

Internet publishing at Statistics Netherlands

Paper submitted by the Central Bureau of Statistics, Netherlands¹

Introduction

1. To an information producer like a statistical office, the Internet is an immense opportunity. With relatively little effort, you can reach very many people. This doesn't mean you will actually reach all of them! And the effort required to get the statistical information on the Internet should not be underestimated. So although the potential of the medium is certainly visible, what exactly should we do with it?
2. This paper describes Statistics Netherlands (SN) ideas and activities with regard to the Internet. What have we done, what works well, where did we go wrong, and what are our plans for the future. Special attention is given to user information and how to organise information effectively. At SN the latter is concentrated around database publishing.

Past and present

3. Statistics Netherlands has had a Web-site since February 1995. Although this is barely three years ago, looking at our first site today it seems very simple, while at the time we were one of the first government organisations in the Netherlands to start a site. It contained some general information and the daily press releases and from the very start we received quite a few visitors. It generated a lot of free publicity and the address of the site is very easy and recognisable in the Netherlands: WWW.CBS.NL (CBS is the Dutch abbreviation of our Bureau). Of course the quality of an Internet site is important to generate traffic, but a recognisable address which is easy to remember and consistent references to the site in other media are equally important. We do realise that the international recognisability of the Dutch abbreviation is limited.
4. After a year and a half we decided to renew the site. The graphic design became an important issue: how to present the information so a user can find what he or she is looking for?

¹ Paper prepared by Lydia van der Hulst & Ahmed Aboutaleb, Statistics Netherlands

5. We are still looking for a definite solution. The form chosen in 1996 followed the familiar presentations of information on paper, and was at least consistent with the other media Statistics Netherlands uses.

6. However, the Internet has its own rules and dynamics, and we came to the conclusion that the translation of existing products to an Internet version is not enough. The site had become a collection of information which had originally been compiled for other media or target groups. At the moment we are in the process of introducing a third version of the site, which will contain products made by SN especially for the Internet, using the newest techniques. These products are described in paragraph 6.

7. A web site is not a stable product. Up to now the web-site of SN has had a life-cycle of a year and a half. Presenting yourself on the Internet means investing continually in technique and content.

In touch with the market

8. To develop and improve a product it is important to keep in touch with the market. Site log-files are an important source of information, but only of quantitative information and only about what is already on offer, not what people would like to see but look for in vain. Mails offer more qualitative information but are time-consuming to analyse. Other sources of market information are market research and desk research studies.

Log-files

9. The log-files show us how many users consult which pages and when. At present there is a lot of discussion about the significance of log information. Should we monitor hits or user sessions, what is a user, what is the effect of the cache problem? These are problems of measurement and as such quite familiar to statisticians: define what you want to measure, and stick to this definition. It is particularly important to analyse developments and comparisons.

10. The number of users of *www.cbs.nl* is increasing all the time. Analyses of the log-files show that most users are interested in figures and press releases. Eighty per cent of the visitors use the Dutch site, 20 per cent the English site. On the English site the page with links to other statistical offices is immensely popular. In general links to other sites are seen as inefficient because in this way you lose a lot of visitors, but the service this page provides is important for our image. We use the log-files to evaluate the content of the site. Popular pages are given brothers and sisters, while less popular ones are removed or placed in a distant corner.

11. Up to now we have refrained from using cookies, except when they are functionally necessary, because of the irritation they provoke, and also, because of this irritation factor, customers disable the cookies and your measure becomes distorted.

12. For a few months now we have also been logging keywords used in the StatLine database. However, apart from some global analyses we have not yet found a way to provide feedback in this respect to the content of the database, mainly due to staff shortages.

Mails

13. Mails give more qualitative information. However, mails from users already on the site are difficult to trace because so many e-mail addresses are made available on the site. We have noticed an increase in the number of e-mails during the last years, but this did not come as a surprise. Most e-mail addresses are used inaccurately. This is very revealing, both in respect of ourselves (too many e-mail addresses are made available, without clear subject links), but also of the users (when

a question arises people use the first e-mail address in sight). The work done at Statistics Netherlands is organised around statistical themes and by trial and error we are looking for a way to communicate all these e-mail addresses to users so that questions are directed at the people who can answer them.

14. Because of the time-consuming aspect of analysis, combined with a shortage of staff to do the job, we have not yet found a way to monitor e-mails in the way we should. Analyses are done only incidentally. An analysis of the e-mails received by the webmaster showed that most concern requests for additional statistical information and technical queries. Surprisingly the number of mails does not seem to be related to the number of visitors to the site: we receive relatively fewer mails as the number of visitors increases. And there are relatively more mails from visitors to the English site than from those to the Dutch site.

Market research

15. Other sources on the market also give information. For instance customers buying other products inform us that they would prefer to receive this information electronically. The Internet is not an isolated activity: its position in the data dissemination process must be continually be evaluated.

16. Questionnaires on the site can give the information about the users which we are lacking in the log-files. SN is very cautious about using this instrument because of the response burden issue in the Netherlands. A questionnaire on a voluntary basis will not always give reliable results. SN plans to set up an Internet panel in a few months, as we expect it will provide more useful. results..

Desk research

17. The Internet is new territory for everyone, not only statistical offices. It is very useful to learn from others, publishers, for instance. What do they do on the Internet? The challenge is to translate these developments to the environment of a statistical office.

Effective information

18. Making an effective Internet site is not an isolated activity. An Internet site can only be effective when it is part of an integrated media mix. This integrated media mix is formulated in the data dissemination policy of SN. As a government office we are obliged to publish our information virtually free of charge; we may only charge reproduction costs. The Internet can be used in different ways within the data dissemination policy (from website to e-mail). Even a website is not a product, but rather a dissemination channel which can be used in different ways for different products and different groups.

19. The latest Internet techniques (channels, agents, dynamic HTML) make it possible to present tailor-made information. The subsequent increasing individualisation of information should not be seen as technology push, but as a consequence of the information overload. People cannot handle the enormous amounts of information the Internet and other media present to them and look for tools to organise this stream. Technology helps them, but on the other hand editors, who select the information, become more important. The disclosure and selection of information becomes one of the most challenging tasks of a statistical office. So the reproduction costs of the Internet are not as low as they seem at first sight. Publishing on the Internet involves making new products. The Internet offers so much information that most customers want someone to make a selection for them.

20. The main task of a statistical office is to make figures accessible in tabulations. The number of accessible tabulations is increased enormously with the Internet. Statisticians can no

longer hide their tables behind difficult texts as everyone must be able to read them. At the same time expert users also use the Internet. The same figures have to be made available to different groups with different needs through the same medium. The minimum solution to this problem is given by meta-data: be exhaustive and organise meta in such a way that every possible selection for specific groups can be made from it.

21. Meta-data is used by users to:

- search the data
- interpret the data

22. For both these tasks, the meta-data required depends on the customer: is he or she an expert or a non-expert, and for what purpose does he need the data. SN has developed a number of directives for meta-data in electronic publications.

- The index is the responsibility of a central editorial board. The index is an important instrument for users to find what they are looking for, and to know where they are during a user session.
- All statistics need to be provided with a description of the population and the survey .
- Every variable needs to be provided with its definition.
- Every figure needs to be provided with its unit of measure.

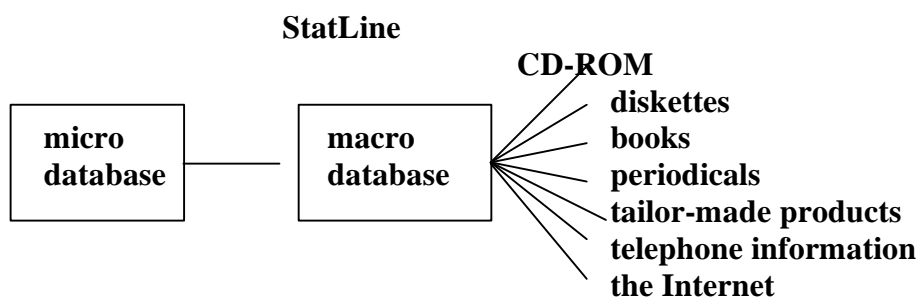
23. These directives are simple but necessary. They are expected to increase in number in the years to come. We even give courses to statisticians on how to make database publications and the importance of meta-data.

24. In the old days meta-data were recorded, but there was no need for anyone except statisticians to read them. Nowadays meta-data has a public function. This development is only a couple of years old. Asking ourselves how we should tackle this transition, we decided that being exhaustive so that later on customer specific selections can be made seemed to be the best solution at the moment. Datawarehouse techniques offer promising solutions.

Database publishing

25. The most important development within SN is not publishing on the Internet, but database publishing. The question is not how to get from print to electronic publishing, but how to organise the publication process to be able to make a paper or electronic publication on demand.

26. At SN database publishing is centred around StatLine. StatLine is the first step towards a flexible standardised medium-neutral output database. Input for StatLine is presently generated from existing publications. In the near future the intention is to turn this process around: output from StatLine will be input for other output sources. This will have consequences for the statistical production process. Before publication is possible all publication data will have to be put into StatLine. To make this easier, the StatLine input process will have to become part of the statistical production process: micro data - StatLine - publication. This will require adjustment of the organisation and of computerised processes.



27. In time StatLine will lead to a more efficient output process. StatLine is much in favour of the tendency towards (standard) tailor-made products and publication on demand (printouts, e-mail); flexibility and efficiency are increasingly important factors and these techniques become more profitable as the format of the output is standardised. Some publications will disappear altogether as their edition is too small, while new ones will be introduced, either on paper or electronic devices. This process will be steered by what the customers want.

28. The Dutch market for statistical information is a niche market, and within it needs are very specific. This means we can only publish small series and tailor-made products. Although we expect the international market to become more important in the future, we do not know how much more. StatLine makes it possible to serve different groups with products which suit them best in terms of content and medium: the whole database or a selection from it on CD-ROM or on-line, publications on paper or diskette, printouts or e-mail on demand, telephone information service. Flexibility pertains not only to electronic publications, but to paper products too.

29. Although StatLine will make the above described processes easier in a technical sense, it is necessary to think ahead. Because of the information overload, selecting information for target groups is more than ever necessary and also has to be provided more rapidly. To keep up with these needs we must constantly analyse market information.

New developments

30. There are a couple of new developments within SN along the lines described in this paper. They consist of technical aspects and of content-oriented projects.

Dynamic HTML

31. Most information on the site is made by hand in static HTML pages. With the use of the newest Internet techniques we are aiming for less manual work and more automated processes. We use Active Server Pages to create dynamic HTML pages from databases, and we have started to generate a dynamic product catalogue from an internal Access database. The next project will be an automated process from press releases in Word to publication in HTML standard layout. The use of database techniques allows us to control the publication process on the Internet more efficiently. The content is made in applications known by the authors. Technology takes care of conversion to HTML. In this process updates of the content and the company style are guaranteed.

StatLine

32. Last year SN put its StatLine database on the Internet, to be consulted free of charge. To open up the information in the multi-dimensional database, whose format was developed by SN itself, advanced search strategy techniques are used. Users on the Internet are used to fuzzy search. In the development of the interface we made a distinction between expert and non-expert users.

Expert users know what they want, they understand the concepts behind the database and are familiar with the vocabulary used. For them we offer an index. On the Internet StatLine is also exposed to non-expert users. The risk of their not finding the data they are looking for is very high. For them SN developed HyperSearch, an intelligent search engine. Users can type one or more keywords (e.g. *bicycle Amsterdam*) and the software creates a table from the database with a maximum score on the combination of these words. If the result is not satisfactory they can go back to the search screen with one click and can alter the keywords (*cars Amsterdam*) or specify (*bicycle Amsterdam 1997*). The target of the HyperSearch program was to reduce the number of user decision points of a user. HyperSearch does a full text search in the meta-data in the database to define a table. This places an extra demand on the meta-data.

Although the online database is only available in Dutch, it is currently being translated into

Web Magazine

33. One example of a product specially made for the Internet is the Web Magazine. This magazine is situated on our homepage. The formula is based on the *news* criterion, and pertains not only to news in the sense of the latest press releases of SN, but also a statistical translation of topical issues in Dutch society. For instance, if there is an outbreak of swine fever in the Netherlands, the Web Magazine would contain an item on pigs, based on livestock statistics. News items are entered daily. This Web Magazine is based on the customers' needs for the selection of information. The introduction is planned for May 1998, so we can say nothing of the results at present.

Tailor-made products

34. Along the same line (customers' needs for a selection of information) SN develops tailor-made products to which customers can subscribe. Push technologies are to be used to get them from the StatLine database to the customer. Other customer friendly ways of disseminating data are target group networks, intranets and extranets. The owners of those networks act like editors who make a selection of the information offered. SN wants to co-operate with these network owners to establish together a selection of statistical information which is useful for the members of these networks.

PDF database

35. SN wants to make all the paper publications accessible through the Internet. For the time being we have opted for a PDF-database (Portable Document File). These textual publications will initially be stored in a separate PDF database, but later on integration through links with StatLine must become possible.

Bilingual

36. Although SN has an English language site, the most valuable information is stored in StatLine. We have started with the development of a bilingual (English/Dutch) database to enter the international market. Naturally this involves inherent meta-data problems.

Conclusions

37. The Internet is a fantastic medium, but we must not forget that customers will dictate how and when it is used. The number of users suggests that the SN site and the e-mail facilities provide a wanted service. We are curious about the results and experiences of other statistical offices with quantitative and qualitative user information.

38. The most important thing we have learned in the development of an Internet site is: don't go faster than the organisation. First organise your information. You have to be sure the information you present on the site is updated regularly. At the same time the back-up office must be ready to take care of the (new) reactions the Internet site will raise. This means not only an e-mail address but also the capacity and the skills to answer the queries. Be aware that an Internet site generates traffic you didn't even know existed before.

39. The Internet has caused an information overload. At the same time the most interesting feature of the Internet is the possibility to generate tailor-made information at an individual level. For a statistical office this is an opportunity to open up the immense amount of information which until recently was stored inside the building. The accessibility of the information requires statisticians to adopt a different attitude towards the publication of their figures. Storage in a database, together with extensive meta-data, from which tailor-made information is produced is the process of the future.

40. As statistical offices we must not only learn from each other, but also help each other. The Dutch site contains a list of other statistical offices world-wide: a very popular page. In this way we not only provide a service to our users but also generate traffic between the statistical sites.

Appendix 1. Contents of WWW.CBS.NL

Bilingual:

- press releases
- short term economic indicators
- key figures
- links to other statistical offices
- general information about SN, its organisation and its services

Dutch only

- StatLine (database with all statistics)
- Web Magazine
- product catalogue
- advertisements
- publications in PDF format

Appendix 2. The new homepage of SN including the Web Magazine

The screenshot shows the homepage of the Central Bureau of Statistics (CBS) in the Netherlands, viewed in Microsoft Internet Explorer. The browser window title is "Centraal Bureau voor de Statistiek - Microsoft Internet Ex". The address bar is empty. The page has a blue and white color scheme.

Navigation and Language: The top navigation bar includes "Home", "Site-map", and "Engels". The left sidebar contains a vertical menu with the following items: CBS-NIEUWS, CIJFERS, INDELINGEN, CATALOGI, BIBLIOTHEEK, BERICHTGEVERS, ORGANISATIE, and AGENDA.

Main Content Area:

- Centraal Bureau voor de Statistiek**
Het Centraal Bureau voor de Statistiek verzamelt, interpreteert en presenteert gegevens over de Nederlandse samenleving
update: 24 maart 1998
- Nieuwe Site CBS**
Dit is een testsite. Bedoeld om de functionaliteit te testen en niet de inhoud. Lees voor meer informatie de [introduce](#).
- Minder werklozen**
De geregistreerde werkloosheid blijft in een hoog tempo dalen. Het einde is nog niet in zicht.
- Einde AIDS-epidemie?**
Een daling van bijna 25 procent in vergelijking met 1995 doet vermoeden dat we het ergste hebben gehad.
- Hoger beroep loont**
Wie wordt veroordeeld tot onvoorwaardelijke gevangenisstraf en daartegen in hoger beroep gaat, krijgt gemiddeld een kwart minder straf.
- Bouw floreert**
De hoogconjunctuur zorgt voor ongekende activiteit in de bouwsector

Right Sidebar:

- StatLine**
Raadpleeg GRATIS de statistische DATABASE over Nederland
- Recent verschenen**
INDEX
Verkiezingen

N.B. dummy-version

Appendix 3. StatLine search screen

