extrakčních nástrojů
- Podpora ruční anotace
- Měření kvality extrakce
- Vizualizace extrahovaných dat

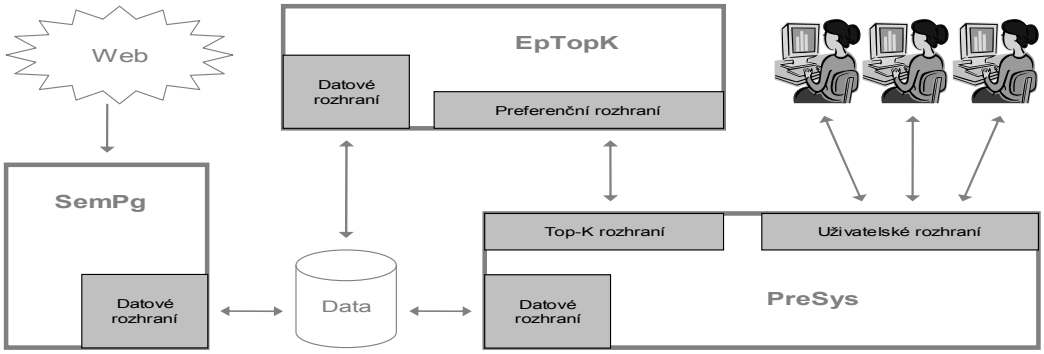
Hlavní přínos projektu spočívá v podpoře dekompozice řešení problému *sémantizace* webu. Platforma Sémantického hřiště pomůže rozdělit celé řešení do komponent a zajistí jejich vzájemné propojení (orchestraci). Zapojení komponent bude variabilní a jednotlivé komponenty bude možné vyměňovat nezávisle na jejich okolí. Žádoucí je též podpora distribuce systému v síťovém prostředí a podpora různých programovacích jazyků v komponentách.

Vlastní vývoj extrakčních nástrojů není součástí tohoto projektu. Řešení by však mělo demonstrovat svou funkčnost implementací (nebo zapojením) alespoň jednoho jednoduchého (exitujícího) extrakčního nástroje. V případě většího týmu bude použito nástrojů více. Přehled existujících extrakčních nástrojů je možné nalézt např. v [2].

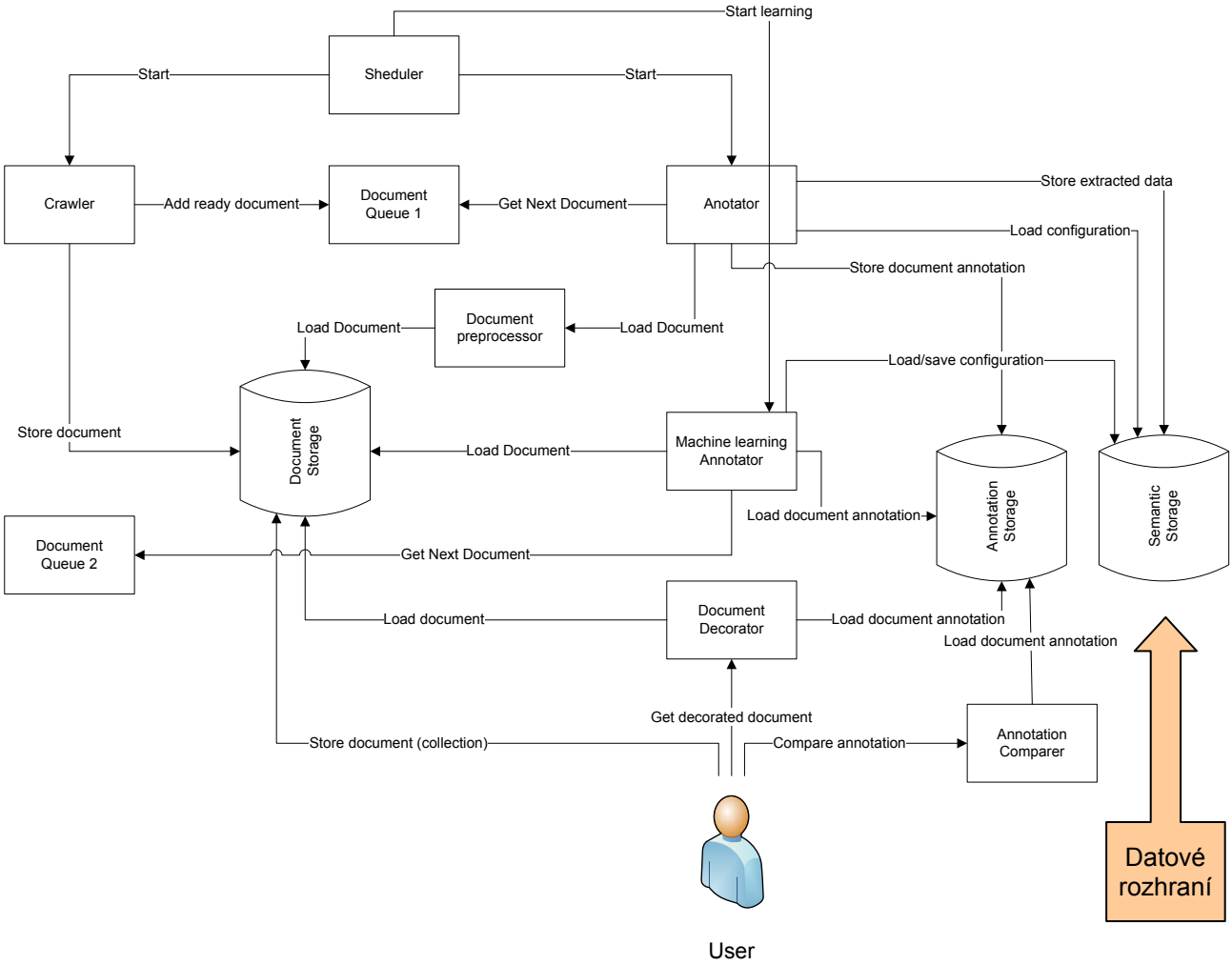
Řešení by mělo co nejvíce vyžívat existující softwarové komponenty. Podle velikosti řešitelského kolektivu je možné některé části systému redukovat, případně odevzdat jen ve formě interfaců. Na projektu bude možné pokračovat v DP.

Tento projekt by měl též poskytovat data a datové rozraní pro související projekty *EpTopK* Matůše Ondreičky a *PreSys* Alana Eckhardta. V budoucnu plánujeme

integraci všech tří projektů do systému *W2U – Web to User*, který je znázorněn na obrázku 1.



Obr. 1 Schéma systému *W2U – Web to User*



→ Šipky v obrázku znázorňují volání funkcí.

Obr. 2 Schéma systému *Sémantické hřiště* (předběžný návrh)

Hrubý návrh řešení

Obrázek 2 představuje předběžný návrh systému Sémantické hřiště. Následuje stručný popis jednotlivých komponent tak, jak si ho v současné době představujeme. Tato představa má sloužit jako zdroj inspirace, nejedná se o závaznou specifikaci.

- **Semantic Storage**

Komponenta, která zodpovídá za správu extrahovaných a sémantických dat.

Load / Save configuration

Používá **Annotator,** když chce načíst svou sémantickou konfiguraci (např. extrakční ontologii).
Metoda **Save** slouží především k ukládání naučené konfigurace v procesu strojového učení anotátoru.

Store extracted data

Slouží k uchování extrahovaných dat, která byla anotátorem získána z anotovaného zdroje. Metoda musí umožnit spolu s daty uchovat (případně zajistit dohledatelnost) informaci o:

- Zdrojovém dokumentu anotace (jeho verzi apod.)
- Anotátoru, který data vyprodukoval
- Konfiguraci anotátoru, která byla v tomto procesu anotace použita
- Případně další informace, například datum a čas

- **Document Storage**

Perzistentní úložiště dokumentů.

Load document

Načte dokument z úložiště.

Store document

Uloží dokument do úložiště.

- **Annotation Storage**

Perzistentní úložiště anotací. Anotací zde rozumíme zobrazení ze základních strukturních jednotek dokumentu (například znaků) na extrahovaná (sémantická) data.

Load document annotation

Načte anotaci daného dokumentu z úložiště.

Store document annotation

Uloží anotaci daného dokumentu dokument do úložiště.

- **Crawler**

Stahuje data z webu.

- **Document Queue**

Fronta dokumentů. Slouží jako seznam zpracovaných respektive na zpracování čekajících dokumentů. Frontu může plnit uživatel nebo jiná komponenta systému – především **Crawler**. Dokumenty z fronty vytahují především **Anotátory**.

- **Annotator (Extraktor)**

Komponenta, která sémanticky anotuje vstupní dokument, případně z něj pouze extrahuje strukturovaná data. V této roli by též šlo zapojit nástroj pro ruční sémantickou anotaci.

- **Machine learning Annotator**

Budoucí **Anotátor** momentálně zapojený v učícím módu. K učení potřebuje referenční anotace a sadu dokumentů, na kterých se bude učit.

- **Document preprocessor**

Složí k předzpracování dokumentu. Například může převádět PDF dokument na HTML nebo HTML dokument na prostý text.

Tyto komponenty by měly zachovat vazby anotací na originální dokument.

- **Annotation Comparer**

Komponenta, která umožní porovnat různé anotace nad stejným dokumentem. Na základě porovnání pak může vyčíslit vzájemné metriky jako například *přesnost, úplnost, F-measure.*

- **Document Decorator**

Komponenta pro zvýraznění (například barevné vyznačení) anotací v originálním dokumentu.

Reference

[1] Jan Dědek, Alan Eckhardt, Leo Galamboš, Peter Vojtáš: **Sémantický Web,** in DATAKON 2008, Brno, ISBN 978-80-7355-081-3, pp. 12-30, 2008.

[2] Chia-Hui Chang, Mohammed Kayed, Moheb Ramzy Girgis, Khaled F. Shaalan: **A Survey of Web Information Extraction Systems**, in IEEE Transactions on Knowledge and Data Engineering, vol. 18, no. 10, pp. 1411-1428, October, 2006.