XML Datatype Constraints Creating Common Alerting Protocol (CAP) Alerts

Sean Donelan – sean@donelan.com

Robustness Principle (RFC1122)

> "Be conservative in what you send, be liberal in what you accept"

Suggested XML Data Type Constraints While Creating CAP Alerts

CAP creators should choose data values conservatively when creating IPAWS messages. CAP processors and consumers should be prepared to parse valid XML data types according to the CAP v1.2 schema and IPAWS v1.0 profile. Exceeding the following data type constraints should be considered warnings, not errors. Elements and values could be larger than the suggested maximum limits. For example, the CAP v1.2 schema allows an unlimited number of fractional digits in geographic coordinates. This document suggests alert origination software limit geographic coordinates to 6 digits after the decimal point. Nevertheless, a geographic coordinate with 7 or more digits after the decimal point would be valid, although a bit nonsensical, according to the CAP v1.2 schema.

CAP alert messages are typically between 5 KiB to 10 KiB without embedded resource files. The maximum size of the XML encoded CAP alert should be less than 2 MiB including embedded resource files <derefUri>. Linked resource files <uri> should be limited to 1.5 GiB each, preferably less than 1.5 MiB. Specific MIME Types may have lower limits, e.g., two minutes for audio files. There is no size limit for linked Information URL <web> resources. Caution: Presidential messages are unlimited, but currently lack an IPAWS specification how those messages would be delivered.

The following table suggests XML data type restrictions when creating CAP messages for IPAWS, which are proper subsets of the datatypes allowed by the CAP v1.2 XML Schema Definition and IPAWS v1.0 Profile.

CAP datatype	Suggested XML data type restrictions while creating IPAWS messages	Element names
Code	<element name="valueName"> restriction base="xsd:NCName" maxLength=32 enumerated values, no spaces. <element name="value"> If <value> in a coded domain restriction base="xsd:token"	eventCode *, parameter *, geocode *

	maxLength=32 enumerated values If <value> in a text domain restriction base="xsd:string" maxLength=2048 (see limits for individual parameters) May include characters from most Unicode blocks allowed by XML.	
	restriction base="xsd:NCName" maxLength=32 enumerated values, no spaces.	**status**, **msgType**, **scope**, **category ***, responseType *, **urgency**, **severity**, **certainty**
	restriction base="xsd:token" maxLength=32 enumerated values	code *, restriction, audience
	restriction base="xsd:language" maxLength=32 enumerated values	language
Data	restriction base="xsd:base64Binary" maxLength=2097152 (1.5 MiB converted to Base64) Individual MIME types have lower suggested size limits, i.e., two minutes for MP3 audio files. Base 64 encoded elements can also contain line-breaks every 64 to 72 characters.	derefUri
	restriction base="xsd:base64Binary" length=28 (160 bits for SHA-1) Should be prepared for longer SHA digests if the CAP standard or IPAWS profile updated with later FIPS Secure Hash standards, e.g., SHA3-512 would be 88 characters in base-64.	digest
Group	restriction base="xsd:token" maxLength=2048	addresses, incidents

	Multiple addresses or incidents are separated by spaces as tokens. If one of the tokens contains a space that token must be enclosed with double quotes. For example: (abc “def ghi” jkl) contains three tokens: ‘abc’, ‘def dhi’, and ‘jkl’. Caution: CAP does not specify how to escape the double quote character in an address or incident token. Avoid using quote characters within an address or incident token.	
	restriction base=“xsd:token” format “([sender],[identifier],[sent])*[sender],[identifier],[sent]” maxLength=2048 (approximately 12 to 24 referenced messages) List identifying previous associated alert messages.	references
	restriction base=“xsd:token” format “([latitude],[longitude]){3,}[latitude],[longitude]” An enclosed geographic area within a simple closed polygon defined by an ordered set of coordinates separated by whitespace. Latitude and longitude values conforming to the WGS84 coordinate reference system. Each polygon contains up to 100 decimal coordinate pairs limited to 6 fractional decimal digits of precision. The polygon’s first and last coordinate pair must have the same value. A valid polygon has a minimum of four coordinate pairs without intersecting edges. Suggest using the fewest, usually less than 20, decimal coordinate pairs completely enclosing the entire alert area with minimum overshoot, and no undershoot.	polygon *
	restriction base=“xsd:token” format “[latitude],[longitude]&#20;[radius]” An enclosed geographic area within a given radius around a geographic point. Latitude and longitude values conforming to the WGS84 coordinate reference system. Each circle contains one decimal coordinate pair limited to 6 fractional decimal digits of precision; and	circle *

	the radius in decimal kilometers with up to 2 fractional decimal digits of precision.	
Identifier	restriction base="xsd:token" (excluding space, comma, & and <) pattern="[_A-Za-z0-9-]+" URL and filename safe alphabet (RFC4648) maxLength=64	**identifier**
	restriction base="xsd:token" (excluding space, comma, & and <) pattern="([.A-Za-z0-9-]+@)?[.A-Za-z0-9-]+" ([optional name@]domain) Internet domain name alphabet (RFC1035) maxLength=64	**sender**
	restriction base="xsd:token" maxLength=128	source, **mimeType**
	restriction base="xsd:anyURI" maxLength=1024 (suggested limit 128) Usually a Hypertext Transfer Protocol (HTTP or HTTPS) web address or URL. Avoid using URIs encoded with obfuscated, unsafe or percent-encoded characters.	web, uri
Integer	restriction base="xsd:positiveInteger" maxInclusive=1610612736 (1.5 GiB) Individual MIME types have lower suggested size limits, e.g., two minutes for MP3 audio files. The maximum is based on a two-minute 1080p H.264 video file, about 1.45GB; which is not currently used in IPAWS. URIs with unknown length, i.e., streaming data, should omit the "size" element. Do not use 0 or MAXINT to indicate unknown or unlimited.	size
Quantity	restriction base="xsd:decimal" totalDigits=6 maxInclusive=328000 (approximately 100 km above sea level) minInclusive=-328000 (approximately 100 km below sea level)	altitude, ceiling

	Measured in decimal (0.0) feet above (or below) sea level per [WGS84] datum. Not typically used in IPAWS.	
Text	restriction base="xsd:string" maxLength=2048 description + instruction <= 1800 bytes, which limits the number of characters depending on UTF-8 encoding. Multi-byte encoded characters reduce the number of available characters. May include characters from most Unicode blocks allowed by XML.	note, description, instruction
	restriction base="xsd:normalizedString" (a string that contains no carriage return, line feed or tab characters) maxLength=256 (see individual elements for suggested limits) May include characters from most Unicode blocks allowed by XML.	senderName, headline, contact
	restriction base="xsd:token" (a normalizedString that contains no leading, trailing or consecutive space characters) maxLength=256 (see individual elements for suggested limits) May include characters from most Unicode blocks allowed by XML.	**event, resourceDesc**, **areaDesc**
Time	restriction base="xsd:dateTime" pattern="\d\d\d\d-\d\d-\d\dT\d\d:\d\d:\d\d[+-]\d\d:\d\d" format "CCYY-MM-DDThh:mm:ssXzh:zm" Caution: the CAP v1.2 schema erroneously includes a comma in its pattern for time elements. However, that comma has no practical effect, because restrictions in XML Schema Definitions cannot expand what is allowed by a data type. Likewise, because the xsd:dateTime base data type only allows ASCII digits [0-9], the "\d" will only match ASCII digits.	**sent**, effective, onset, expires

Notes on string data types

Plain Text strings use the data types xsd:string and xsd:normalizedString in the CAP data types table above. Plain Text is a pure sequence of character codes; plain Unicode-encoded text is therefore a sequence of Unicode character codes. Plain Text does not contain rich-text markup. The CAP Language identifier in the Info block does not restrict Unicode characters

used in strings to a specific language. In a multicultural society, scripts and symbols are often used and shared between multiple languages such as proper names.

Coded Text strings use the data types xsd:token and xsd:NCName in the CAP data types table above. Coded Text is a structured string, usually containing enumerated values, used for machine processing, parsing or matching. Unicode characters used in Coded Text are not related to the Language identifier. The IPAWS Profile implies Coded Text elements should only use the Unicode Basic Latin character block (ASCII).

CAP creators should write Coded Text elements with only Unicode Basic Latin (ASCII) characters, except when a Coded Text element specifically allows characters from other Unicode blocks. There are a few semi-Coded Text elements, which use both reserved keywords processed as Coded Text and strings rendered in different languages as Plain Text. Semi-Coded Text elements can use characters from other Unicode blocks. The semi-Coded Text elements are Event Type <event>, Resource Description <resourceDesc>, Area Description <areaDesc>, and Parameter <value> in specified <valueName> domains.

Notes on languages and string data types

The CAP Alert block does not have a Language identifier. The CAP Alert block primarily contains Coded Text strings used for machine processing, except the Note element is a Plain Text string. The Note element is used for operator or system messages. The Note element should be written in a language understood by alert system operators. The Note element does not need to be in the same language as any associated Info blocks. The Note element is normally not rendered as part of a public alert but is not considered private.

The CAP Info block does have a Language identifier, as the preferred language for that block and sub-block(s). The CAP Info block, Resource sub-blocks and Area sub-blocks include both Coded Text strings used for machine processing and Plain Text strings used for rendering alerts to the public.

CAP creators should always write their own public alert text for the public in each Info block language using the Plain Text elements. CAP creators should not rely on exchange partners to translate Coded Text strings into default text for public alerts.

CAP creators should limit characters in Plain Text strings to characters used in each Info block language. Nevertheless, Plain Text strings can use the full XML Unicode character set, regardless of the Info block preferred language. Characters from multiple languages are sometimes needed for proper names of people or places, or multicultural words in an alert text.

CAP creators should write the following Plain Text elements using the preferred language of each Info block.

- Sender Name <senderName>
- Headline <headline>
- Event Description <event>
- Instructions <instructions>
- Parameter <valueName> = “CMAMtext”
- Parameter <valueName> = “CMAMlongtext”

Note: IPAWS currently processes CMAMtext and CMAMlongtest only from Info blocks with U.S. English (en-US) and U.S. Spanish (es-US). They can be included in Info blocks with other language identifiers but will not be forwarded to WEA. Each national variant of CMAS/WEA has different mandatory and alternate languages, e.g., Canada uses English (en-CA) and French (fr-CA).

CAP creators should use a controlled vocabulary for the following semi-Coded Text elements. The controlled vocabulary may have different language specific translations. For example, the Resource Description reserved keyword “EAS Broadcast Content” is used in all Info block languages. While general-purpose Resource Descriptions may use a language specific vocabulary, e.g., “Missing Person” and “Persona desaparecida.”

- Event Type <event>
- Resource Description <resourceDesc>
- Area Description <areaDesc>