

Voorstel: Basis Datatypen baseren op internationale standaarden (versie 2014-02-28)

Inleiding

Korte beschrijving van het onderwerp.

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.

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 voor stelsel standaarden met betrekking tot gebruik basis datatypen (waardetype):
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.

Dit zijn in feite drie standaardisatie onderdelen over hetzelfde onderwerp. Omdat ze echter sterk met elkaar te maken hebben zijn ze verwoord in één voorstel.

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/xmlschema-2/>) en daarin: primitive datatypes:
<http://www.w3.org/TR/xmlschema-2/#built-in-primitive-datatypes>
- ISO 19136: Geographic information - Geography Markup Language

In principe zou een verwijzing van de stelselstandaard naar deze standaarden voldoende moeten zijn. Om de toepassing in meer detail uit te leggen is een tabel gemaakt van internationaal gestandaardiseerde typen die nu relevant zijn voor de basisregistraties.

Onderstaand is een niet limitatieve opsomming van internationaal gestandaardiseerde typen (ter completering):

Semantiek			XML implementatie
Waardetype/syntax	Definitie	Nederlandse term	
Numeriek:			
Number	Supertype of Decimal, Real and Integer.		xs:double
Decimal	An exact decimal data type. NOTE This differs from float, as float is an approximate value and Decimal is exact.		xs:decimal
Integer	A signed integer number, the length of an integer is encapsulation and usage dependent.		xs:integer
Real	A signed real (floating point) number consisting of a mantissa and an exponent, the length of a real is encapsulation and usage dependent.		xs:double
Tekst:			
CharacterString	A CharacterString is an arbitrary-length sequence of characters including accents and special characters from repertoire of one of the adopted character sets.		xs:string
Waarheid:			
Boolean	A value specifying TRUE or FALSE		xs:boolean
Logical	A value specifying TRUE or FALSE or MAYBE (UNKNOWN).		not supported: Een implemenatie is een enumeratie met de waarden: true, false, maybe
Probability	A value between 0.0 and 1.0 to describe probability.		Not supported
Datum en Tijd:			
Date	A date gives values for year, month and day. Character		xs:date

	encoding of a date is a string which shall follow the format for date specified by ISO 8601. Format: YYYY-MM-DD (eg 1997-07-16)		
Time	A time is given by an hour, minute and second. Character encoding of a time is a string that follows the ISO 8601 format. Time zone according to UTC is optional.		xs:time
DateTime	A DateTime is a combination of a date and a time type. Character encoding of a DateTime shall follow ISO 8601.		xs:dateTime
Day	Format: DD = two-digit day of month		xs:gDay
Month	Format: MM = two-digit month (01=January, etc.)		xs:gMonth
Year	Format: YYYY = four-digit year (eg 1997)		xs:gYear
Collecties:	A collection type is a template type that contains multiple occurrences of instances of a specific type.		
Set	A Set is a finite collection of objects, where each object appears in the collection only once. A set shall not contain any duplicated instances. The order of the elements of the set is not specified.		not supported
Bag	A bag may contain duplicate instances. As with a Set, there is no specified ordering among the elements of a bag. Bags are most often implemented through the use of proxies or reference pointers.		xs:all ?
Sequence	A Sequence is a Bag-like structure that orders the element instances. This means that an element may be repeated in a sequence.		xs:sequence
Coördinaten:			
GM_Point		Punt. 0-dimensionale geometrie.	gml:Point
GM_MultiPoint		Multipunt. Verzameling van punten die gezamenlijk één object vormen. (instanties van GM_Point).	gml:MultiPoint
GM_Curve		Lijn. 1-dimensionale geometrie.	gml:Curve
GM_MultiCurve		Multilijn. Verzameling van lijnen die gezamenlijk één object vormen (instanties van GM_Curve).	gml:MultiCurve
GM_Surface		Vlak. 2-dimensionale geometrie.	gml:Surface
GM_MultiSurface		Multivlak. Verzameling van vlakken die gezamenlijk één object vormen (instanties van GM_Surface).	gml:MultiSurface
GM_Solid		Volume. 3-dimensionaal geometrietype.	gml:Solid
GM_MultiSolid		Multivolume. Verzameling van volumes die gezamenlijk één object vormen (instanties van	gml:MultiSolid

		GM_Solid).	
Etc. zie ISO 19107			

3) Restricties op waardetypen.

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. Voorgesteld wordt om hier één systematiek voor te gebruiken. De methodiek van reguliere expressies is hiervoor de meest voor de hand liggende keuze. 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.

Aanpak

Hoe komt de standaard voor dit onderwerp tot stand?

Wat moet er gebeuren? Wie gaat het doen en hoeveel inspanning kost het?

Deze standaard is middels een bureaustudie opgesteld omdat er gebruik gemaakt wordt van bestaande internationale standaarden. Dit voorstel omvat ook al de invulling van dit onderwerp. Wat nu nog nodig is, is een review en daarna publicatie.