Voorstel: Basis Datatypen baseren op internationale standaarden (versie 1.0 2014-08-29)

Inleiding

Korte beschrijving van het onderwerp.

Deel 1: Basisdatatypen.
Basisregistraties zijn semantische standaarden. Per registratiedomein wordt de semantiek (betekenis) van begrippen beschreven en gekoppeld aan de syntax (het begrip). Voor efficiëntie en hergebruik is het van belang dat er harmonisatie is tussen registraties. Eén van de onderdelen waarop relatief eenvoudig gestandaardiseerd kan worden is de semantiek en syntax van datatypen en specifiek de basis waardetypen. Voorbeelden hiervan zijn 'numerieke waarde', 'alfanumerieke waarde', 'geheel getal', 'datum'. Internationaal zijn de equivalenten hiervan gestandaardiseerd. Dit voorstel omvat het aansluiten op de internationaal gestandaardiseerde termen.

Deel 2: Restricties op datatypen.
Een tweede onderdeel van het voorstel behandelt een manier voor het beschrijven van restricties op numerieke en alfa-numerieke datatypen.

Deel1: Basis datatypen.

Analyse

Korte beschrijving van de huidige situatie.

Huidige situatie.
Op dit moment is er geen expliciete afspraak over het gebruik van standaard datatypen. Elke standaard gebruikt zijn eigen definities. Een bekend voorbeeld is het format waarin het tijdstip datum wordt beschreven. Maar ook waardetypen als 'getal', 'alfanumeriek', 'getal met max drie decimalen', 'N4', 'AN15' zijn termen die gebruikt worden. In de regel wordt het correct overbrengen van de semantiek van hetgeen men wil beschrijven wel gerealiseerd maar het is geen garantie voor eenduidige interpretatie. Zeker waar de semantiek in berichten geïmplementeerd wordt met hun eigen (XML) syntax is een eenduidige terminologie van belang.

Internationaal zijn de basis datatypen gestandaardiseerd. Semantiek en syntax zijn vastgelegd. Het is een relatief eenvoudige stap het gebruik van deze typen als uitgangspunt te nemen voor waardetypen in stelselstandaarden.

Bijkomend voordeel:
De relatie met internationale standaarden wordt hierdoor op een praktische wijze op de kaart gezet en gestimuleerd.

Voorstel

Korte beschrijving van het voorstel.

Regel:
Stelselstandaarden maken voor basis datatypen (of waardetypen) daar waar mogelijk gebruik van internationale standaarden. Dit geldt voor de semantiek, de syntax en de implementatie (encoding) daarvan in berichtenverkeer.

Het is niet zinvol om in dit voorstel een lijst van gestandaardiseerde datatypen op te nemen. Er wordt verwezen naar de standaarden die daar betrekking op hebben.

1) Semantiek en syntax.
Semantiek en syntax worden vastgelegd in de semantische modellen van de basisregistraties. Voor de standaard datatypen zijn onderstaande internationale standaarden van belang.

- ISO/IEC 19501:2005 Information technology – Open distributed Processing Modelling Language (UML) Version 1.4.2
- ISO/IEC 11404: Information technology – General Purpose Datatypes (GPD)
- ISO 19107: Geographic information – Spatial schema

2) Implementatie in berichtenverkeer.
Conform de in de basisregistraties vastgelegde semantiek worden datasets gecreëerd en worden data via berichtenverkeer uitgewisseld. Bij de implementatie van semantische datatypen in berichtenverkeer gelden andere standaarden. Op XML gebaseerd berichtenverkeer is het basisuitgangspunt. Een voor de hand liggende regel is dat voor datatypen dan ook de overeenkomstige XML typen worden gebruikt. De twee relevante standaarden zijn:

- W3C XML schema standaard: XML Schema Part2 Datatypes. (http://www.w3.org/TR/2012/REC-xmlschema11-2-20120405/) en daarin: primitive datatypes: http://www.w3.org/TR/2012/REC-xmlschema11-2-20120405/#built-in-primitive-datatypes
- ISO 19136: Geographic information - Geography Markup Language

Enkele voorbeelden van gestandaardiseerde datatypen staan in de bijlage.

Datum Tijd als voorbeeld:
Het xml datatype voor datum en tijd is dateTime: http://www.w3.org/TR/2012/REC-xmlschema11-2-20120405/#dateTime

dateTime volgt het format yyyy-mm-ddThh:mm:ss+-hh:mm. Het laatste deel '+-hh:mm' geeft de tijdzone relatief ten opzichte van UTC (Coordinated Universal Time of (bijna) Greenwich Mean Time). Aangeven van de tijdzone is optioneel maar geeft wel universele duidelijkheid over een tijdstip.
Het voorbeeld 2014-08-28T12:00:00+01:00 geeft het moment 28 augustus 2014 om 12 uur aan in de Nederlandse tijdzone.

Deel 2: Restricties op numerieke en alfanumerieke waardetypen.

Analyse:

Op de datatypen voor numerieke en tekstwaarden is het vaak van belang om restricties op aantal cijfers en karakters of type karakters vast te leggen in het semantische model. In de modellen van de basisregistraties worden daar op dit moment verschillende methoden voor gebruikt. Dit is verwarrend in de communicatie naast dat de definities niet altijd sluitend zijn .

Oplossing:

Voorgesteld wordt om hier één systematiek voor te gebruiken. De methodiek van reguliere expressies is hiervoor de meest voor de hand liggende keuze. Voor informatie over reguliere expressies zie: http://www.regular-expressions.info/

De volgende lijst zijn voorbeelden van restricties op datatypen zoals die nu beschreven worden en hoe ze als reguliere expressie opgenomen kunnen worden.

Huidige notatie	Tekst	Voorstel: Reguliere expressie
AN40	Maximaal 40 alfanumerieke karakters. Hoofdletters, kleine letters of cijfers.	[A-Za-z0-9]{1,40}
N3 (-99 tot +999)	Geheel getal tussen -100 en 1000	
N1 (-5 tot +5)	Geheel getal tussen -6 en 6	[-+]?[0-5]
N6 (0 – 999999)	Geheel getal tussen -1 en 1000000	^\d{0,6}$
Formaat: AN6 Waardenverzameling: 1000AA – 9999ZZ	Postcode format: 4 cijfers en 2 hoofdletters.	[1-9]{1}[0-9]{3}[A-Z]{2}

In de implementatie hiervan moet rekening gehouden worden met de leesbaarheid van reguliere expressies voor gemiddelde lezers. Het kan daarom zinvol zijn om de reguliere expressie te combineren met een beschrijvende tekst.

Bijlage: Voorbeelden van gestandaardiseerde datatypen.

Waardetype/syntax	**XML implementatie**
CharacterString	xs:string
Number	xs:double
Integer	xs:integer
Real	xs:double
Boolean	xs:boolean
Date	xs:date
Time	xs:time
DateTime	xs:dateTime
Day	xs:gDay
Month	xs:gMonth
Year	xs:gYear
GM_Point	gml:Point
GM_MultiPoint	gml:MultiPoint
GM_Curve	gml:Curve
GM_Point	gml:Point
GM_MultiPoint	gml:MultiPoint

Bijlage: Over dit voorstel van het project 'Utrecht'

Dit voorstel is een product van het project Utrecht. De aanleiding voor het project Utrecht was de behoefte van basisregistraties aan een verzameling technische afspraken die zij kunnen toepassen bij de berichtuitwisseling met al hun afnemers. Deze verzameling afspraken heeft de naam 'Gemeenschappelijk Afspraken Berichtstandaarden' (GAB) gekregen. Het idee van deze GAB is dat zowel de basisregistraties als de drie sectorstandaarden StUF, SuwiML en NEN 3610 en, indien van toepassing, ook Digikoppeling zich aan de GAB gaan houden, zodat binnen de Nederlandse overheid harmonisatie van berichtuitwisseling op technisch niveau ontstaat.

Het Project Utrecht heeft vijf voorstellen opgeleverd voor elementen van de GAB. De voorstellen zijn opgesteld door een werkgroep bestaande uit vertegenwoordigers van NEN 3610, StUF, SuwiML, Digikoppeling en een aantal basisregistraties, met ondersteuning vanuit het Bureau Forum Standaardisatie (BFS) en de stelselarchitecten van het iNUP-programma.

De vijf voorstellen zijn een selectie van 'laaghangend fruit' uit een lijst met potentiële gemeenschappelijke afspraken. De verwachting is dat deze vijf voorstellen relatief eenvoudig te adopteren zijn door de basisregistraties en de drie sectorstandaarden. De voorstellen zijn door de werkgroep van het project Utrecht breed uitgezet ter review. De definitieve versies van de voorstellen worden aangeboden aan de beheerders van de drie sectorstandaarden en bij de basisregistraties met het verzoek ze als wijzigingsverzoek in behandeling te nemen.

De sturing op het project Utrecht lag tijdens het i-NUP-programma bij de stuurgroep Digikoppeling, namens de Programmaraad Stelsel van Basisregistraties (PSB). Voor het vervolg op het project Utrecht na i-NUP wordt een Federatief Overleg opgericht waarin naar verwachting KING, BKWI, Geonovum, Logius en een aantal basisregistraties zitting zullen nemen. Dit Federatief Overleg wordt verantwoordelijk voor de sturing van het vervolg op het project Utrecht.