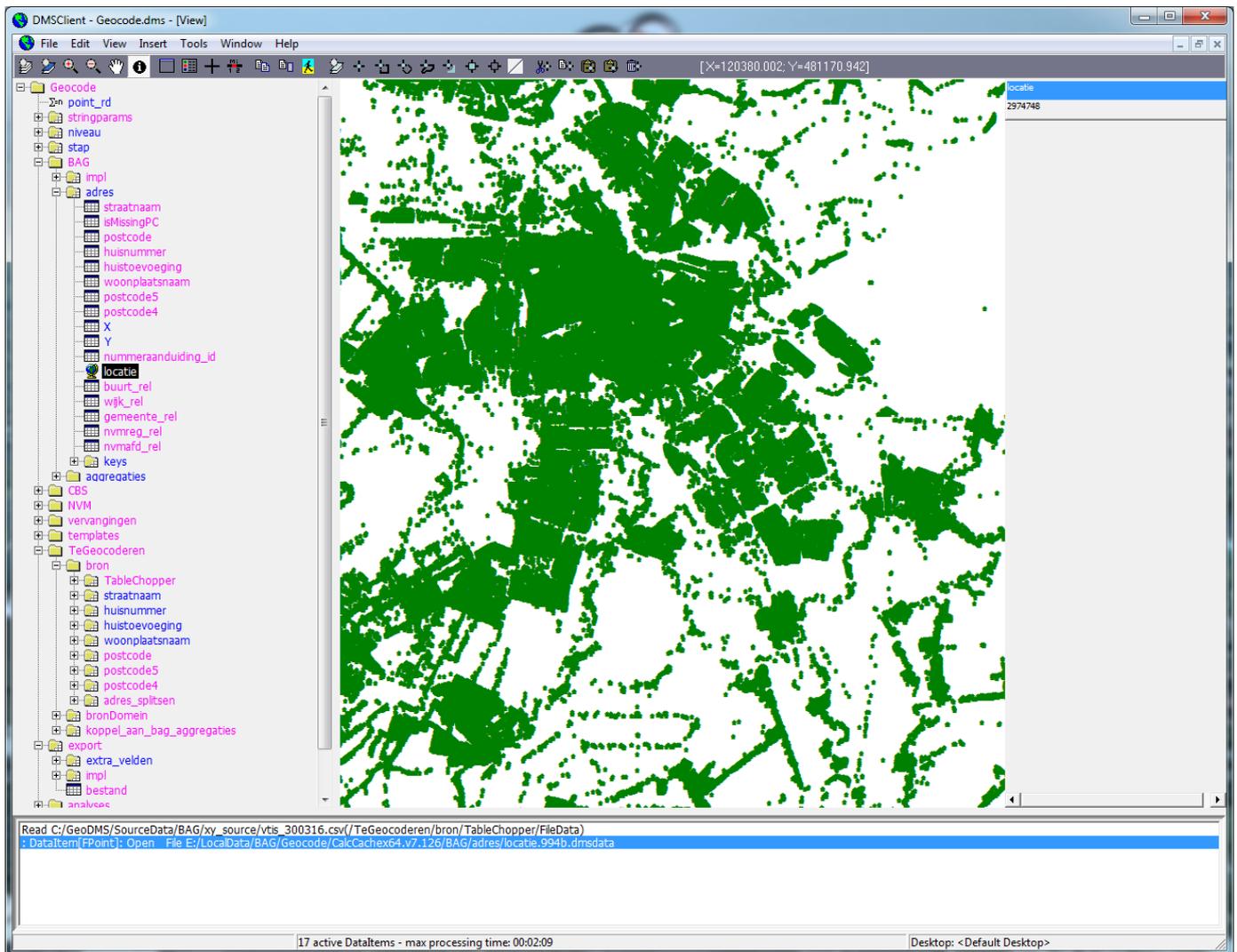


Geocoding with the GeoDMS using BAG data

Making Snapshots from and Geocoding against the Dutch Basic Administration of Addresses and Buildings (BAG)



Dr. Jasper Dekkers (Spatial Information Laboratory | VU)
and Martin van der Beek MSc. (Object Vision B.V.)

Vrije Universiteit Amsterdam

Table of contents

| | |
|--|----|
| Project information..... | 2 |
| Document information | 3 |
| Download links | 3 |
| Introduction | 4 |
| Licensing, citing and referring | 4 |
| Preparatory steps | 5 |
| Section 1 - Geocoding your data with an existing BAG-snapshot..... | 7 |
| Section 2 - Creating your own BAG-snapshot..... | 11 |
| 2A) XML2FSS | 11 |
| 2B) MakeSnapshot.dms | 12 |

Project information

| | | | |
|----------------------------------|--|--|---|
| Project name: | Geocoding address datasets with the GeoDMS using BAG Data | | |
| Project leader: | Jasper Dekkers (Vrije Universiteit) | Phone: Email: | 020-5982211 j.dekkers@vu.nl |
| Technical implementation: | Jasper Dekkers (Vrije Universiteit) | | |
| | Martin van der Beek (Object Vision BV) | Email: | mailto:mtbeek@objectvision.nl |
| Organisations: | Vrije Universiteit FEWEB-RE / SPINlab De Boelelaan 1105 1081 HV Amsterdam Room 9A-91 spinlab.vu.nl | Object Vision BV De Boelelaan 1085 1081 HV Amsterdam www.objectvision.nl | |

Document information

| Version – Date | Edited by | Remarks |
|-----------------------|--|--|
| 0.1 17-01-2016 | Martin van der Beek | |
| 0.2 07-09-2016 | Jasper Dekkers | Document format revised and contents updated to include latest changes |
| 0.3 13-09-2016 | Martin van der Beek and Jasper Dekkers | Final editing, addition of some technical background information, license information added |
| 0.5 16-09-2016 | Jasper Dekkers | Easy download option and tree structure added |
| 0.6 21-09-2016 | Jasper Dekkers | Minimum hardware requirements added |
| 0.7 16-03-2017 | Jasper Dekkers | Updated configuration, new BAG-snapshot added, CBS data updated, bounding box-functionality added, coordinate restriction on the Netherlands added |
| 0.8 07-04-2017 | Jasper Dekkers | Updated configuration, path to BAG-snapshot corrected, Province-data added in sourcedata and in export |

Download links

This manual, the software, the configuration and the source data can be downloaded via <https://spinlab.vu.nl/support/tools/geodms-bag-geocoder>.

More information about this toolkit, it's scripts and technical details involving the GeoDMS-software can be found at <http://www.objectvision.nl/geodms/projects/bag-toolkit>. This website also contains a downloadlink for the BAG-Geocoder configuration (or toolkit as it is called there), but without any of the source data.

Introduction

This manual explains how you can use the GeoDMS software in combination with a model configuration developed by us to geocode datasets on your own computer, locally, using the the Dutch Basic Administration of Addresses and Buildings (BAG).

The BAG is a database with actual and historic information about addresses and buildings. It is very useful for lots of purposes (e.g. geocoding, determining building types, calculating accessibility indicators). It is open data and is available as a set of XML files.

Using the GeoDMS-BAG configuration we developed, you can geocode your own csv-files with address data on your own local computer. next to the advantages that these local calculations are much faster than via a web service, especially for large datasets and the fact that it's for free, you also get lots of meta-information about the geocode-level of your data (we have included no less than 16 different levels). Both main addresses and secondary addresses ('nevenadres' in BAG-terminology, relates to an object with multiple entrances), are included in our BAG-snapshot data.

This manual describes all necessary steps to geocode your own data. If you have any questions, or you have suggestions for further developing the tooling, feel free to contact us. Since the configuration is open source, you are free to develop it further by yourself. We kindly request you to share your developments with the community, so that others may profit from it as well.

Licensing, citing and referring

The GeoDMS software is available under the following license conditions:
www.objectvision.nl/geodms/software/license-and-disclaimer.

The geocoding configuration is available under GNU-GPL version 3 license, see www.gnu.org/licenses/gpl-3.0.html for more information.

The included CBS shapefiles with administrative region boundaries of municipalities, quarters and neighbourhoods (in Dutch the 'Wijk en buurtkaart') are available under CC-BY 3.0 license. When the CBS shapefiles are used to visualize area boundaries or data per region, you are obliged to include the following reference: © Kadaster / Centraal Bureau voor de Statistiek, 2016.

The shapefile containing NVM-regions has been constructed by Jasper Dekkers from SPINlab and may be used without referring.

When you use our geocoding configuration to geocode your data and subsequently use this data in your work, we kindly request you to recognize our work and refer to our geocoding configuration in the following way:

Dekkers, J.E.C. and Van der Beek, M., *Geocoding with the GeoDMS using BAG data. Making Snapshots from and Geocoding against the Dutch Basic Administration of Addresses and Buildings (BAG)*, Vrije Universiteit/Object Vision B.V., Amsterdam, 2017.

Jasper Dekkers and Martin van der Beek
September 2016-April 2017

Preparatory steps

Before you continue with this manual, please make sure you have done the following things:

1. Check the **minimum hardware requirements**: The BAG is a large dataset and the geocode tool uses many aggregations. Therefore the tool needs sufficient internal memory. We experienced that on machines with 4Gb internal memory the tool becomes slow and might result in memory allocation errors.
We therefore advise to use the tool on 64 bits machines, with 64 bits versions of the GeoDMS and 16Gb internal memory or more. This also applies to the Xml2FSS and MakeSnapshot projects. Additionally, running the configuration on a Solid State Disk (SSD0 instead of a regular rotating hard disk will speed up the calculations as well;
2. Download and install the Geo Data and Model Server (GeoDMS) software (version 7.125 or higher) from the website of Object Vision (www.objectvision.nl/geodms/software);
3. Choose either the Easy or the Advanced download option:
 - **Easy download option**
With the Easy download option, you download one zip-file containing both the tool and all the source data. Extract the zip-file to *c:/GeoDMS/* and you are ready to go! Your directory installation on the C-drive now should look like the one depicted in Figure 1.

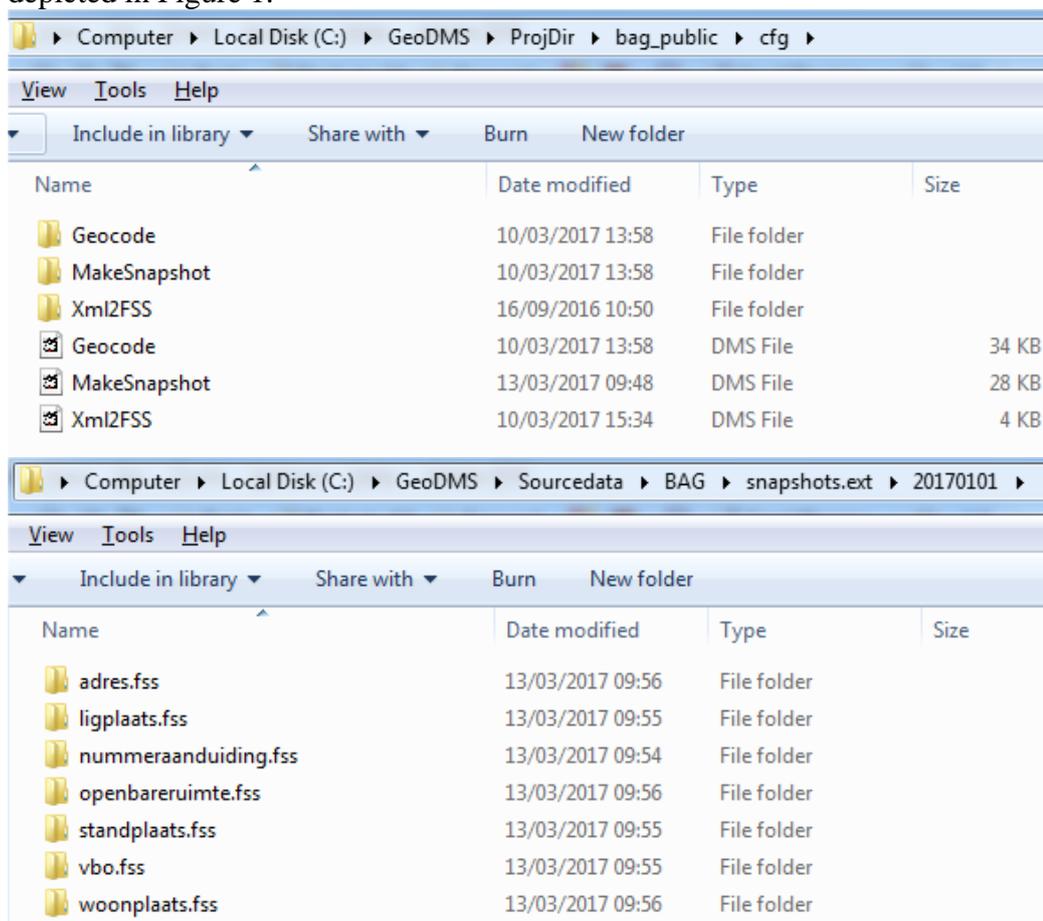


Figure 1 - Directory structures on your local hard drive (C:)

Now, read steps 6 and 7 on the next page and then continue to Section 1 for the Geocoding.

- **Advanced download option**

If you select the Advanced download option, we assume you already have one or more other GeoDMS model configurations on your computer and you want to be able to select the directories where you put the tool (ProjDir) and the data (Sourcedata) yourself. If you follow this path, carry out steps 3 to 5 as well. After having extracted the zip-files, you will have to adjust some path names in Bron.dms, Geocode.dms and MakeSNAPSHOT.dms. We assume you know how to do this.

4. Download the BAG-configuration from the website and extract it to a configuration directory. We suggest you use *c:/GeoDMS/ProjDir/BAG_public/cfg* for that;
5. Download CBS and NVM source data and extract the contents to SourceData directories. We suggest you use *c:/GeoDMS/Sourcedata/CBS/2016/* respectively *c:/GeoDMS/Sourcedata/NVM/2015* for that.
6. Download or create BAG-snapshot-data and administrative area data: Before you can use the GeoDMS software to geocode your data against the BAG source, two things need to be taken care of:
 - 1) transforming the BAG XML-files into the native data format of the GeoDMS and
 - 2) taking a snapshot of the BAG, i.e. an image of the BAG of a certain date.

For your convenience, we have already taken these steps for you and have made available for download a BAG-snapshot on our website. We have also included some administrative area data already in the configuration, i.e. the CBS municipalities, quarters and neighbourhoods plus the NVM housing market region areas.

You can copy the snapshot data (dated 1 January 2017, or 20170101 in GeoDMS format) into the data directory, then forget about the two steps described in sections 2A and 2B of this manual and immediately go to the geocoding process (Section 1). We suggest that you download the BAG-snapshot data and store it in a directory. We suggest you use *c:/GeoDMS/Sourcedata/BAG* for that (if you unzip the data in this directory, your fss-subdirectories will then effectively be stored in *c:/GeoDMS/Sourcedata/BAG/snapshots.ext/20170101*).

If, however, you want to create your own BAG-snapshot, for instance a snapshot from another date, you can follow the steps described in Section 2 (A and B) to do so.

Further steps

7. You can see (and edit) the path names for your SourceDataDir and your LocalData directory (where the results of your geocoding actions will be written to) via the GeoDMS GUI, menu-item Tools > Options > Paths (see Figure 2);
8. In the same Options-menu (see, again, Figure 1), you can also link to external programs, such as an ASCII/Text editor. In our case, we have linked to the Crimson Editor. You can use the same settings on your own computer. The Crimson Editor can be downloaded from: www.objectvision.nl/geodms/software/system-requirements.

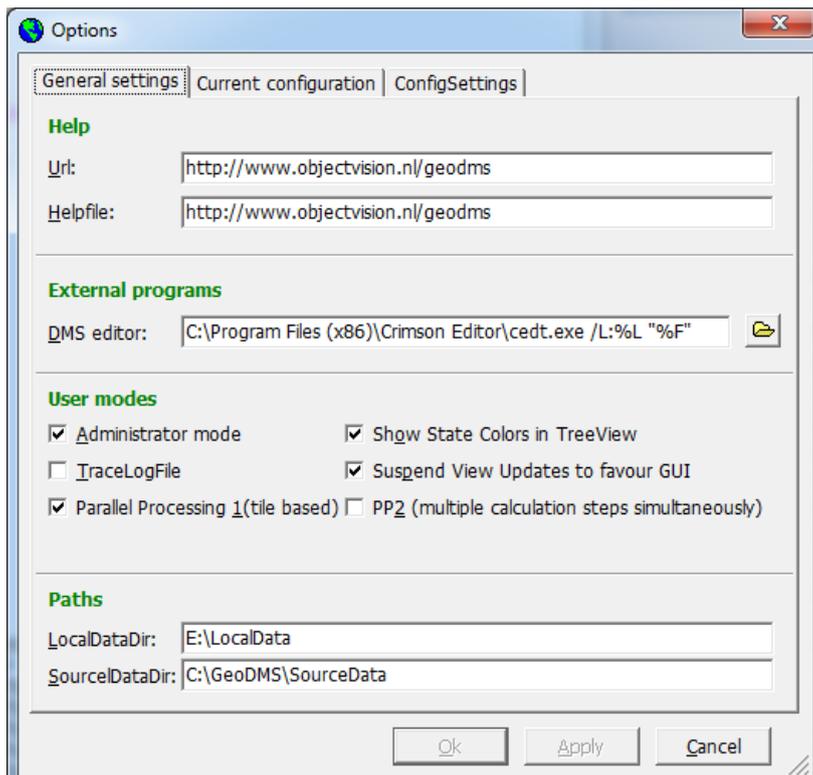


Figure 2 - Tools > Options in GeodMS GUI - setting Paths on your local hard drives

Section 1 - Geocoding your data with an existing BAG-snapshot

With the Geocode.dms script, a data set in CSV-format containing addresses can be geocoded using a BAG-snapshot.

1. Open the file *Geocode.dms* file from the subdirectory *%ProjDir%/BAG_public/cfg* in an ASCII/text editor (for example Notepad, Crimson Editor).
2. In the container *BAG > impl*, under the parameter *SnapshotDir* (see rule 113), set the correct BAG-snapshot reference by adjusting the date (the date format is *yyyymmdd*), see Figure 3. In the next lines in the file you can see that this scripts fetches the address table (*adres.fss*) from the snapshotdir and uses it for geocoding purposes.

```

Crimson Editor - [C:\GeoDMS\ProjDir\bag\cfg\Geocode.dms]
File Edit Search View Document Tools Macros Window Help
[Icons]
bron Geocode
94  >>> >>> ,1
95  >>> >>> ,0
96  >>> >>> ];
97  >>> >>> }
98
99  >>> unit<uint32> stap:= Subset(niveau/code > int8(0))
100 >>> {
101 >>>   attribute<int8>   code    := niveau/code[nr_OrgEntity];
102 >>>   attribute<string> volgorde := niveau/volgorde[nr_OrgEntity];
103 >>>   attribute<string> naam     := niveau/naam[nr_OrgEntity];
104 >>>   attribute<uint8>  aggtype  := niveau/aggtype[nr_OrgEntity];
105 >>>
106 >>>   attribute<string> expr_key := replace(volgorde,',','_' + '_' + ');
107 >>> }
108
109 >>> container BAG
110 >>> {
111 >>>   container impl
112 >>>   {
113 >>>     parameter<string> SnapshotDir:= '%SourceDataDir%/BAG/snapshots.ext/20160101';
114 >>>
115 >>>     unit<uint32> adres
116 >>>     := StorageName = "%SnapshotDir + '/adres.fss'"

```

Figure 3 - SnapshotDir-parameter in Geocode.dms

- Open the file *bron.dms* from the subdirectory `%ProjDir%/BAG_public/cfg/Geocode` in the ASCII/text editor. Next, in the container *TableChopper* adjust the file name-parameter (*BestandsNaam*, see line 21) so that it contains the name of the source data CSV-file you need geocoded. We normally put our data in the source data directory under BAG in a sub-directory called *xy_source* (`%SourceDataDir%/BAG/xy_source/`).
Note: sometimes, people get an error on the source-file stating that there's an illegal character in the first column name. This usually is solved by not saving the csv-source data in utf-8 format but in local system format instead.
- Adjust, if necessary, the field separation character (*VeldScheidingsTeken*-parameter) that you use in the CSV-file in *bron.dms* line 27 (see again Figure 4). By default, this character is set as comma-separated, but you can also use a semi-colon.

```

17 container bron
18 {
19   container TableChopper // Deze wordt gebruikt ipv de gdal.vect storage manager omdat de TableChopper
20   {
21     parameter<String> BestandsNaam := '%SourceDataDir%/BAG/xy_source/vtis_300316.csv';
22   }
23   parameter<uint32> nrElems      := strcount(FileData,stringparams/newline)-1;
24   unit<uint32>      Domein        := range(uint32,0,nrElems);
25 }
26 parameter<string> VeldScheidingsTeken := ','
27 IntegrityCheck = "VeldScheidingsTeken = ',' || VeldScheidingsTeken = ';'";
28

```

Figure 4 - FileName- and FieldSeparation-parameter in Bron.dms

- Check in rules 61 to 79 whether the column names for the different address components in your CSV-file (e.g. street, postcode, house number) match with the names of these components in these rules (*data/<ColumnName>*). When they do not exactly match, either adjust the *<ColumnName>* in the *bron.dms* file or adjust the column names in your CSV-file (the latter option being easier to adjust).

The default address names in the configuration are:

- *straat* (street)
- *huisnr* (house number)
- *huisnr_toev* (house number addition)
- *postcode* (zipcode)
- *plaats* (place of residence)

- Start the GeoDMS GUI and activate the item *export > bestand* (export file, see Figure 5).

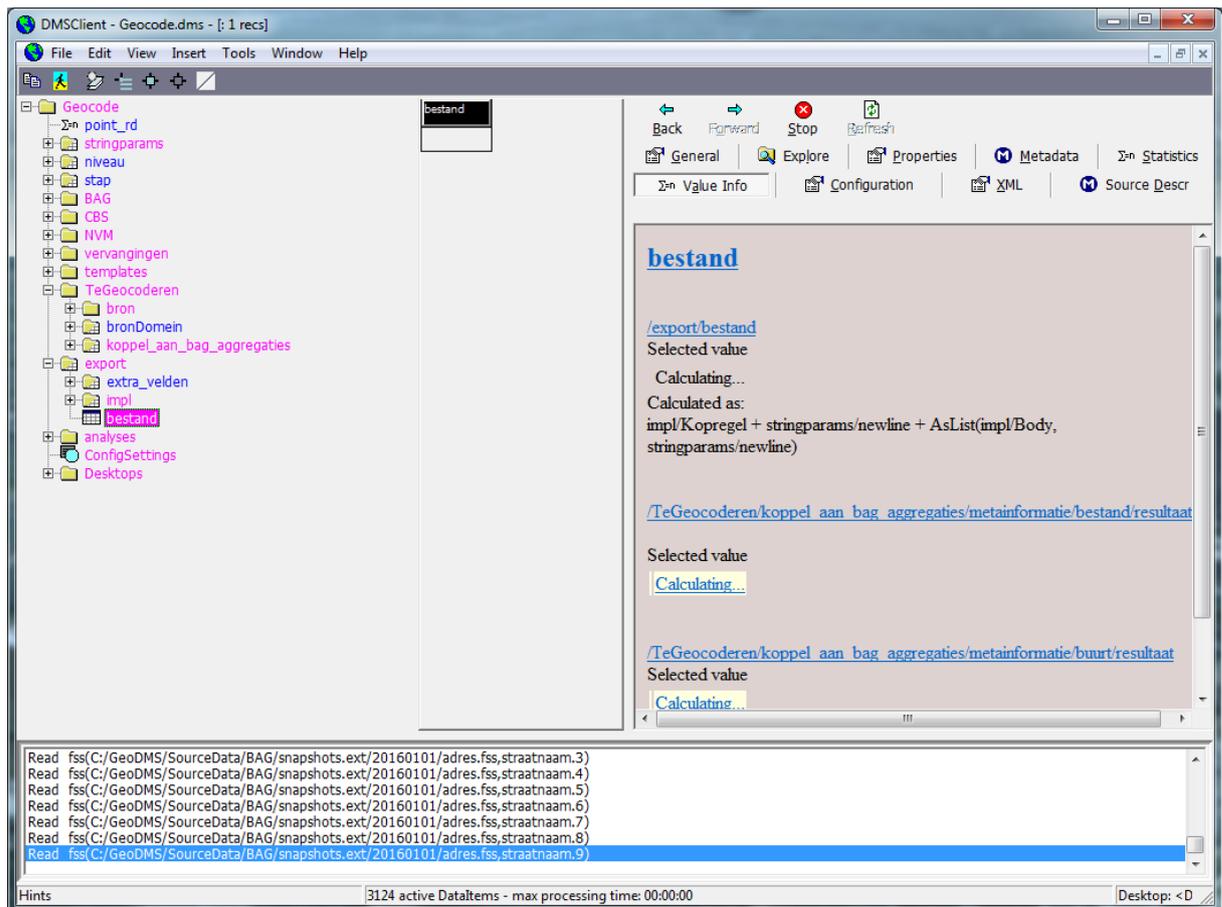


Figure 5 - Activate Export > Bestand in GeoDMS GUI

The geocoding process is now executed and a *result.csv* file will be written in the *LocalData*-directory that belongs to your BAG-configuration (in our case: *%LocalData%/bag/Geocode*). This file contains exactly the same data as your source data file, extended with a geocode level indication (ranging from 1 to 16), X- and Y-coordinates (in the RD_new coordinate system), BAG-address fields and some statistics relating to the geocoding process (e.g. standard deviation in case a coordinate has been calculated using multiple BAG addresses). the full list of added variables:

- CBS-administrative unit indications from 2016: *buurt_code* , *wijk_code*, *gemeente_code*
- NVM-administrative unit indication from 2015: *nvmafd_code*
- Geocode level: **niveau_code**, ranging from the best possible (#1 to only an average geocode on place of residence (#16).
- Coordinates in RD_New coordinate system: X, Y
- Statistical variables: *count*, *var* (variance)
- BAG-variables: *nummeraanduiding_id*, *bag_postcode*, *bag_straatnaam*, *bag_huisnummer*, *bag_huisletter*, *bag_toevoeging*, *bag_woonplaatsnaam*

Next to the *Resultaat.csv* file, a meta-file (*Resultaat_meta.csv*, see Figure 6) is created, containing meta-information about the quality of the geocodes:

- 'code' means the niveau-code variable and tells you what geocode level an observation has been geocoded on;
- 'omschrijving' describes the geocode level (i.e. indicates what address fields are included in this geocode level);
- 'aantal' gives the number of observations in your data set that have been geocoded on the different levels;

- 'perc' gives the cumulative percentage of observations)

| 1 | code | omschrijving | aantal | perc | gemiddeld aantal adressen | gemiddelde variantie |
|----|------|--|---------------|------------|---------------------------|----------------------|
| 2 | 1 | postcode straatnaam huisnummer huistoewoeging woonplaatsnaam | 50,873 | 63 | 1.06491 | 46.3804 |
| 3 | 2 | postcode straatnaam huisnummer huistoewoeging | 2,763 | 3 | 1.06732 | 6.01E+07 |
| 4 | 3 | postcode huisnummer huistoewoeging woonplaatsnaam | 592 | 1 | 1.08784 | 9.29408 |
| 5 | 4 | postcode huisnummer huistoewoeging | 35 | 0 | 1.17143 | 1.58E+09 |
| 6 | 5 | straatnaam huisnummer huistoewoeging woonplaatsnaam | 247 | 0 | 1.05263 | 365.552 |
| 7 | | | 54,510 | 67 | | |
| 8 | 6 | postcode straatnaam huisnummer woonplaatsnaam | 8,304 | 10 | 4.22387 | 438.458 |
| 9 | 7 | straatnaam huisnummer woonplaatsnaam | 129 | 0 | 4.21705 | 114.949 |
| 10 | 8 | postcode huisnummer woonplaatsnaam | 1,470 | 2 | 2.36259 | 599.339 |
| 11 | 9 | postcode huisnummer straatnaam | 600 | 1 | 3.28167 | 7.81E+08 |
| 12 | | | 10,503 | 13 | | |
| 13 | 10 | postcode huisnummer | 111 | 0 | 2.01802 | 1.58E+09 |
| 14 | 11 | postcode straatnaam woonplaatsnaam | 3,337 | 4 | 32.7438 | 28420.4 |
| 15 | | | 3,448 | 4 | | |
| 16 | 12 | straatnaam woonplaatsnaam | 266 | 0 | 109.008 | 116822 |
| 17 | 13 | postcode | 12,193 | 15 | 34.712 | 2.31E+09 |
| 18 | 14 | postcode5 | 56 | 0 | 320.286 | 1.56E+09 |
| 19 | 15 | postcode4 | 117 | 0 | 2611.85 | 8.61E+08 |
| 20 | 16 | woonplaatsnaam | 27 | 0 | 50684.8 | 3.74E+06 |
| 21 | | | 12,659 | 16 | | |
| 22 | -1 | geen koppeling mogelijk | 14 | 0 | 0 | 0 |
| 23 | | | 81,134 | 100 | | |
| 24 | | | | | | |

Figure 6 - Resultaat_meta.csv

As Figure 6 shows, the first 5 geocode-levels more or less give a unique match between an address from the source data and the BAG. You can see that in the column 'gemiddeld aantal adressen' which stands for the average number of BAG-addresses used to calculate a geocode. Often, this is a little bit higher than 1, for instance because of student dormitories containing multiple address in one building, or elderly homes having the same issue.

Other result files describe the number of geocoded objects per administrative unit (by default, CBS neighbourhoods, quarters and municipalities from the year 2016 and NVM areas from 2015 are included in the configuration).

Section 2 - Creating your own BAG-snapshot in two steps

For the two steps depicted below in sections 2A and 2B, two separate GeoDMS scripts are available:

- Xml2FSS.dms: Processes the set of BAG XML files, resulting in an FSS file for each BAG object type. FSS is the internal binary native format used by the GeoDMS to make fast calculations.

The BAG and the resulting set of FSS files contain history. The identifications per object type therefore do not have to be unique (i.e. one address can be listed multiple times, for instance because an old building has been demolished at that address and a new building has been erected);

- MakeSnapshot.dms: Makes a snapshot of the BAG FSS files for a specific moment in time. In this set, the identification per object type is unique. In the snapshot process the addresses table (derived from multiple other tables in the BAG) used for geocoding is constructed. Also other attributes are calculated for the snapshot data (not geocoding-related).

New in the March 2017 release of the tool is that you can also predefine an extent for which you want to create a snapshot. For instance if you are only interested in the Greater-Utrecht region, you can define a bounding box using x and y-coordinates. The advantage of this step is that you will have a much smaller (and thus faster) snapshot tailored specifically to your region of interest. In section 2B, step 3 is described how to do this.

2A) XML2FSS

1. Open the BAG XML-file which you can download from <http://geodata.nationaalgeoregister.nl/inspireadressen/atom/inspireadressen.xml>. The XML file contains a reference to the zip-file with the BAG data, the download size and the update date. Download the referred zip-file into a directory of your choice.
2. The downloaded zip-file contains a zip file per BAG object type. Extract all these files (their names contains the following abbreviations: LIG, NUM, OPR, PND, STA, VBO, WPL) and put them into one so-called XML folder, so without subdirectories. As a naming convention for the directory, we advise *%SourceDataDir%/BAG/%date%/NGR*, in which:
 - *%SourceDataDir%* is the GeoDMS placeholder for the physical path of your SourceData directory on your computer. You can see (and edit) this path name via the GeoDMS GUI, menu-item Tools > Options > Paths;
 - *%date%* is the date of the downloaded BAG (i.e. the update date from the inspireadressen.xml file), with the date convention *yyyymmdd*.
3. Open the GeoDMS-file Xml2FSS.dms in an ASCII/text editor (e.g. Notepad, Crimson Editor). If you have followed the naming conventions mentioned under step 2, then only adjust the item date (line number 27, see Figure 7) so that it matches the date of the downloaded BAG-XML. If you have not followed the exact naming convention, then also adjust the XmlDir in rule 31. Save your adjustments.

If you set another extent, we advise you to adjust the label in rule 56 as well since this label determines the name of the subdirectory in the Snapshot-folder where the results are written to.

Finally, the selection-extent only works if you change the value of the boolean parameter on rule 57 from False to True.

4. Start the GeoDMS GUI and activate the item *snapshot_%shapshotdate% > MaakVolledigeSnapshot* (for example by double-clicking the item) or, if you only are interested in an address table, activate the item *MaakAdressenTabel* instead. You will see that the variable *%shapshotdate%* shows the date you have chosen in step 1. For geocoding purposes, create an address table only will suffice.

The GeoDMS will now write output data in a folder called *%SourceDataDir%/BAG/snapshots.ext/%date%.new* (ext is added to the name because new information is added to the BAG-source data). Depending on your choice (full snapshot versus address table only), there will be more or less FSS-directories and –files.

5. When the writing of data in step 4 is done, close the GeoDMS and then adjust the name of the shapshot.ext-subdirectory from *date%.new* in *%date%*. This snapshot data is now ready to be used for, for instance, geocoding purposes.