

Metadata portal in Statistics Norway

Anne Gro Hustoft¹ and Jenny Linnerud, Statistics Norway

Summary

Statistics Norway's metadata strategy was approved in 2005. The strategy as such advocates for metadata (systems) in Statistics Norway, and several measures were recommended to support this, e.g. the development of a metadata portal.

Implementation of the metadata portal began late 2005, and the portal was released on the Internet in the spring of 2008. The portal now displays our key metadata concepts and the contents of our master metadata systems making them more accessible and easier to use both for researchers and external metadata experts, and for internal users. Metadata managers can use the portal to follow up the coverage and quality of the contents of the underlying metadata systems. The design is flexible so that the contents of other metadata systems, e.g. a questionnaire server, can be added as these become available.

2. Metadata portal

2.1 Purpose and content of the metadata portal

The home page of the Internet version of the metadata portal in English is shown below <http://www.ssb.no/english/metadata/>



Fig. 1 Metadata portal – home page

¹ The presentation will be given by Anne Gro Hustoft

The overall purpose of the metadata portal is to make Statistics Norway's metadata systems more accessible and easier to use. Both internal and external users get easier access to the metadata by displaying the contents of these systems in a common web page. Our work within this area has been inspired by the corresponding web pages of Statistics Canada and Statistics New Zealand.

The main purpose of the metadata portal is to give access to information stored in the metadata systems and delivered by web services, but the page also contains links to other relevant metadata. At present the portal gives access to classifications, variable definitions, codelists, file descriptions, register descriptions and file variables collected from our different metadata systems. The file descriptions, register descriptions and file variables are only shown in the internal version, among other things because they contain sensitive information. In addition to making this information accessible for internal users, this version also gives the user a chance to check the quality of the file description. This is done by an automatic program that checks the metadata in the file descriptions with the data in the permanent files and gives messages like; "X records longer than expected", "Y records shorter than expected", "The variable has values not listed in the code list", "Expected value range is not documented" etc.

The metadata portal also contains metadata that are not yet stored in metadata systems (e.g. definitions of statistical units), links to other relevant Statistics Norway web pages (e.g. About the statistics and Statbank) and external links to relevant international metadata web pages. Under Definitions/Concepts you will, amongst other things, find definitions of our key metadata concepts. The establishment and documentation of key concepts related to metadata was an important part of the implementation of our metadata strategy. The document was made in close contact with those working in statistical methods, IT, production and dissemination of statistics, and it was subject to a hearing round in all these departments. The document was also discussed in our metadata forum, in the steering group for our metadata strategy and in our standards committee. Finally, it was approved by the director general

2.2 Status reports

The metadata portal makes it possible for managers to check the progress of the metadata work. The status reports in fig. 3 are available on the Internet, while the internal version has several more status reports to satisfy the needs of internal users (e.g. Number of variable definitions by subject area, Number of variable definitions linked to file descriptions, Number of variable definitions linked to tables in Statbank and Classification versions approved for internal use, but not for external use – at present a total of 25 status reports).

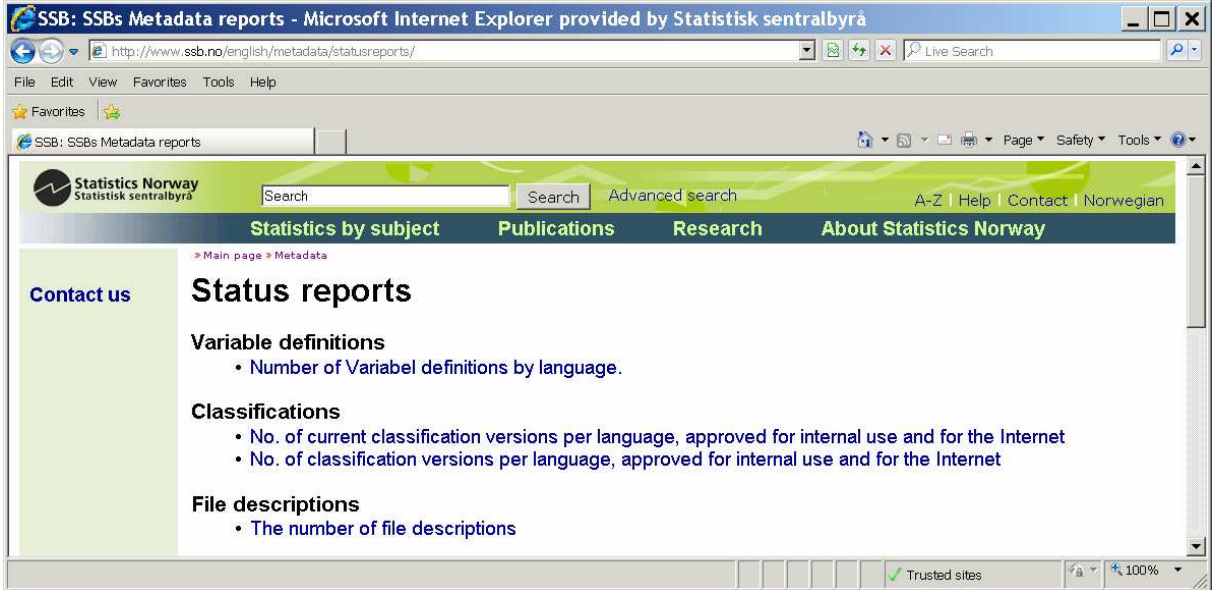


Fig. 2 Available status reports

If we choose the status report for variables, we get the following information:

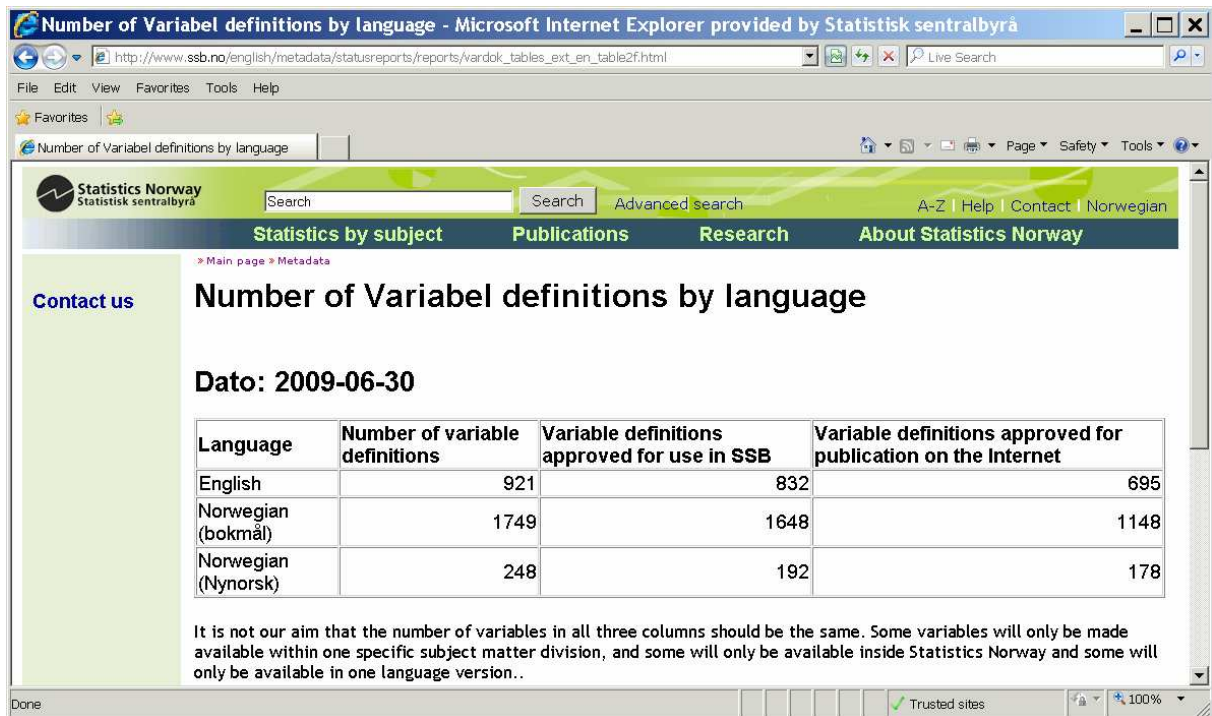


Fig. 3 Status for Variable definitions

As the table shows, there is still some translation work to be done. In the internal portal version the status for the different metadata systems is given per subject matter division which makes it a more relevant tool for the managers.

2.3 Variables

If the users are searching for a variable definition, they can click on the Variable definitions link, and they are then linked to the window below where they can search for a variable definition by using different search criteria (name, word in definition etc.), or use the list of variable definitions in alphabetical order.

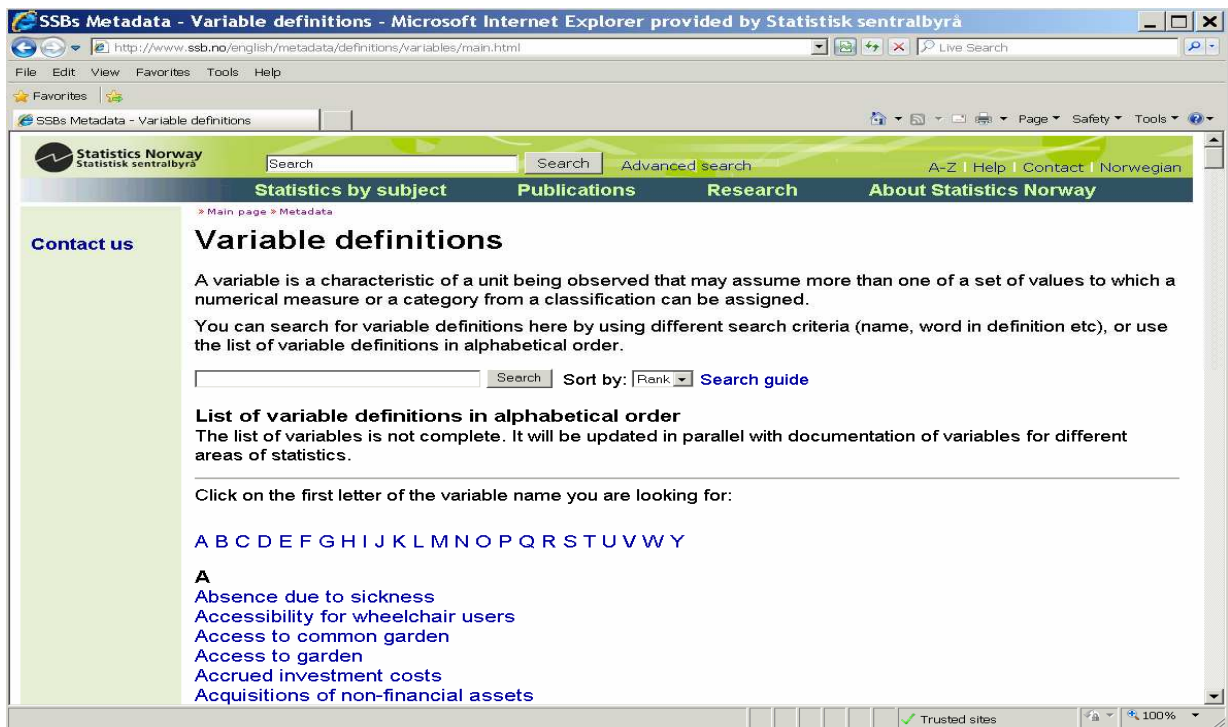


Fig. 4 Variable definitions

If we choose agricultural area as our variable, we get the information in fig.5. By using the two clickable links at the top right in fig 5 the user can also get access to the variable documentation in both versions of Norwegian (Bokmål and Nynorsk).

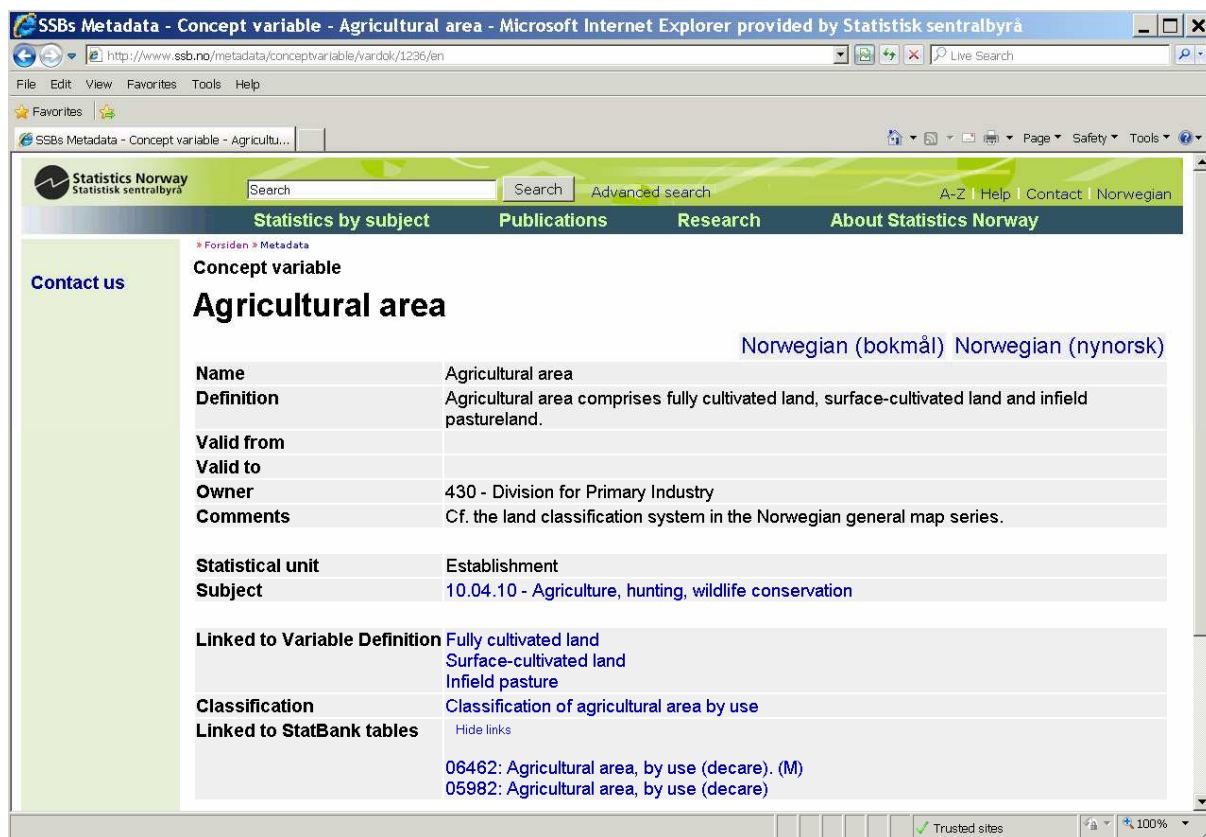


Fig. 5 Documentation of agricultural area

For this specific variable, there is no validity period because the definition has “always” been like this, but for most of the variables, we will have a “valid from”-date for the definition.

2.4 Links

Often a definition will contain other variables (fully cultivated land, surface-cultivated land and infield pasture land) and the definitions of these variables are linked in the “Linked to Variable Definition”-field. If the user clicks on the linked variables, he/she will be able to see the definition of the linked variables.

If the variable chosen is a categorical variable, it will be linked to the relevant classification or classification version in the Classification database, and by clicking on the link in the Classification field, the user will get access to this (see fig. 6).

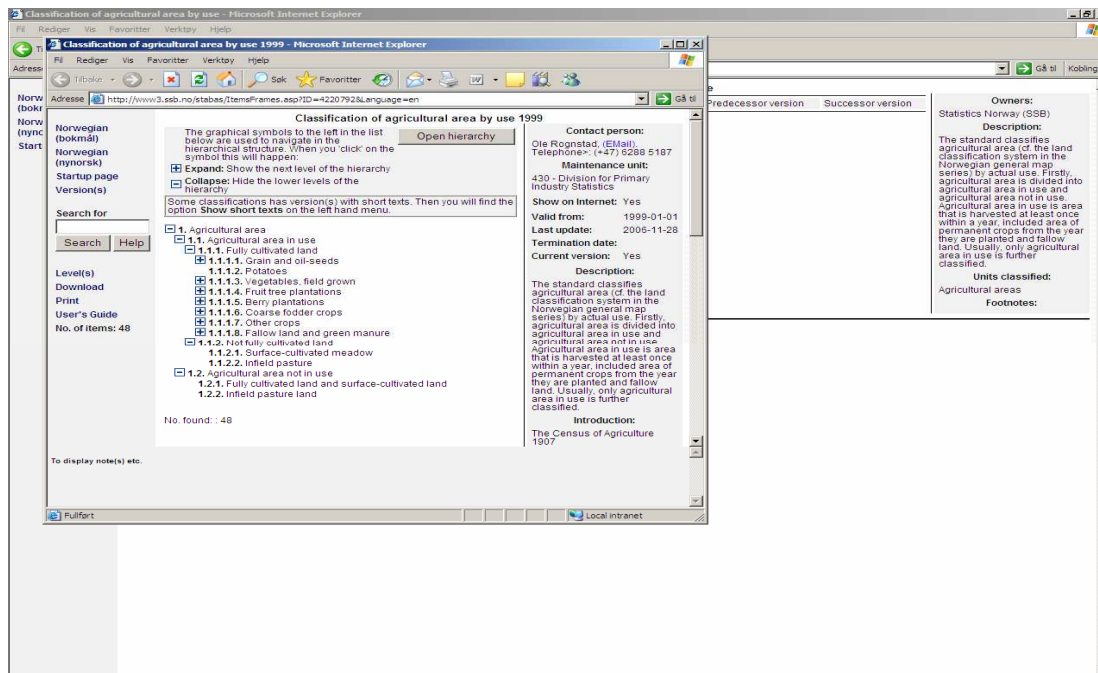


Fig. 6 The classification linked to Agricultural area

It is useful to see the links between the metadata and the data. We link our variable definitions to our dissemination system (Statbank). The linking work is done within Statbank, but in the metadata portal we can see which Statbank-tables the different definitions are linked to (fig. 5). It is quite motivating for the owners of the variables to see that the variables are used on the Internet (also by other subject matter divisions), and it is also useful to know who uses their variables if the owner chooses to do some changes to the definitions (then the other users should be informed so they can consider if they still want to link this definition).

If the user clicks on one of the Statbank-tables, he/she will be linked to this table (as shown in fig. 7),

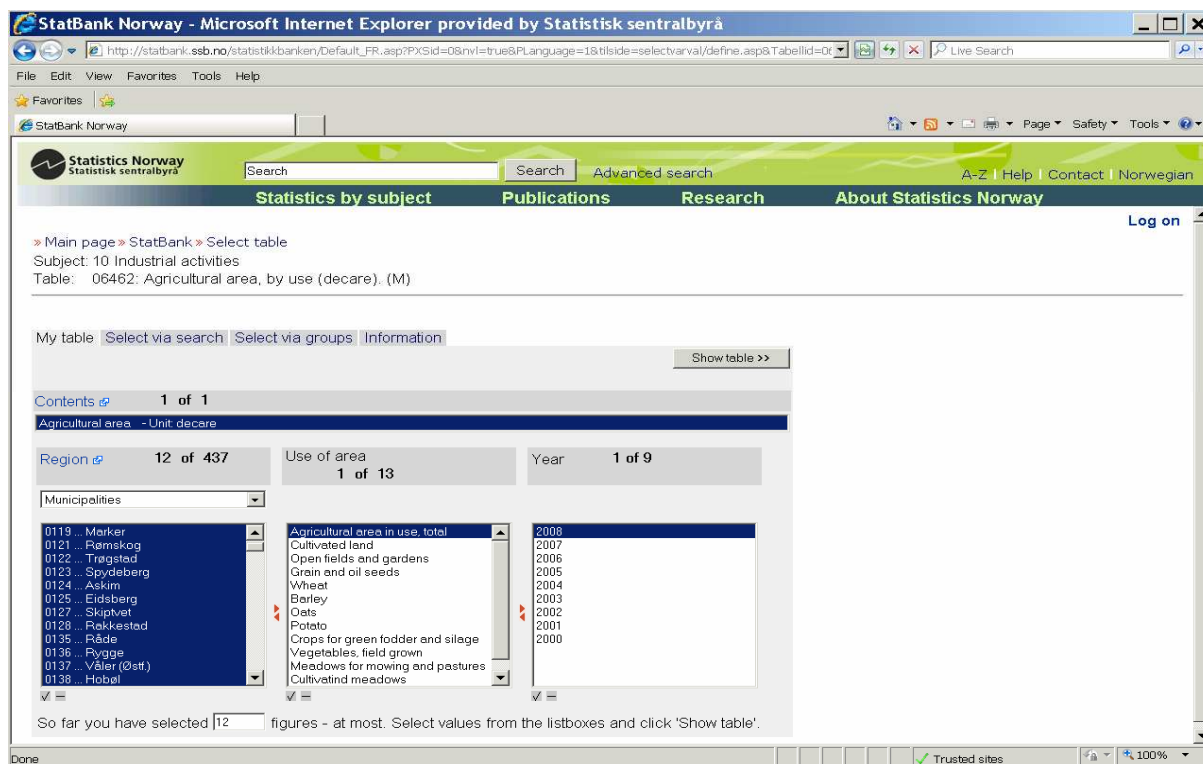


Fig. 7 Statbank-table with link to variable definition

and will be able to make a statistical table showing the relevant data (fig. 8).

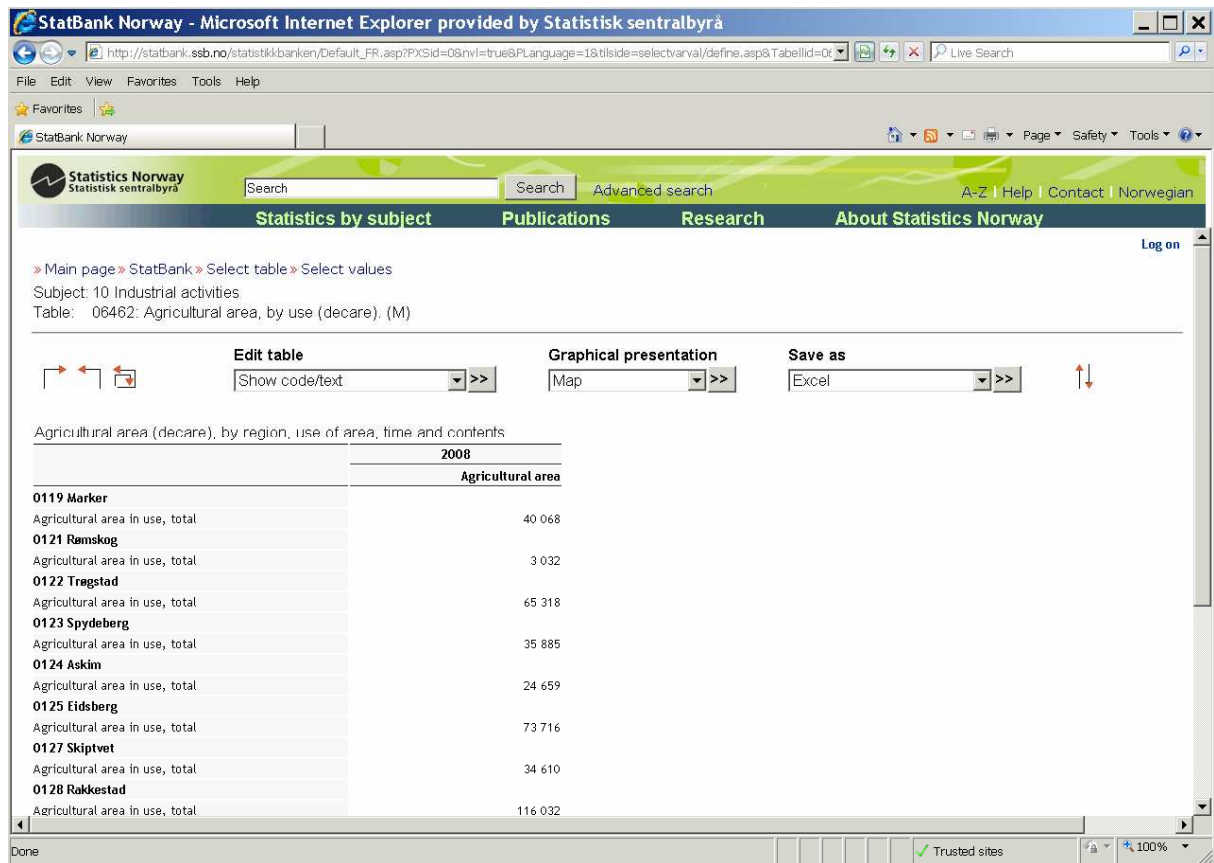


Fig 8 One of the possible tables made from the selection criteria in fig. 7

A user who enters Statbank (with or without using the metadata portal), can click in the Contents-field, and get the linked variable definition shown in fig. 10.

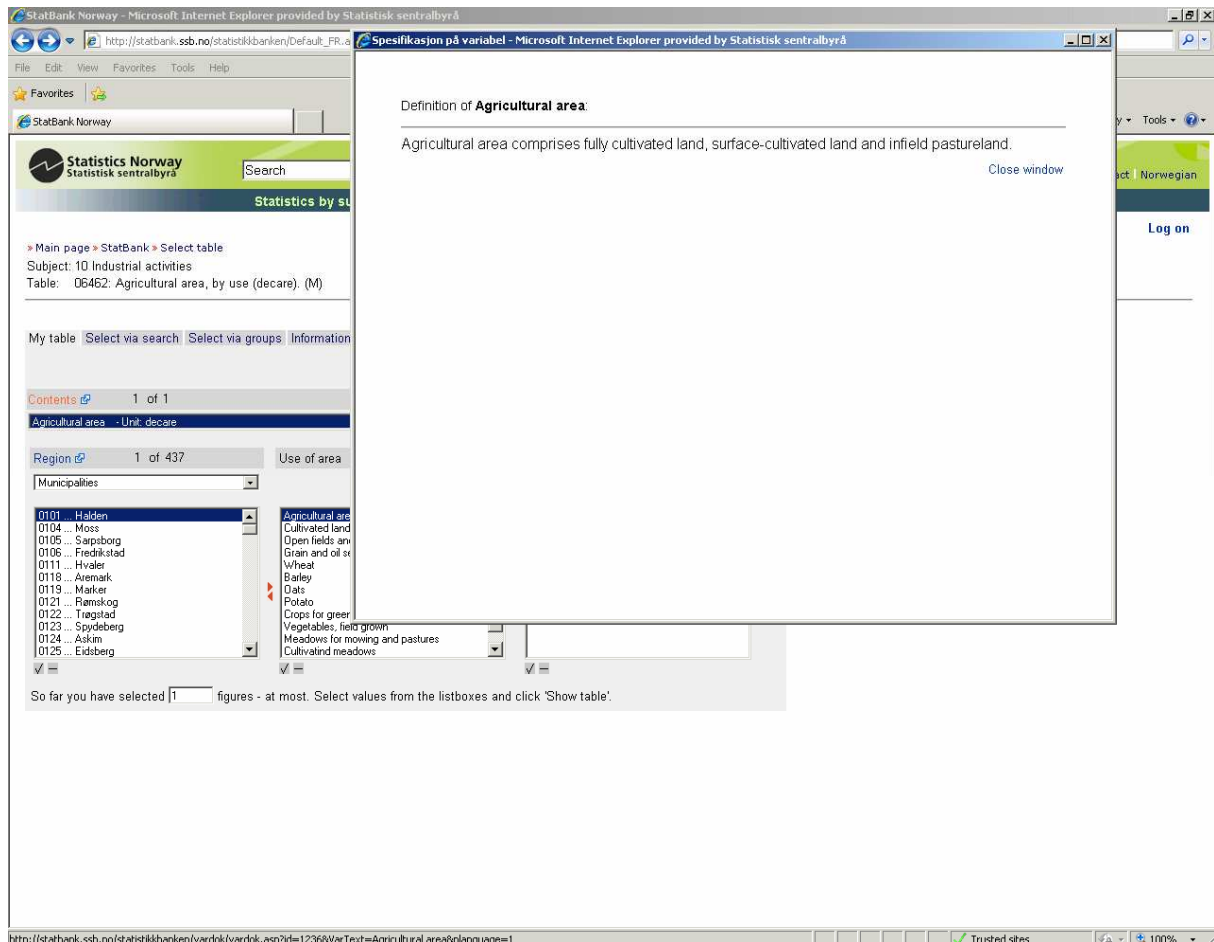


Fig. 9 Variable definition linked to Statbank

2.5 Tool for harmonisation

Some variables have the same name, but are defined differently. This is visible in the alphabetical list where you can see the number of definitions belonging to the same variable name in parenthesis behind the variable (fig. 10). Most of the times that same name/different definition occurs, it is either because they are different versions of the same variable, or because different laws/regulations require the same variable name but define the variables differently according to different subject areas. Sometimes, however, these duplicates do not arise from real differences, but just from lack of coordination. Then we ask the relevant subject matter divisions to look into the variables in question, and see if they can harmonise their definitions and reduce them to one. This is one of the ways that we can use the portal as a tool for harmonisation.

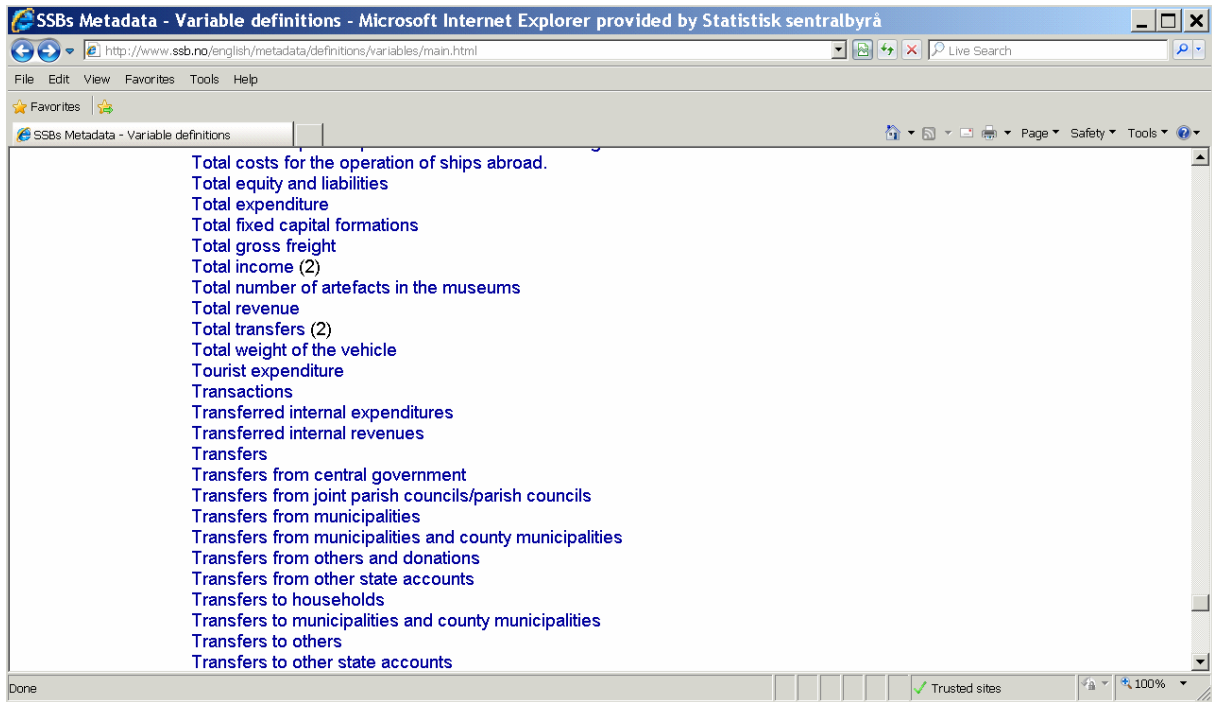


Fig. 10 Same variable name, different definitions

If the two different definitions are due to two versions of the same variable, a click on the variable name will result in the following picture

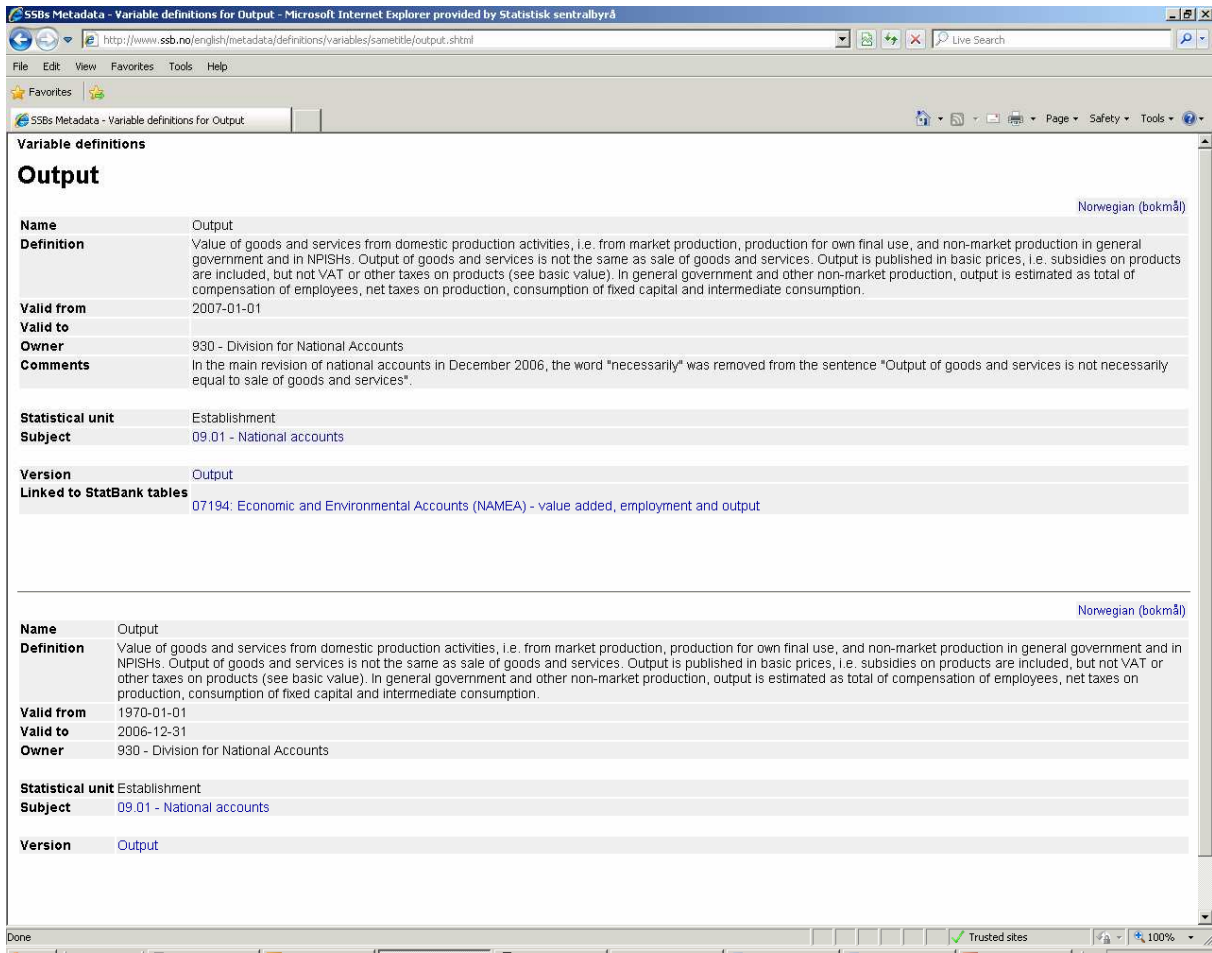


Fig. 11 Different versions of the same variable

2.5 Searching across metadata types

We can also use the home page for searching across the different metadata types. If we search, e.g. for “reclaimed land”, without ticking off any of the metadata type boxes, we will get hits among all metadata types (concept variable, classification version and codelist) where “reclaimed land” is found.

The screenshot shows a web browser window displaying the search results for "reclaimed land" on the Statistics Norway metadata portal. The search results are sorted by rank and include the following items:

- 1 Reclaimed land -- Rank: 1000**
Land reclamation comprises land transferred to arable land (fully cultivated land) or surface-cultivated land. Clearing of infield pastureland is not included.
Statistical unit: Establishment
Subject area: 10.04.10 - Agriculture, hunting, wildlife conservation
Type: Concept variable
Division: Division for Primary Industry
Language: en
- 2 Classification of codes for PRODCOM 2005 -- Rank: 404**
PRODCOM is the title of the EU production statistics for Mining and Quarrying and Manufacturing. Production is to be recorded according to the product headings of ...
Type: Classification Version
Division: Division for Manufacturing Statistics
Language: en
Valid from-to: 2005-01-01 - 2006-12-31
- 3 Classification of codes for PRODCOM 2006 -- Rank: 404**
PRODCOM is the title of the EU production statistics for Mining and Quarrying and Manufacturing. Production is to be recorded according to the product headings of ...
Type: Classification Version
Division: Division for Manufacturing
Language: en
Valid from-to: 2006-01-01 - 2007-12-31
- 4 Classification of codes for PRODCOM 2008 -- Rank: 319**
Type: Classification Version
Division: Division for Manufacturing
Language: en
Valid from-to: 2008-01-01 -
- 5 PRODCOM 2005 (Classification of codes for PRODCOM 2005) -- Rank: 251**
PRODCOM is the title of the EU production statistics for Mining and Quarrying and Manufacturing. Production is to be recorded according to the product headings of ...
Type: Codelist
Division: Division for Manufacturing Statistics

Fig. 12 Search result for “reclaimed land”

2.6 Concluding remarks

Version 1 of the metadata portal was released both on the Intranet and the Internet in February 2008. There still remains some development work related to extended functionality (e.g. file descriptions), but due to lack of IT-resources, this work has been put on hold. As a consequence, much of the work related to the metadata portal in 2009 has been focused on improving the quality of the content in the underlying metadata systems that are being accessed through the portal.

Regarding the Internet version of the metadata portal, we hope that it can make a contribution to semantic interoperability at the national level. At the international level we hope it will make a contribution to fruitful discussions on standardisation and harmonisation of metadata.