ADDRESS VERIFY WORLD 2 NODE

This document will help one understand the steps needed to configure and use the new Address Verify (World 2) node.

Table of Contents
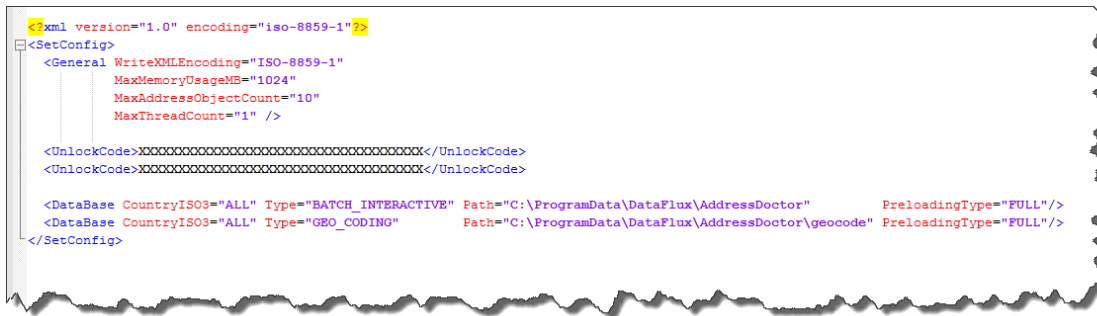
Configuration & Use

Configuration

Unpack the contents from dfpower-fix19534-win32.zip.

Modify the SetConfig.xml in the install/root/etc/ location. In this add the Address Doctor license code as well as the location of the data.

```xml
<?xml version="1.0" encoding="iso-8859-1"?>
<SetConfig>
    <General WriteXMLEncoding="ISO-8859-1"
             MaxMemoryUsageMB="1024"
             MaxAddressObjectCount="10"
             MaxThreadCount="1" />

    <UnlockCode>XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX</UnlockCode>
    <UnlockCode>XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX</UnlockCode>

    <DataBase CountryISO3="ALL" Type="BATCH_INTERACTIVE" Path="C:\ProgramData\DataFlux\AddressDoctor"         PreloadingType="FULL"/>
    <DataBase CountryISO3="ALL" Type="GEO_CODING"        Path="C:\ProgramData\DataFlux\AddressDoctor\geocode" PreloadingType="FULL"/>
</SetConfig>
```

There are some items worthy of note from the SetConfig.xml file. First in the General section:

- WriteXMLEncoding: Do not change this value from the default of "ISO-8859-1".
- MaxMemoryUsageMB: The maximum number of megabytes of memory which the engines is allowed to allocate dynamically
- MaxAddressObjectCount: Do not change this value from the default of "10".
- MaxThreadCount: Do not change this value from the default of "1".
- CacheSize: Using this attribute, the amount of memory reserved in such a way may be controlled - valid settings are NONE, SMALL, LARGE. Using the standard setting of "LARGE" is always recommended, unless all reference data needed is preloaded (so that "NONE" may be used) or the memory footprint needs to be reduced via the "SMALL" or "NONE" setting.

The unlock codes is the place for the AddressDoctor license. There is one line per Address Doctor license code provided.

The Database lines indicates

- The country that is to be loaded. "ALL" is a valid option and should be used for "BATCH_INTERACTIVE" type processing. If one country wants to be entered then the ISO alpha-3 codes should be used. An example of this would be "USA", GBR", "FRA", etc.
- The type of data. The only valid options recognized are BATCH_INTERACTIVE and GEOCODING.
- The path to the data. Where it resides on disk.
- How the data should be preloaded. Valid options are "FULL", "PARTIAL" and "NONE".

A quick note about preloading data. Partial preloading will load the metadata and indexing structures into memory. The reference data itself will remain on the hard drive. Partial preloading offers some performance enhancements and is an alternative when not enough memory is available to fully load the desired databases. Partial preloading may not be supported for all databases.

Full preloading will move the entire reference database into memory. This may need a significant amount of memory for countries with large databases such as the USA or the United Kingdom, but it will increase the processing speed significantly.

Since large amounts of memory may be allocated during preloading, with significant data amounts moved into memory, it might take some time to load the databases into memory. Databases will be preloaded in the order they are defined in the SetConfig.xml.

Here is a summary of tips to optimize the performance of validation:

- Install as much memory as possible to allow country databases to be fully pre-loaded into memory. At least as much memory as the size of the most often used country databases combined plus 256MB should be available. If all countries available from AddressDoctor are to be used simultaneously, add more memory to cover the entire size of all databases.
- Preload at least the databases of frequently used countries with the proper parameters set in the SetConfig.xml.
- When full preloading is not an option, store the database files on a fast hard disk or even better a SATA Solid State Disk (ideally exceeding 200MB/sec read transfer rate - for development purposes, high-speed USB or FireWire flash modules exceeding 30MB/sec read transfer rate might suffice). Especially the access latency (average seek time) should be minimized: Internal AddressDoctor benchmarks for "PreloadingType=NONE" with an Intel X25M G2 SATA SSD have shown a typical performance increase of a factor 20.
- Keep the AddressDoctor reference databases on a separate hard drive. Read and write address data from other drives. Make absolutely sure to keep the database files defragmented, internal tests have shown that performance may easily decrease by as much as 35% when the files are heavily fragmented.
- The AddressDoctor engine is very data-intensive, with a significant amount of non-localized memory accesses during processing: As such, it greatly benefits from direct multi-channel memory access (e.g. via Quick Path Interconnect or HyperTransport) with high bandwidth and low latency, combined with large processor caches, such as found in top-of-the line server processors.
- Use high performance multi-core processors, like Intel Xeon X55xx/65xx/75xx and higher, AMD Opteron 24xx/84xx and higher or IBM POWER7 and higher. Provided there is enough memory available for full preloading, the processor clock frequency will directly determine the speed of address processing. See http://www.spec.org/cpu2006/results/rint2006.html for a comparison of integer processing throughput between different processor architectures.
- When running batch processes without having a sufficient amount of memory installed, try to process records ordered by country with intermittent re-initialization of AddressDoctor using
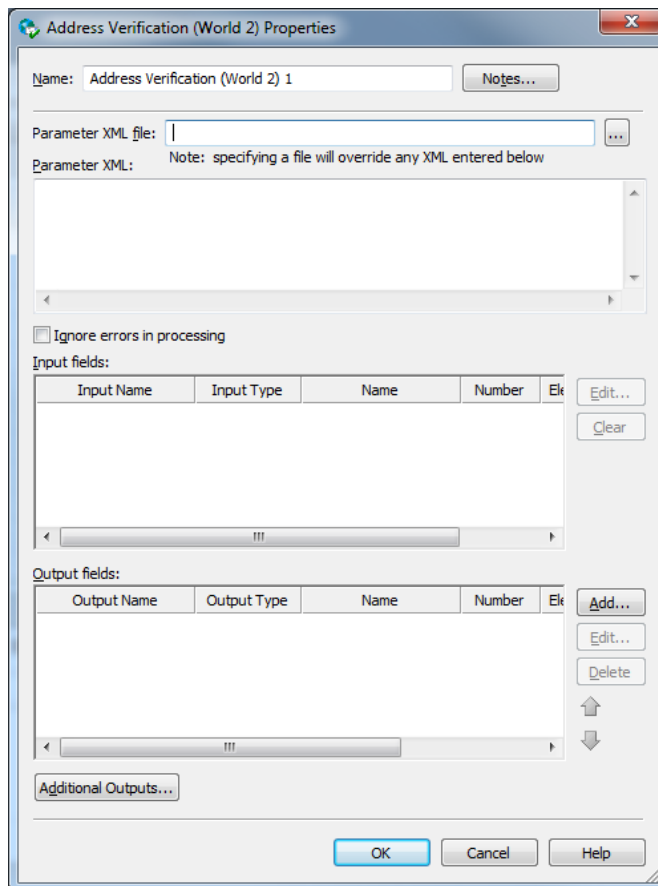
the appropriate pre-loading settings (see chapter 5.26). The engine will also benefit from internal and OS caches for addresses sorted by country as compared to addresses in random order, as they would for instance occur in a Web Service environment.

How to use

- Drop the node on the canvas.
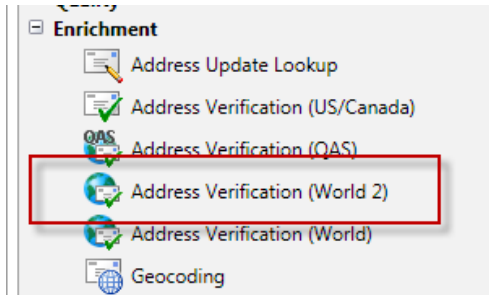
This will open up the properties.

- Set the "Parameters XML File" location equal to install/root/etc/parameters.xml. Or XML can be pasted into the "Parameter XML" box if something different needs to be used then the default. If a file is provided this is the only thing that will be used.
- Next map the input fields to the inputs needed by the node. This is done in the "Input Fields" list box

  o By default the inputs will already be listed. The user just needs to map these to an input type. This is done by selecting the "Edit" button.


  o The options for Input types are "ELEMENT", "LINE" and "COMPLETE".
    ▪ ELEMENT should be used if the input field represents a piece of the address.
    ▪ LINE should be used if the input field represents a line of the address
    ▪ COMPLETE should be used if the line represents the complete address.
  o Once an input type is selected, the "Name", "Number" and "Type" needs to be selected where permissible.
    ▪ The "Name" field will classify the input filed in such a way that the address verification engine knows how to use it. If the "Input Type" is defined as ELEMENT then examples for "Name" are Street, Number, Building, etc.
    ▪ The "Number" field helps classify the version of the input. The options change based on the "Name" that is picked. Users will probably be fine with leaving this as "1".
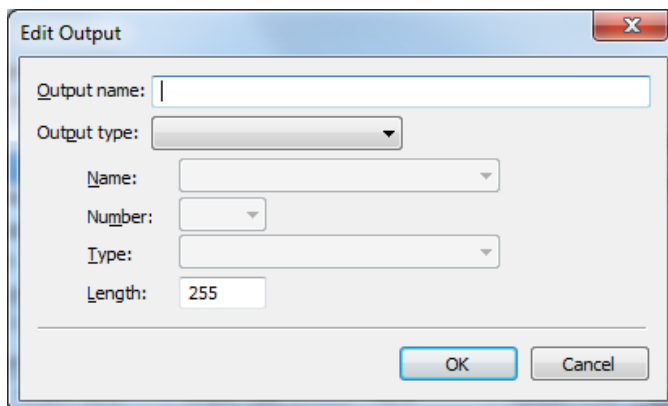
- The "Type" field helps define how the data looks. For example, if the input is Street, the "Type" field has the following options in it which helps define what data is in the input field.
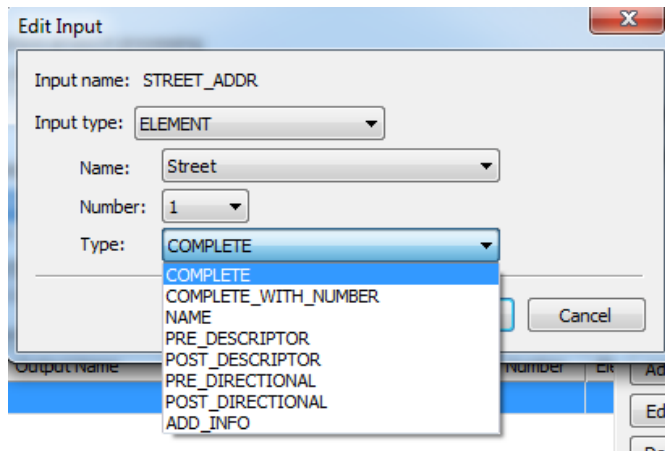
- Once all the inputs are defined, then the output needs to be selected. By default the outputs are blank and left to the user to fill in.
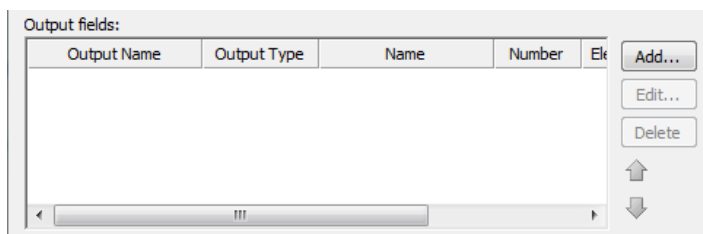
  o To add outputs, select the "Add" button.

- o Give the field a name
- o Select the type of output desired. The options are "XML", "PARAMETER", "DATAPARAMETER", "ELEMENT", "LINE", "COMPLETE", "ENRICH_DATAPARAMETER", "ENRICH_ELEMENT". Selecting one of these output types will drive the selection of Name, Number and Type.
  - XML will provide the entire response in its raw format
  - PARAMETER provides the data specific around how the data was processed.
  - DATAPARAMETER provides the specific values for response codes. Select these values to be able to determine how well the address was validated.
  - ELEMENT provides unique pieces of the address, like Street, Building Number, etc.
  - LINE provides all the elements that make up one line of the address.
  - COMPLETE provides the entire address in one field.
  - ENRICH_DATAPARAMETER is used to retrieve the status field in case any enrichment actions are allowed. These actions are driven by the AddressDoctor license.
  - ENRICH_ELEMENT is used to get the enrich data that the engine returned.
- o It is recommended during the initial design time to use XML outputs to get an understanding of the data that is being returned. From there, the outputs can be added that match what is desired.

When designing a job that will use the new World node, make sure that all selected output values will be returned for each address. If a field is not returned, then this could create an error condition when the job is running. An example of where this could happen is with Geocoding. If the Geocode output elements are included as output but a country has not been licensed for Geocoding results, the node will throw an error. Make sure that enough data has been run through the node before promoting a job to production.

Geocode

If Geocoding has been purchased from AddressDoctor modify the SetConfig.xml file to include the license and to point to the data and update the parameters.xml file to include the process mode of geocoding.

Status Values

Process Status Values

The process status values returned describe the result output quality of a call to AddressDoctor. This value is returned by selecting an output field that is of type PARAMETER and has the name "ProcessStatus":

V4	Verified – Input data correct - all (postally relevant) elements were checked and input matched perfectly
V3	Verified – Input data correct on input but some or all elements were standardised or input contains outdated names or exonyms
V2	Verified – Input data correct but some elements could not be verified because of incomplete reference data
V1	Verified – Input data correct but the user standardisation has deteriorated deliverability (wrong element user standardisation – for example, postcode length chosen is too short). Not set by validation.
C4	Corrected – all (postally relevant) elements have been checked
C3	Corrected – but some elements could not be checked
C2	Corrected – but delivery status unclear (lack of reference data)
C1	Corrected – but delivery status unclear because user standardisation was wrong. Not set by validation.
I4	Data could not be corrected completely, but is very likely to be deliverable – single result (e.g. One result found, but too poor for a C status. e.g. HNO is identified as wrong in batch mode.)
I3	Data could not be corrected completely, but is very likely to be deliverable – multiple results (e.g. HNO is wrong, several HNO generated in interactive mode.)
I2	Data could not be corrected, but there is a slim chance that the address is deliverable
I1	Data could not be corrected and is pretty unlikely to be delivered.
N1	Validation Error: No validation performed because country was not recognized
N2	Validation Error: No validation performed because required reference database is not available
N3	Validation Error: No validation performed because country could not be unlocked
N4	Validation Error: No validation performed because reference database is corrupt or in wrong format
N5	Validation Error: No validation performed because reference database is too old – please contact AddressDoctor to obtain updated reference data
Q3	FastCompletion Status – Suggestions are available – complete address
Q2	FastCompletion Status – Suggested address is complete but combined with elements from the input (added or deleted)
Q1	FastCompletion Status – Suggested address is not complete (enter more information)
Q0	FastCompletion Status – Insufficient information provided to generate suggestions
S4	Parsed perfectly
S3	Parsed with multiple results
S2	Parsed with Errors – Elements change position
S1	Parse Error – Input Format Mismatch
RB	Country recognized from Abbreviation
RA	Country recognized from ForceCountryISO3 Setting

R9	Country recognized from DefaultCountryISO3 Setting
R8	Country recognized from name without errors
R7	Country recognized from name with errors
R6	Country recognized from territory
R5	Country recognized from province
R4	Country recognized from major town
R3	Country recognized from format
R2	Country recognized from script
R1	Country not recognized - multiple matches
R0	Country not recognized

It is strictly recommended to only accept records with Vx or Cx status for automated data updates. Ix records need to be reviewed manually before using these results for any data update whatsoever.

Mailability Scores

AddressDoctor provides an estimate of how likely successful delivery of mail to an address might be. This is a simplification of the process status values and gives a measure to determine whether an address should be bothered with for mailing in a specific usage scenario. This is returned by selecting an output type of DATAPARAMETER with the name of "MailabilityScore":

- 0: Undeliverable
- 1: Risky
- 2: Fair Chance
- 3: Should Be Fine
- 4: Almost Certain
- 5: Completely Confident

Addresses with a mailability of 5 and 4 may always be considered for sending mail, while 0 or 1 should not be used independent of the scenario. Addresses marked with 2 or 3 may be used, but should be treated with caution: 2 indicates that the results are not corrected and therefore may still contain an incorrect address component. 3 indicates a correction which may require a review before sending the mail piece.

If there is a requirement to understand exactly what was validated or corrected in the address, the ProcessStatus value, ElementInputStatus and ElementResultStatus fields should be used instead of the MailabilityScore.

Score	Description	What it means
5	Completely Confident	All relevant elements of the address that have been entered were checked in the processing and have been verified in the process.
4	Almost Certain	An address is considered to be *Almost Certain* when one of the following two scenarios is present. Scenario 1: Some of the relevant elements of the address could not be checked due to reference data and the rest of the address have been verified in the process. Scenario 2: All relevant elements have been entered and some of the relevant elements of the address have been corrected in the process with a very high confidence. This only happens if the match was unique and the number of discrepancies was very low.
3	Should Be Fine	Some of the relevant elements of the address have been corrected in the process. A correction only happens if the match was unique and the number of discrepancies was acceptable.
2	Fair Chance	The address could not be corrected or validated in the process based on two scenarios. Scenario 1: A candidate match could not be made that had sufficient confidence. Scenario 2: The address matching ended with multiple candidates with similar confidence levels (multi-match situation). The input address, therefore, has a *Fair Chance* to be mailable as the relevant elements exist.
1	Risky	The address entered could only generate a partial match.
0	Undeliverable	The address entered is either missing too many components or a majority of the components could not be verified as they generate no matches against the reference data.

Element Status and Relevance Values

Element status values give a detailed explanation of the outcome of the validation operation. They are only meaningful after a validation operation has been performed, even though some information is available after a parsing operation for the "ElementInputStatus" value.

Twenty address elements are covered in both, "ElementInputStatus" and "ElementResultStatus". The former provides per element information on the matching of input elements to reference data, while the latter categorizes the result in more detail than the overview ProcessStatus values. These values are returned by setting an output type of DATAPARAMETER and a "Name" of "ElementInputStatus" and "ElementResultStatus".

The element positions (from left to right) are, where level 0 pertains to the Item 1 status information, while level 1 summarizes the status information on Items 2-6:

1	PostalCode level 0
2	PostalCode level 1 (e.g. ZIP+4 – Plus 4 addition)
3	Locality level 0
4	Locality level 1 (e.g. Urbanisation, Dependent Locality)
5	Province level 0
6	Province level 1 (e.g. Sub Province)
7	Street level 0
8	Street level 1 (e.g. Dependent street)
9	Number level 0
10	Number level 1
11	Delivery service level 0 (e.g. PO Box, GPO, Packstation, Private Bags)
12	Delivery service level 1
13	Building level 0
14	Building level 1
15	SubBuilding level 0
16	SubBuilding level 1
17	Organisation level 0
18	Organisation level 1
19	Country level 0 (Mother country)
20	Country level 1 (e.g. Territory)

Input Status

The possible values for validation are:

0	empty	
1	not found	
2	not checked (no reference data)	
3	wrong - Set by validation only	The reference database suggests that either Number or DeliveryService is out of valid number range. Input is copied, not corrected for batch mode.
4	matched with errors in this element	
5	matched with changes (inserts or deletes)	For example: • Parsing: Splitting of house number for "MainSt 1" • Validation: Replacing input that is an exonym or dropping superfluous fielded input that is invalid according to the country reference database
6	matched without errors	

For parsing, the following values are possible:

0	empty
1	element had to be relocated
2	matched but needed to be normalized
3	matched and OK

Result Status

Is only set after validation as an indication whether verification ("verified") or correction ("changed") were possible or not, the potential values are (for all address elements apart from country):

0	empty
1	not validated and not changed. Original is copied.
2	not validated but standardized.
3	validated but not changed due to invalid input, database suggests that number is out of valid ranges. Input is copied, not corrected – this status value is only set in batch mode.
4	validated but not changed due to lack of reference data.
5	validated but not changed due to multiple matches. Only set in batch mode, otherwise multiple suggestions that replace the input are marked as corrected (status value 7).
6	validated and changed by eliminating the input value
7	validated and changed due to correction based on reference data
8	validated and changed by adding value based on reference data
9	validated, not changed, but delivery status not clear (e.g. DPV value wrong; given number ranges that only partially match reference data).
C	validated, verified but changed due to outdated name
D	validated, verified but changed from exonym to official name
E	validated, verified but changed due to standardization based on casing or language.

Validation only sets this status if input fully matches a language alternative.

F	validated, verified and not changed due to perfect match

For Country (position 19 & 20), the following values are possible:

0	empty
1	Country not recognized
4	Country recognized from DefaultCountryISO3 setting
5	Country not recognized - multiple matches
6	Country recognized from script
7	Country recognized from format
8	Country recognized from major town
9	Country recognized from province
C	Country recognized from territory
D	Country recognized from name with errors
E	Country recognized from name without errors
F	Country recognized from ForceCountryISO3 setting

Relevance

In addition to the element status values described previously, information is available on which of the address elements of the address processed are actually relevant from the local postal operator's point of view. This is retrieved by adding an output type of DATAPARAMETER with a name of "ElementRelevance". The possible values for each address element are "1" for relevant and "0" otherwise.

For any given address, all address elements with a value of "1" must be present for an output address to be deemed valid by the local postal authority. "ElementRelevance" may well vary from address to address for countries with different address types, e.g. rural versus metropolitan addressing. Furthermore, AddressElements that have actually been validated against reference data (i.e. with a ElementResultStatus of 7 and higher) may override the default ElementRelevance value defined for that AddressElement.

Please note that "ElementRelevance" is really only meaningful for a "ProcessStatus" value of Cx or Vx.

ResultPercentage Values

The "ResultPercentage" value gives an indication how similar a result is to the parsed input, values close to 100% imply high similarity. They are mainly provided to allow for filtering out too extensive corrections in records in master data management environments with very stringent data quality requirements.
It is discouraged using "ResultPercentage" values for any other use case scenarios than the two described above. To get the value add an output type of DATAPARAMETER with a name of "ResultPercentage".

Geocoding Status Values

AddressDoctor 5 introduces Geocoding for selected countries: This means the Version 5 API will provide the option to enrich a validated address by the respective geo-coordinates in WGS84 (http://wikipedia.org/wiki/WGS84) format.
The quality of coverage will naturally vary from country to country and while AddressDoctor shall strive to provide geo-coordinates on house number or building level, depending on data availability, only street or even locality level geo-coordinates might be available.
The corresponding status values returned with the processing result via AD_GetResultXML()or AD_GetResultParameter() are:

EGCU: Geocoding database not unlocked
EGCN: Geocoding database not found
EGCC: Geocoding database corrupt
EGC0: No Geocode available
EGC1..3: Reserved for future use
EGC4: Geocode with partial postal code level accuracy (e.g. 795xx)
EGC5: Geocode with postal code level accuracy
EGC6: Geocode with locality level accuracy
EGC7: Geocode with street level accuracy
EGC8: Geocode with house number level accuracy (interpolated approximation)
EGC9: Reserved for future use