# DATABÁZE V PROSTŘEDÍ WEBU

**Radek Dvořák**

ČVUT Praha, Elektrotechnické fakultě, Katedra počítačů, Karlovo náměstí 13, 121 35 Praha 2
DvorakR1@fel.cvut.cz

**Abstrakt**
Cílem projektu je nalezení universálního modelu datového zdroje při přístupu k heterogenním datovým prostorům a k heterogenním typům informací na Internetu/webu. Stěžejní oblastí, datových zdrojů dnešního světa jsou databáze. Přistupovat k nim je ovšem z důvodu rozmanitosti jejich charakteru, nejednotného pohledu na data a přístupu k datům velice složitý. Proto vhodně vytvořený model datového zdroje umožní snadný a hlavně jednotný pohled na datové zdroje a to jak uložené v klasických (relační) či specializovaných databázích (objektové, XML a další), tak i ostatních datových zdrojích. Jedná se o mezičlánek mezi datovým zdrojem a uživatelem (člověk, počítač), který odstraní nejednotnost při přístupu k datům a zajistí uživatelsky příjemný prostředek spolupráce s datovým zdrojem. Formalismus modelu bude zahrnovat i nestandardní metody vyhledávání jako jsou přibližné vyhledávání, kontextové vyhledávání, vyhledávání netextové informace.

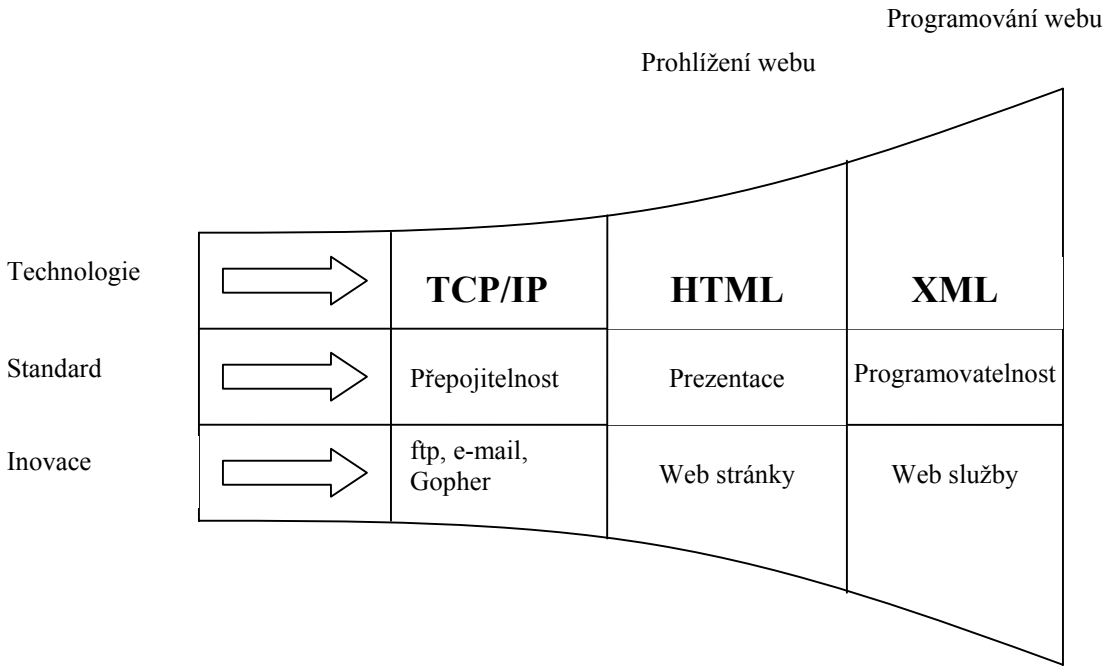
## 1. Vývoj webových technologií

Pojmem inženýrství se obecně rozumí systematická aplikace vědeckých znalostí při návrhu a tvorbě cenově efektivních řešení praktických problémů. Jestliže použijeme tuto obecnou definici inženýrství i na oblast webu, pak můžeme definovat pojem webového inženýrství jako aplikaci systematického, disciplinovaného, kvalifikovaného přístupu k vývoji, provozu a údržbě webových aplikací.

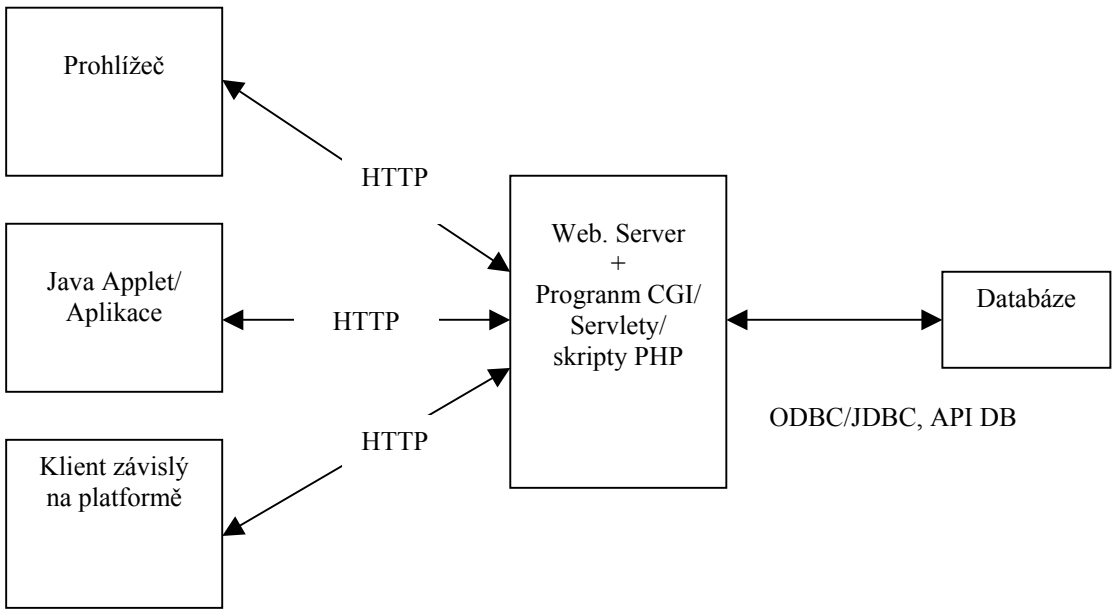
Web sám o sobě má svůj vývoj jak naznačuje obrázek, od využití TCP/IP spojení - statické webové prezentace, přes interaktivní webové prezentace založené na technologiích klient-server, až po dnes dynamicky se rozvíjející oblast webových služeb. Poslední vývojový stupeň uvažuje nejen o interakci člověk - stroj, ale i stroj - stroj. Proto je kladen velký důraz na kvalitu popisu informací tak, aby ji byl schopen správně interpretovat jak člověk tak i stroj. ( Norma pro popis webové služby je WSDL, protokol vzájemné komunikace objektů SOAP je založený na textovém formátu XML a seznam webových služeb UDDI.)

## 2. Webové aplikace dneška

Dnešní webové aplikace jsou postaveny nad třívrstvým modelem. Za první vrstvu se považuje klientské rozhranní - web klient, nejrozšířenější je webový prohlížeč (Internet Explorer, Mozila,..). Tento klient spolupracuje s druhou vrstvou která je pro webové aplikace příslušný web server rozšířený o podporu programování (CGI+API DB, PHP, ASP,…)



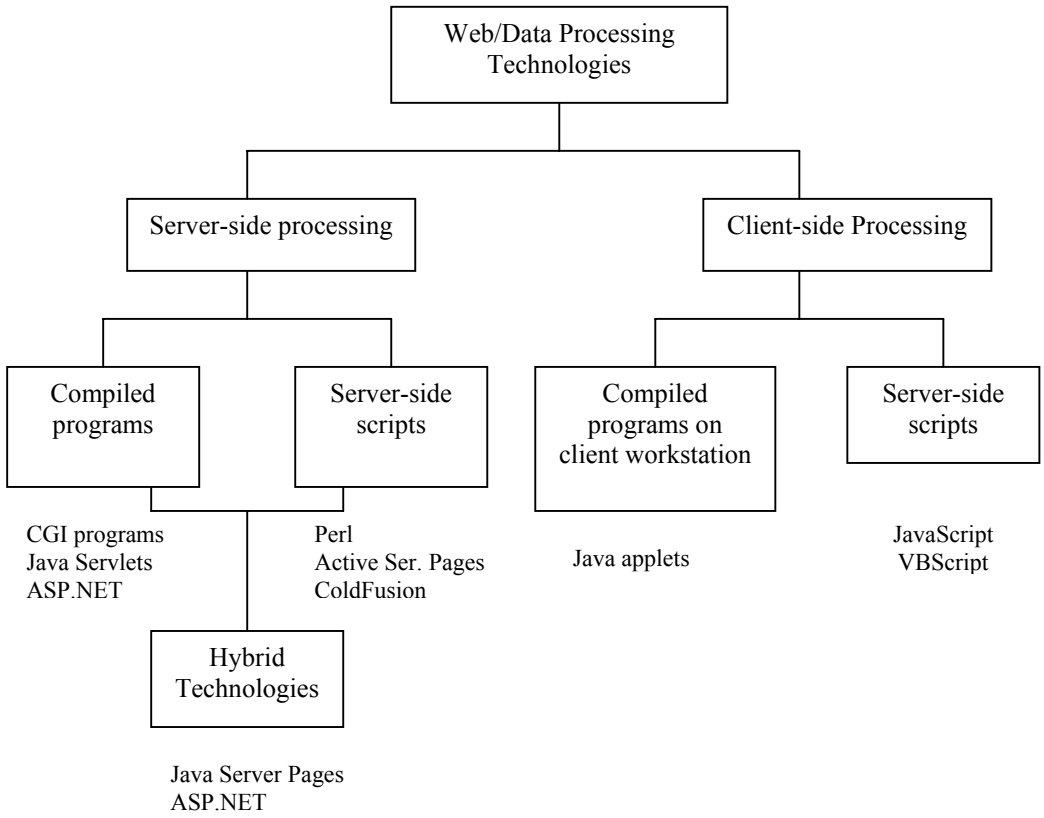

Obr.1 Etapy vývoje Webových technologií




Obr.2 Model dnešních aplikací

Aplikace pracující ve druhé vrstvě potřebuje spolupracovat s nějakým zdrojem dat ať už ve "skutečných" databázích, ale i nestrukturovaně v různých formátech, na různých platformách. Poslední vrstvou tedy většinou je databáze, ve které jsou uloženy datové informace aplikace.

Spojení mezi programovou logikou a databází se dnes řeší většinou rozhraním ODBC, JDBC či specializovaným API daného aplikačního rozhranní k databázi (proprietální).




Obr.3 Rozdělení programových prostředků v prostředí webu

Programové prostředky využívané při programování ve webovém prostředí popisuje obrázek 3. Webové aplikace je možno rozdělit na procesy běžící na straně serveru a klienta. Každá tato oblast se dále dělí na kompilované programy a skripty. Dnes se začínají čím dál více prosazovat hybridní technologie, které spojují dobré vlastnosti kompilovaných programů a skriptů. Nejznámější příklady všech jazyků jsou uvedeny v obrázku. Tyto jazyky většinou umožňují přístup k databázím, a to buď s využitím ODBC/JDBC či specializovaným API pro příslušné databáze.

## 3. Datové zdroje

V dnešním světě kde v každý okamžik vzniká velké množství dat jsou databázová technologie již neodmyslitelným nástrojem pro jejich efektivní shromáždění, zpracování a opětné vyhledání informací. Na základě evoluce informačních technologií a zvláště databází vzniklo několik typů databází. Mezi dnes nejrozšířenější bezesporu patří relační databáze, označované jako RDBMS (Relational Database Management Systems). Jejich zástupcem můžeme například jmenovat MySQL (http://www.mysql.com/). Druhou neméně důležitou skupinou jsou objektové databáze, označované jako ODBMS (Object Oriented DBMS). Jejich zástupcem je databáze GameStone (http://www.gamestone.com/).V dnešní době se ale začínají obě skupiny prolínat a dnešní databáze se snaží přijmout z obou skupin ty lepší vlastnosti, označujeme je jako ORDBMS (Object Relational databáze). Jejich zástupcem může být databáze Oracle (http://www.oracle.com), která umožňuje jak relační tak objektový

přístup k datům s jejich výhodami či nedostatky. Vedle těchto základních a nejznámějších databází existuje i velká řada systémů, které řeší práci s daty specifickým způsobem. Jejím příkladem může být databáze Cache (http://www.intersystems.com/cache/technology/), označovaná jako postrelační databáze. Databáze Cache je datovým uložištěm, které je schopno za jistých okolností přistupovat ke stejným datům relačně i objektově. Ovšem dnes se vyvíjí i databáze označované jako XML, které mají jako nativní prostředí strukturu XML.
Spolu s databázemi, které představují největší ”zásobárnu” dat je ale i možné ukládat data do souborů se specifikovanou strukturou, jakou jsou například HTML, PDF, XML (většina formátů je standardizována organizací W3C, http://www.w3c.org/) a další.

Všechny výše uváděné způsoby uložení dat budeme souhrnně nazývají datová uložiště. Skupina datových uložišť nebude nikdy konečná, neboť se neustále nové vytvářejí a jiné zanikají.

Mimo výše uváděných odlišností mezi jednotlivými datovými uložišti je ještě dnes třeba brát v úvahu platformu nad kterou je systém postaven a to nejen jako OS (operační systém), ale i "protokol", kterým se k datům přistupuje.

A jaký je vůbec důvod k přístupu k informacím na webu prostřednictvím databází. Důvode je velice jednoduchý a plyne ze základních specifiky webového inženýrství. Chceme aby data byla bezpečně uložena, byla aktuální a jejich update se provedl bez zásahu administrátora. Všechny tyto výhody právě přináší využití databází na webu

Při spojení obou oblastí se ovšem jeví jako kritický problém zajištění *univerzálního přístupu* k datovým uložištím z prostředí webu, resp. Internetu. Dnešní trendy při řešení situace přístupu k datům z webu je využití rozhranní ODBC a JDBC, ale to má nemalé problémy a to nejen v omezenosti pouze na základní příkazy jazyka SQL (dnes se zápis příkazů téměř na každém databázovém zdroji odlišuje) ale i tato řešení nejsou zcela efektivní diky nutnosti přístupu přes další prostředky. V případě rozhranní JDBC existují 4 způsoby, kde některé z nich vyžadují přístup současně přes JDBC i ODBC . Pro každé datové uložiště dnes již většinou existuje API - přímé rozhraní umožňující přístup k datům, ale tato rozhraní jsou velice specifická a podporují pouze různé úrovně přístupu k datům z daného programového prostředí do příslušného datového zdroje. Současný výzkum se snaží o lokální řešení dílčích problémů jako je například převod statických stránek do databází, ale globální pohled stále chybí. Dnes na odborných pracovištích vznikají prostředky sblížení různého obsahu datových uložišť, jedná se hlavně o hledání efektivní transformace jednoho typu datového uložiště na jiné, při zachování výhod příslušných datových zdrojů. Z předchozích bodů vyplývá, že chybí univerzální model, pomocí kterého by bylo snadno a bez znalosti specifik datového úložiště možné přistupovat k datům.

Cílem výzkumu je nalezení universálního modelu datového zdroje při přístupu k heterogenním datovým prostorům a k heterogenním typům informací na Internetu/webu. Stěžejní oblastí, datových zdrojů dnešního světa jsou databáze. Přistupovat k nim je ovšem z důvodu rozmanitosti jejich charakteru, nejednotného pohledu na data a přístupu k datům velice složitý. Proto vhodně vytvořený model datových zdrojů umožní snadný a hlavně jednotný pohled na datové zdroje a to jak uložené v klasických (relační ) či specializovaných databázích (objektové, XML a další), tak i ostatních datových zdrojích. Bude se jednat o mezičlánek mezi datovým zdrojem a uživatelem (člověk, počítač), který odstraní nejednotnost při přístupu k datům a zajistí uživatelsky příjemný prostředek spolupráce s datovým zdrojem.

První krok návrhu bylo provedení analýzy dnes dostupných databázových systémů. V první fázi bylo nutné nalézt databázové stroje. Internetové vyhledávače nabídli v celku velké množství systémů a projektů, které tyto otázky řešili ovšem při bližším zkoumání těchto systém se velice často narazilo na neexistující odkaz, či nulové informace o projektu, který nebyl dokončen. Po odfiltrování tohoto šumu byla vytvořena databáze "databází". Její struktura byla volena s ohledem na analýzu možností a požadavků dnešních systémů. Jedná se hlavně o základní množina podporovaných datových typů, podporovaná rozhranní, podporované systémové prostředí a další doplňující informace. Databáze datových zdrojů je otevřeným systémem, ve kterém je možné informace aktualizovat a je dostupná na adrese http://db.dvorakconsulting.net. Z této analýzy je patrná standardizace datových typů, rozhranní menších databází k velkým, které udávají směr vývoje.

## 4. Výsledky průzkumu

Podpora programového prostředí pro databáze je ve většině případů velice omezená. Relační databáze často podporují pouze PL/SQL. Objektové databáze většinou podporují pouze vlastní nativní prostředky (např. pro Cache se jedná o Caché ObjectScript, Caché Basic)

Neméně důležitou otázkou je interface k databázi mezi standardní patří jenom podpora pro rozhranní ODBC a JDBC. Mimo těchto rozhraní je pro velké množství databází podpora v jazycích Java, .NET, C++ a PHP, jedná se většinou o nativní prostředky. Ke XML databázím je nejrozšířenější interface XML:DB API.

Podpora SQL v relačních databází je dnes většinou pouze na úrovní verze 92. Ale jsou databáze, které nepodporují ani tu, některé systémy podporují vybraná rozšíření dle normy 99. Při analýze podporovaných operačních systémů je vidět rozmanitost dnešního světa IT, ve kterém jsou databáze implementovány na většině dnes dostupných implementací.

Většina síťových databází podporuje TCP/IP protokol, některé ze systémů ještě podporují Novell síťové prostředí IPX/SPX.

Rozložení typu databází stále vládnou relační databáze s téměř 60 %, další velkou skupinou jsou XML, následované kombinovanými (13%) a objektovými databázemi(9%)

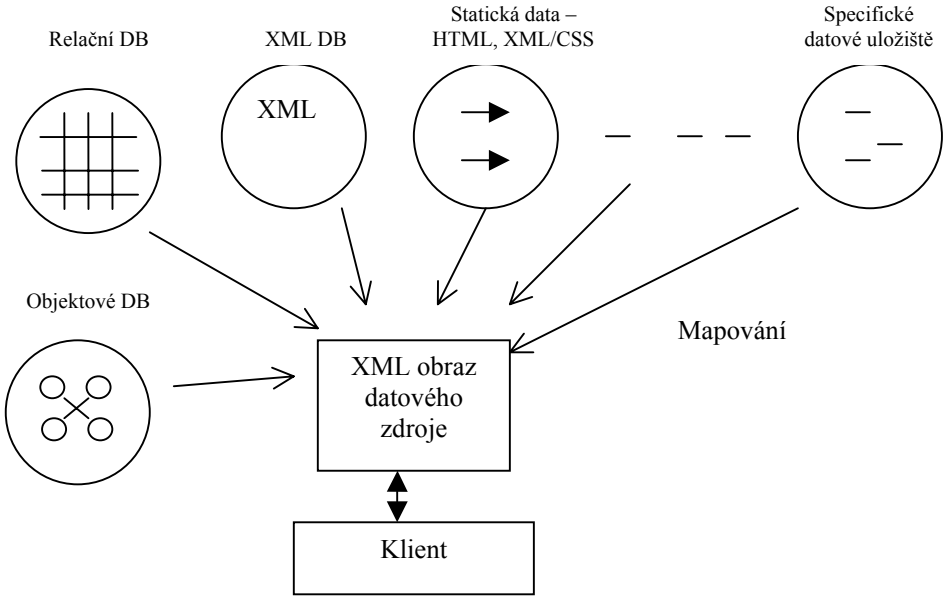
## 5. Stanovení podmínek pro univerzální model - data v datových zdrojích

Z předchozí analýzy byly vysledovány tyto závěry pro datové elementy. Data, která se dnes vyskytují v databázích je možné rozdělit na dvě základní skupiny dynamické a statické. Základním prvkem dynamických jsou metody a procedury - algoritmus a "znalosti" - logická pravidla. Mezi statické data patří jednak " klasické" jako jsou například number, char,.. Mezi klasickými daty je dnes vhodné rozlišovat zda se jedná o atomické či strukturované datové typy. Další významnou skupinou mimo " klasických " dat jsou texty, visuální komponenty (bitmapová a vektorová grafika a video) a nakonec ještě zvukové datové typy.
Z tohoto výčtu je patrné že rozmanitost dnešních datových typů je velká a přístup a práce s nimi ne zcela jednotná. Každý z těchto typů vyžaduje po uživateli jeho znalost a znalost práce s ním.

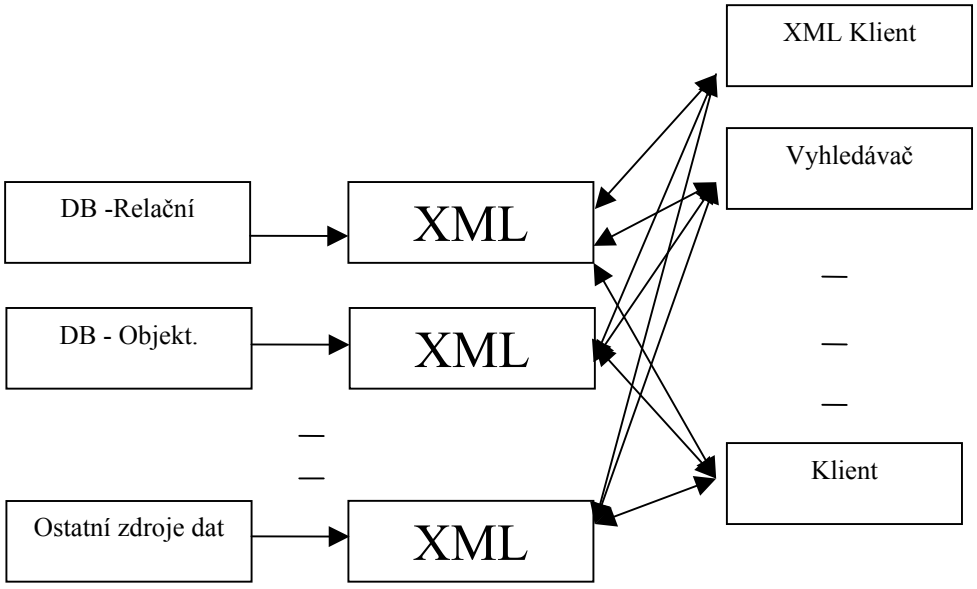
Perspektivním řešení se jeví vytvoření modelu datových zdrojů, do které by bylo možné všechny takovéto informace uložit/namapovat a pak bez znalosti příslušného datového zdroje k nim přistupovat. Představa by měla být podobná jako je tomu u jazyka Java, kde kód je také přeložen do „meziproduktu“ byte kódu, který je pak libovolně přenositelný na všech platformách.

Takovouto strukturou se dnes jeví XML, které jednak podporuje velké množství datových elementů, které se v databázích využívají, ale splňuje i další podmínky jako je efektivní přístup k datům pro lidský mozek tak i pro stroj. XML struktura také umožňuje velice sofistikovaný přístup k datům, který je řešen stromovou strukturou XML. S masovým rozšiřováním XML se dostává do podvědomí uživatelů způsob dotazování na tato data. Mezi další důvody výběru XML je jeho možnost použít doplňkové informace k různým datovým elementům nosným informacím modelu, které XML struktura také umožňuje a není zcela jednoduché ji zachytit v databázových systémech.

Tato volba také velice odráží dnešní směr ukládání dat, které se řeší buď vlastním uložením v databázi nebo dnes se rozvijí velké množství projektů, které hledají optimální metody transformace/ukládání dat do XML dokumentů, které ovšem bude model akceptovat. Při sledování dnešních systémů jejich vývoj k těmto způsobům konvergují.
XML struktura je také velice vhodná svoji velkou samopopisující schopností.




Obr.4 Pohled klienta na datové zdroje




Obr.5 Pohled na systém

Na volbě struktury a obsahu tohoto stromu by záležela i možnosti rozšiřujícího hledání. V neposlední řadě jazyk XML je dobře podporová při zobrazení informací přes webové rozhranní.

## 6. Shrnutí výhody navrhovaného řešení

- Jazyk XML se dnes stal již standardem při komunikace člověk-stroj tak i stroj-stroj
- XML představuje standardní způsob reprezentace dat
- Univerzálnost přístupu (Umožnit uživateli jednotný pohled při práci s neznámým datovým zdrojem, vyhledání informací )
- Jednoduchá komunikace - založena na XML
- Umožňuje přístup k datům bez znalosti fyzického uložení dat
- Modelu by měl zahrnovat i nestandardní metody vyhledávání
- Bude brát v úvahu lokální specifika - kulturní,…
- Bude brát v úvahu proprietární technologie.
- Konzistentní rozhranní
- Nezávislé na prostředí
- Prostředí založené na XML a databázích umožňuje snadnou transformaci do podoby, kterou je schopné zobrazit koncového zařízení (PDA, Mob. Tel., ..)

## 7. Budoucí práce

Dalším cílem bude vhodně navrhnout strukturu univerzálního XML rozhranní (XML obraz datového zdroje), do kterého by se mapovali stávající zdroje, ale model by měl být natolik obecný, aby bylo možné do něj zakomponovat nové typy datových zdrojů
U vybraných zdrojů implementovat mapování a testovat sjednocení přístupu
Přitom brát v úvahu lokální specifika, - brát zřetel mimo funkční závislosti také psychologické vnímání lidského mozku při přístupu k datům

Umožnit uživateli jednotný pohled při práci a vyhledávání informací na webu, nejen pracujících nad relačními, objektově-relačními databázemi, ale také i na XML dokumenty, které mohou být též umístěny různým způsobem, ale i na informace induktivní a velké množství dalších. Nalezením jednotné metodiky vyhledávání nad těmito bázemi, by mohl být řešen transformačními datovými strukturami, nebo nějakým funkčním mechanismem zajišťující transparentnost uchovávaných dat.

**Literatura:**


1. Chaudhri, Akmal B. Web, Web-services, and database systems : NODe 2002 Web- and database-related workshops Erfurt, Germany, October 7-10, 2002 : revised papers Berlin : Springer, 2003, pp. 206–220.
2. Kaliničenko, L. A. Advances in databases and information systems : 7th East European conference, ADBIS 2003,Dresden, Germany, September 3-6, 2003 : proceedings, Berlin : Springer, 2003
3. Murugesan, San, Deshpande Yogesh. Web Engineering, Springer-Verlag Berlin, 2001
4. Carbenell J.G., Siekmann,J. Managing WEB-Based Data, IEEE IC, July August, 2002