

SCHÉMOVÉ JAZYKY

Přednáška z předmětu KMA/POK

Otakar ČERBA

Západočeská univerzita v Plzni

Schémové jazyky

- Jazyky pro popis dokumentu
- XML schémata
- XML Schema Languages

Schematron



Rick Jelliffe

RELAX NG



James Clark

DSDL



Eric van der Vlist

RELAX NG



Murata Makoto

W3C XML Schema



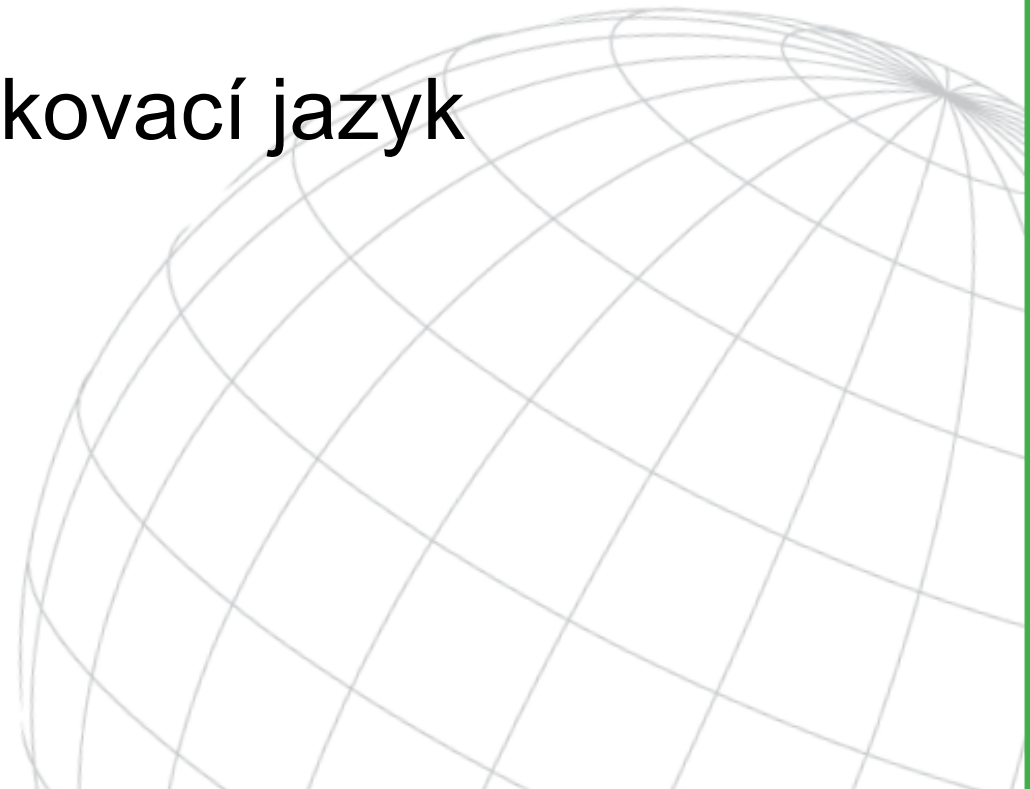
Henry S.
Thompson

Schémové jazyky

- **Formální definice nového značkovacího jazyka** = definování podmnožiny SGML/XML
- Umožňují definovat jednotlivé prvky XML dokumentu (elementy, atributy, entity...) a způsoby jejich použití (přípustné kombinace, omezení, datové typy, vzájemné vazby...)
- Schémata umožňují plné využití všech možností značkovacího jazyka, především kontroly dokumentu z hlediska sémantiky (syntaxi zajišťuje především **parser**)
- Kontrolu dokumentu z hlediska schématu zajišťují **validátory**
- Schémata mohou být chápány jako filtry chránící systémy a především uživatele před širokou různorodostí XML formátů

Proč schémata?

- Formální definice značkovacích jazyků
- Validace dokumentů
- Lepší manipulace s dokumenty prostřednictvím XML editorů
- Dokumentace pro značovací jazyk
- Databinding



Kontrola dokumentů

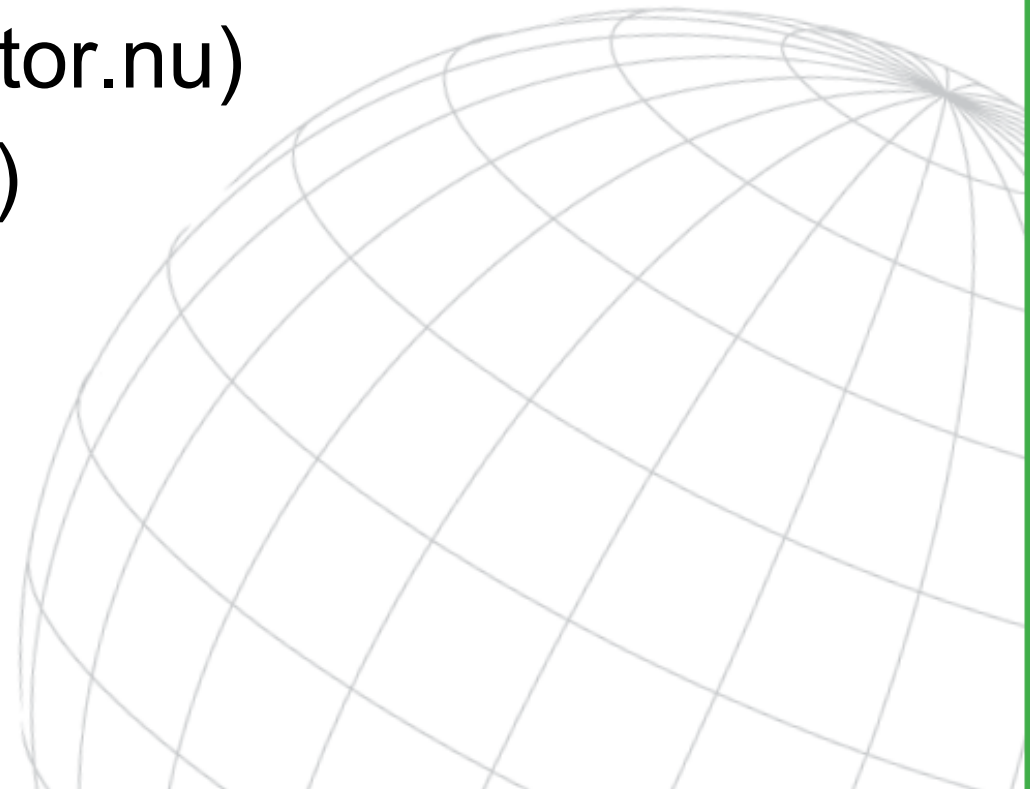
- **Správně formátovaný dokument** (well-formed document) - dokument splňuje veškeré náležitosti XML dokumentu (správné vnořování prvků, hlavička...) → **syntaxe**
- **Platný (validní) dokument** (valid document) - dokument odpovídá příslušnému souboru zapsaném v jazyce pro popis dokumentů → **syntaxe + sémantika**

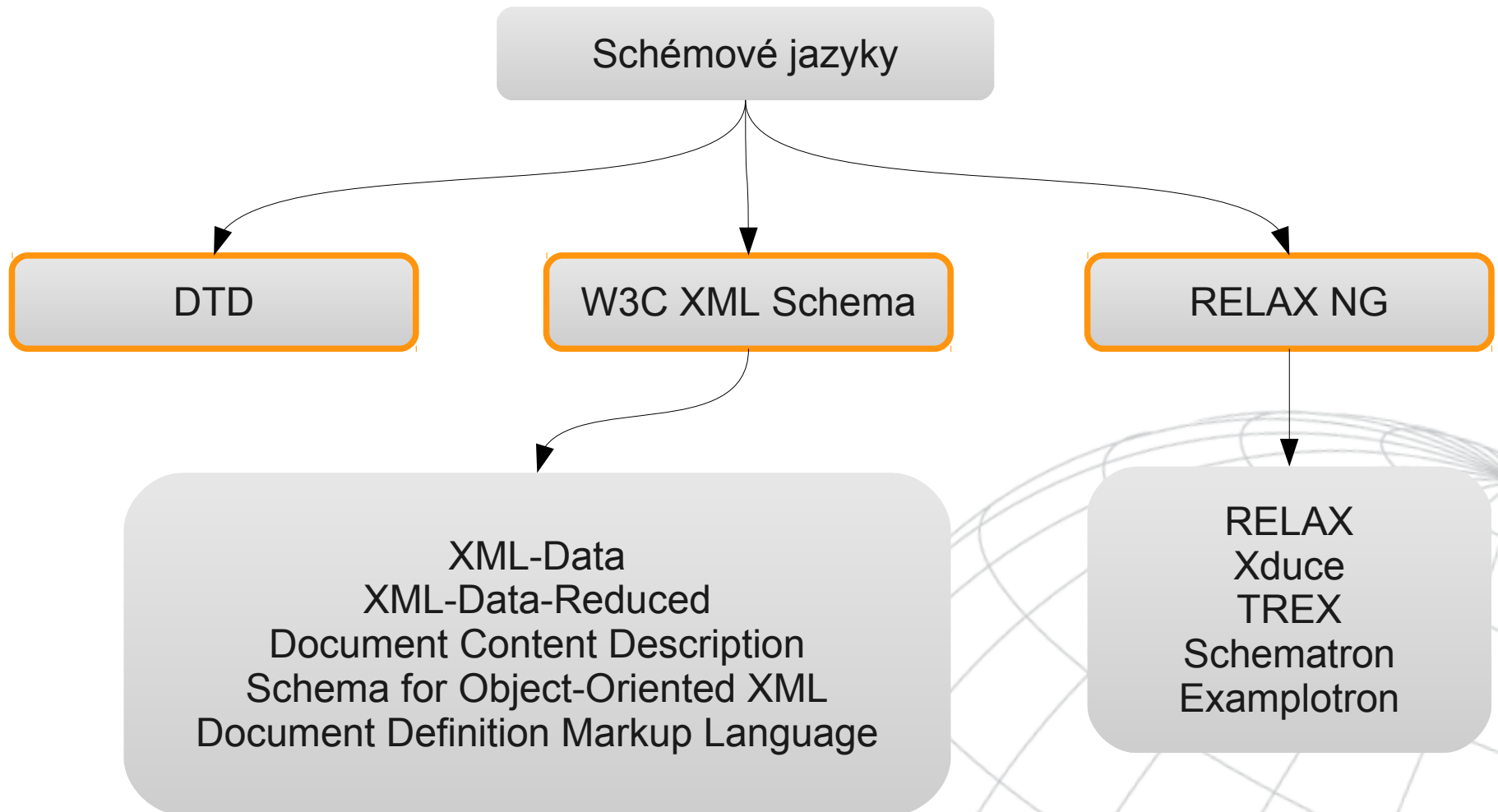
Úrovně validace dokumentu

- Struktura - testování prvků XML dokumentu (elementů, atributů...), ale nikoli jejich obsahu.
- Datové typy - vytvoření vazeb mezi značkováním a obsahem - kontroluje se obsah elementů a atributů bez ohledu na jejich vzájemné vazby.
- Integritní omezení - kontrola všech vazeb (např. odkazů) uvnitř dokumentu nebo mezi dokumenty.
- Business rules - kontrolují se omezení typu „datum narození musí být menší než datum úmrtí“ apod.

Validátory

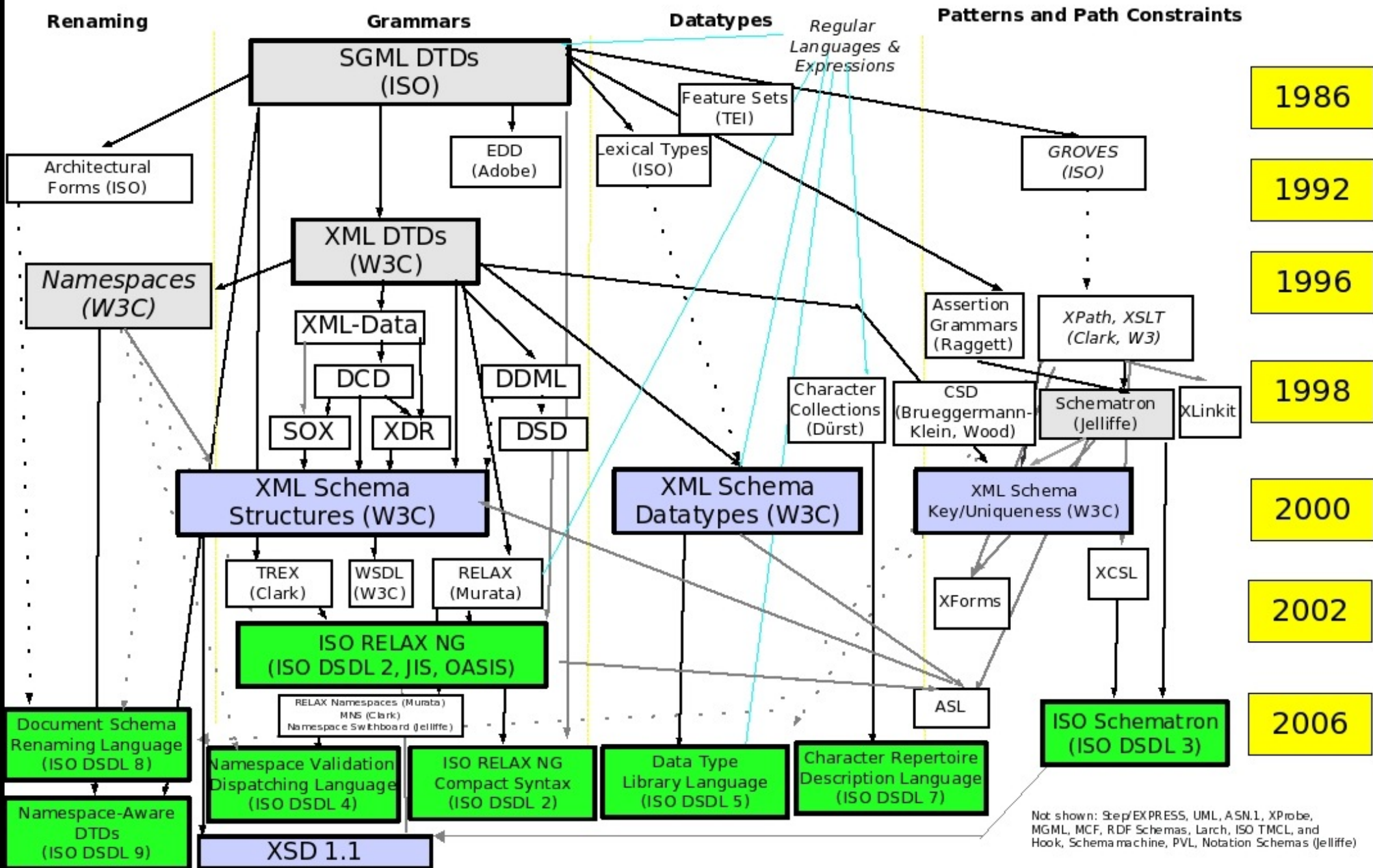
- **Jednoúčelové** (kontrola konkrétního schématu)
vs. **víceúčelové** (kontrola konkrétního schémového jazyka)
- Integrace do prohlížečů, editorů
- Webové služby (Validator.nu)
- Externí programy (Jing)





Family Tree of Schema Languages for Markup Languages

Rick Jelliffe © 1999, 2006. Permission granted to use this in any way providing this attribution is kept. (v.5)



DTD

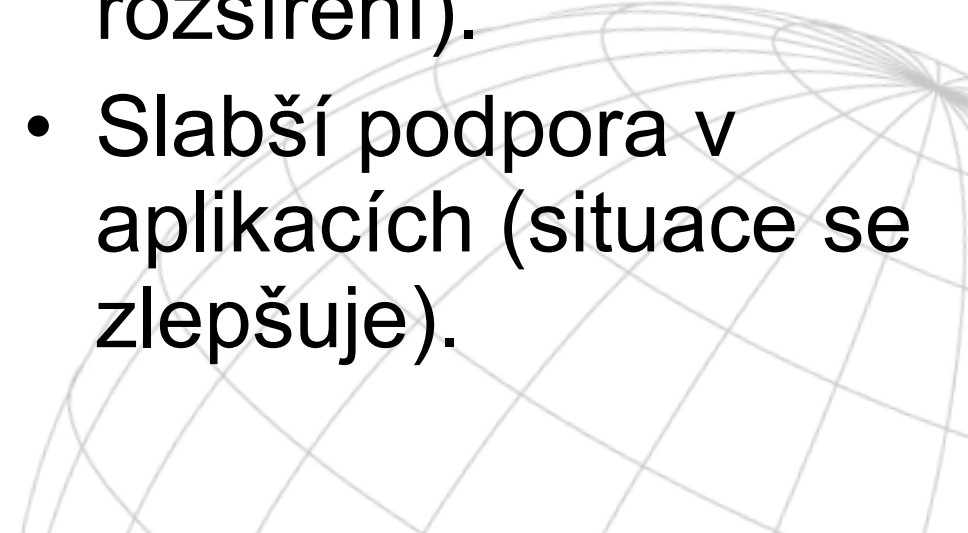
- Podpora SGML
- Podpora velkého množství aplikací
- Kompaktní, přehledné a čitelné
- Neodpovídá současným potřebám
- Neexistuje podpora jmenných prostorů
- Nedokáže definovat povolený obsah elementů
- Nemožnost určení většiny datových typů
- Nestandardní zápis
- Na DTD se nedají aplikovat transformační XSLT styly

W3C XML Schema

- Druhý nejpoužívanější schémový jazyk – podpora v komerčních aplikacích (MS, IBM, Sun, Oracle...).
- Podpora datových typů.
- Práce se jmennými prostory.
- Referenční integrita, zajištění klíčových hodnot.
- Objektová orientace.
- Modularizace schémat.
- Tvorba dokumentace schématu
- Složitá specifikace
- Komplexnost



RELAX NG

- RELAX NG je k dispozici ve dvou formách zápisu - XML a RNC (kompaktní textová syntaxe). Obě varianty nabízí poměrně úsporný a jednoduchý zápis.
 - Jednoduchý a „příjemný“ jazyk.
 - Propojení RELAX NG a jazyka Schematron
 - Neexistuje vestavěná podpora datových typů (pro doplnění této funkce je nutné využití rozšíření).
 - Slabší podpora v aplikacích (situace se zlepšuje).
- 

Schematron

- Umožňuje deklarace specifických omezení (např. obsah jednoho elementu musí být větší než obsah jiného, vazby mezi daty uloženými v několika dokumentech apod.).

```
<schema xmlns="http://www.ascc.net/xml/schematron" >
  <pattern name="Obsahuje element Jazyk element Navez?">
    <rule context="Jazyk">
      <assert test="Navez or Rok">Elementy
chybí.</assert>
      <report test="Navez and Rok">Elementy jsou
přítomny.</report>
    </rule>
  </pattern>
</schema>
```



schematron

A language for making assertions about patterns found in XML documents

XML Schema – předchůdci

- Document Definition Markup Language (DDML, dříve XSchema) - velice jednoduchá aplikace, která znamenala převedení DTD do podoby kompatibilní s XML. To znamená, že tento jazyk také nepodporuje různé datové typy. Hlavní využití je konverze mezi různými DTD. Editor formátu DDML jsou Ronald Bourret, John Cowan, Ingo Macherius a Simon St.Laurent (leden 1999).

XML Schema – předchůdci

- XML-Data - tento jazyk pochází z dílen firem Microsoft, DataChannel, Arbotext, Inso Corporation a University of Edinburgh (W3C Note - leden 1998). Šlo o jeden z prvních pokusů vytvořit schématický jazyk na bázi XML (i když vazba na XML nebyla striktní). Jazyk obsahoval možnost definování interních a externích entit, objektově orientovanou strukturu a možnost mapování pomocí RDF (Resource Description Framework).
- XML-Data Reduced (XDR) - zjednodušená verze jazyka XML-Data, která vznikla v roce 1998 (W3C Note - červenec 1998). Tento jazyk vyvinula firma Microsoft a University of Edinburgh. XDR je používán ve všech aplikacích firmy Microsoft a díky jejímu velmi silnému postavení na trhu se XDR stalo druhým nejpoužívanějším schématickým jazykem po DTD.

XML Schema – předchůdci

- Schema for Object-oriented XML (SOX) - formát, který používá například firma Commerce One má oproti XDR některé vylepšené vlastnosti - dědičnost nebo možnost sdílení a dalšího využívání jednotlivých částí schémat. Je patrný vliv DTD. Jazyk vytvořily společně firmy Commerce One a Veo Systems - W3C Note, září 1998; druhá verze červenec 1999.
- Document Content Description (DCD) - na vzniku aplikace se podílely společnosti Microsoft, Textuality a IBM (W3C Note - červenec 1998). Jedná se podmnožinu XML-Data.

RELAX NG – předchůdci

- RELAX (Regular Language for XML, Regular Language Description for XML) - jednoduchý jazyk založený na matematické teorii hedge automatů aplikovanou na XML stromy byl vytvořený v roce 2000 jako japonský ISO standard (autor Murata Makoto).
- TREX (Tree Regular Expressions for XML) - byl vyvinutý Jamesem Clarkem v lednu 2001. TREX je založený na bázi XDuce a XML syntaxe. TREX se snadno čte, téměř jako běžný text.
- XDuce - jazyk vytvořený v březnu 2000. Nejde o schéma, ale spíše o regulérní programovací jazyk.

Ostatní schémata

- Document Structure Description (DSD) - zajímavá aplikace, která vznikla ve firmě AT&T. Použitím DSD dojde nejen k automatické kontrole dokumentu, ale také k přetransformování původního XML do jiného XML dokumentu, do kterého jsou na místo chybějících hodnot doplněny implicitní hodnoty.
- DT4DTD, DTD++
- Examplotron, Hook, DocBook NG
- Namespace Routing Language (NRL)

Jaký schémový jazyk?

- **DTD** - nepotřebujeme-li využívat jmenné prostory a datové typy.
- **W3C XML Schema** - potřebujeme-li pracovat se jmennými prostory a datovými typy a zároveň chceme využívat převážně komerční aplikace.
- **RELAX NG** - potřebujeme-li pracovat se jmennými prostory a datovými typy a zároveň chceme využívat převážně open source aplikace.

GML 3.2.1

XHTML 2.0

SVG 1.1

XSLT 2.0

XHTML 1.0

HTML 4.0

DocBook 5.0

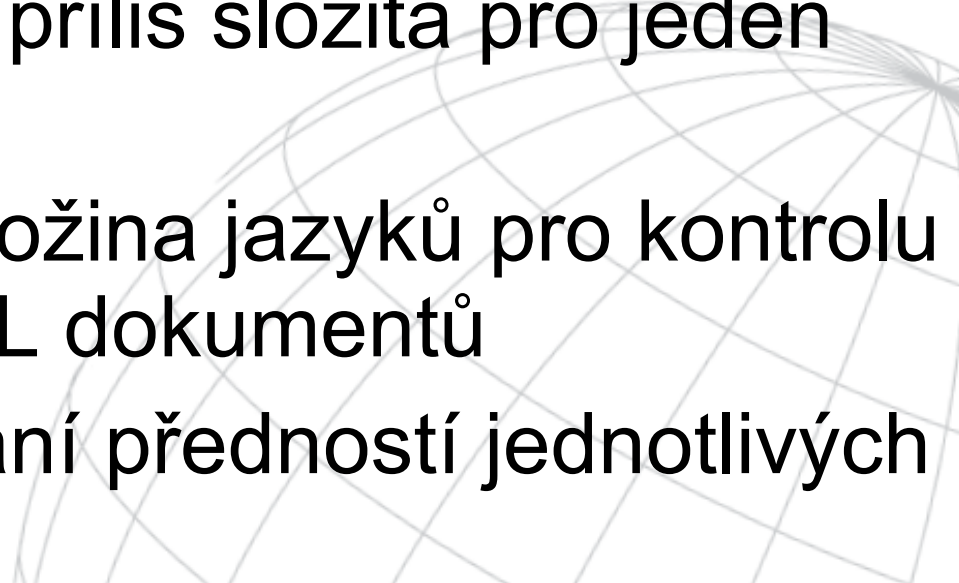
SVG Tiny 1.2

Existuje jen jedna validace?

- Oblast schémových jazyků je poměrně komplikovaná.
- Dokumenty obsahující několik jmenných prostorů, je potřeba validovat nejen pomocí několika schémat, ale také pomocí několika schémových jazyků.
- Uživatel požaduje kontrolu gramatiky, „business rules“, datových typů, integrity mezi jednotlivými prvky dokumentu, znakových sad apod.
- Vzhledem k neexistenci ideálního jazyka pro popis typu dokumentu je v současnosti téměř nutné vytvoření prostředí (rámce) pro komplexní validaci XML dokumentů.

Document Schema Definition Languages (DSDL)



- <http://dsdl.org/>
 - Standardizováno ISO – ISO/IEC 19757
 - Validace dokumentů je příliš složitá pro jeden schémový jazyk
 - DSDL = prostředí a množina jazyků pro kontrolu kvality a správnosti XML dokumentů
 - Hlavní přínos = využívání předností jednotlivých validačních technologií
- 

Komponenty DSDL

- **Regular-grammar-based Validation – RELAX NG; struktura dokumentu**
- **Rule-based Validation – Schematron; business rules**
- **Selection of Validation Candidates – Namespace-based Validation Dispatching Language (NVDL); dekompozice dokumentu**
- **Datatypes – Data Type Library Language (DTLL); alternativa k W3C XML Schema**
- **Path-based Integrity Constraints – ve vývoji; integrita mezi elementy a atributy**

Komponenty DSDL

- Character Repertoire Validation – Character Repertoire Description Language (CRDL); specifikace používaných znaků
- Declarative Document Architectures – Document Schema Renaming Language (DSRL), jazyk, který dokáže vzájemně mapovat XML struktury
- Namespace and Datatype-aware DTDs – ve vývoji
- Validation Management – stmelující část standardu – xvif/outie (XML Validation Interoperability Framework)

Výsledek DSDL

- Výsledkem by měla být kodifikace schémových jazyků, například Schematronu nebo RELAX NG, do podoby ISO norem a především vytvoření jednotného prostředí pro validaci XML dokumentů.
- Cílem DSDL je vytvoření rámce, v němž by bylo možné validovat jeden XML dokument pomocí více validačních prostředků.
- Hlavním přínosem bude především využívání předností jednotlivých validačních technologií.

RELAX NG (RNC)

```
<skola>  
  <nazev_VS/>  
  <stat/>  
  <obor/>  
  <studovane_predmety/>  
  <studium_v_jazyce/>  
  <typ_studia/>  
  <pocet_semestru/>  
</skola>
```

```
element skola {  
  element nazev_VS {empty},  
  element stat {empty},  
  element obor {empty},  
  element  
    studovane_predmety  
    {empty},  
  element studium_v_jazyce  
    {empty},  
  element typ_studia {empty},  
  element pocet_semestru  
    {empty}  
}
```

TRANG

James Clark




RELAX NG (RNG)

```
<?xml version="1.0" encoding="UTF-8"?>
<element name="skola" xmlns="http://relaxng.org/ns/structure/1.0">
  <element name="nazev_VS" <empty/> </element>
  <element name="stat" <empty/> </element>
  <element name="obor" <empty/> </element>
  <element name="studovane_predmety" <empty/> </element>
  <element name="studium_v_jazyce" <empty/> </element>
  <element name="typ_studia" <empty/> </element>
  <element name="pocet_semestru" <empty/> </element>
</element>
```


DTD

```
<!ELEMENT skola  
  (nazev_VS,stat,obor,studovane_predmety,studium_v_jazyce,  
    typ_studia,pocet_semestru)>  
<!ELEMENT nazev_VS EMPTY>  
<!ELEMENT stat EMPTY>  
<!ELEMENT obor EMPTY>  
<!ELEMENT studovane_predmety EMPTY>  
<!ELEMENT studium_v_jazyce EMPTY>  
<!ELEMENT typ_studia EMPTY>  
<!ELEMENT pocet_semestru EMPTY>
```



W3C XML Schema

```
<?xml version="1.0" encoding="UTF-8"?>
<xs:schema
  xmlns:xs="http://www.w3.org/2001/XMLSchema"
  elementFormDefault="qualified">
  <xs:element name="skola">
    <xs:complexType>
      <xs:sequence>
        <xs:element ref="nazev_VS"/>
        <xs:element ref="stat"/>
        <xs:element ref="obor"/>
        <xs:element ref="studovane_predmety"/>
        <xs:element ref="studium_v_jazyce"/>
        <xs:element ref="typ_studia"/>
        <xs:element ref="pocet_semestru"/>
      </xs:sequence>
    </xs:complexType>
  </xs:element>
```



```
<xs:element name="nazev_VS">
  <xs:complexType/>
</xs:element>
<xs:element name="stat">
  <xs:complexType/>
</xs:element>
<xs:element name="obor">
  <xs:complexType/>
</xs:element>
<xs:element name="studovane_predmety">
  <xs:complexType/>
</xs:element>
<xs:element name="studium_v_jazyce">
  <xs:complexType/>
</xs:element>
<xs:element name="typ_studia">
  <xs:complexType/>
</xs:element>
<xs:element name="pocet_semestru">
  <xs:complexType/>
</xs:element>
</xs:schema>
```

