



aan LS
van A.H. Kroese
CBS
onderwerp onderzoek 'remote access to microdata'
datum 30 juni 2021

Wat is remote access?

Het CBS biedt de mogelijkheid aan geautoriseerde onderzoeksinstituten zelf onderzoek te doen met de databestanden van het CBS. Onderzoekers kunnen vanuit een veilige werkplek via een beveiligde internetverbinding onderzoek doen op deze databestanden die koppelbare data bevatten op persoons-, bedrijfs- en adresniveau. Uiteraard zijn hier eerst direct identificeerbare gegevens uit verwijderd. Dit wordt remote access (RA) genoemd. Hiervoor krijgt de onderzoeker een persoonsgebonden token in bruikleen. De onderzoeker heeft alleen toegang tot de bestanden die nodig zijn voor zijn onderzoek. Het is onder voorwaarden ook mogelijk om eigen bestanden te uploaden en te koppelen aan CBS-data. De onderzoeker krijgt een afgeschermd werkomgeving tot zijn beschikking waarin hij tussenbestanden, syntaxen en output kan opslaan. Alle data blijven binnen de beveiligde omgeving van CBS. Als een onderzoeker (tussen)resultaten buiten de beveiligde omgeving wil brengen, dan controleert het CBS of de resultaten geen onthullingsrisico bevatten.

Introductie

Het CBS verzamelt data met als doel om betrouwbare statistische informatie te leveren voor de overheid, het bedrijfsleven, de wetenschap en burgers. Tal van statistieken worden volgens een vast programma als nieuwsbericht en als open data via de website van het CBS gepubliceerd. Daarnaast maakt het CBS op verzoek van (met name) overheidspartijen ook vele maatwerkstatistieken. De data die het CBS publiceert via zijn website zijn niet herleidbaar tot individuele personen, bedrijven of instellingen.

De wet staat toe dat ook externe organisaties voor het doen van wetenschappelijk of statistisch onderzoek gebruik kunnen maken van de databestanden waar het CBS over beschikt. Door externe onderzoekers de kans te geven met de data te werken profiteert de samenleving optimaal van de veelheid aan data die het CBS beheert. Er is op deze manier immers veel meer onderzoek mogelijk dan het CBS zelf zou kunnen doen. De databestanden waar deze onderzoekers mee werken, kunnen privacygevoelige informatie bevatten. Daarom is het van groot belang dat bij dergelijk gebruik de veiligheid goed op orde is en onderzoekers zorgvuldig met deze data omgaan, zodat de privacy van burgers en bedrijven niet geschonden wordt.

Organisaties die aan alle eisen van de Beleidsregel toegang instellingen tot microdata CBS voldoen, kunnen via remote access (onderzoek op afstand) toegang krijgen tot de data die het CBS heeft verzameld. Dit gebeurt onder strikte voorwaarden. Voor externe organisaties gelden dezelfde eisen



op het gebied van veiligheid en privacy als voor het CBS. De remote access-voorziening is uitgegroeid tot een belangrijke dienst van het CBS. Het is internationaal een toonaangevend voorbeeld.

Met het oog de snel groeiende gebruikersaantallen, nieuwe technische ontwikkelingen en ontwikkelingen op het gebied van informatieveiligheid en privacy heeft het CBS onderzoek laten doen naar de toegang tot CBS-databestanden via remote access. Dit onderzoek is in 2020 uitgevoerd door een commissie van onafhankelijke wetenschappelijke experts onder voorzitterschap van Prof. Dr. B. van den Berg (Hoogleraar Cyber Security Governance, Universiteit Leiden). Het onderzoek richtte zich op de volgende twee vragen:

- Welke potentiële cybersecurity- en privacy-risico's zijn in de toekomst mogelijk bij de huidige vorm van remote access tot databestanden van het CBS?
- Welke maatregelen zou het CBS kunnen nemen, of welke (beleids)keuzes zou het CBS kunnen maken om de juiste balans te garanderen tussen enerzijds het algemeen belang van het brede gebruik van de gegevens die al verzameld zijn, en anderzijds het belang dat burgers hebben in de veiligheid en privacy van hun gegevens?

Het onderzoek

De evaluatiecommissie is in maart 2020 begonnen met haar werk, dat uit vijf stappen bestond:

1. Verzamelen van informatie over de werking, functionaliteit en toegangsprocedures van de CBS remote access. Deze informatie werd ook vergeleken met remote access-faciliteiten van statistische organisaties in andere landen.
2. Diepte-interviews met partijen aan zowel de gebruikerskant als aan de CBS-kant.
3. Literatuurstudie op zowel technisch, juridisch, als sociaalwetenschappelijk gebied, die kennis over de belangrijkste risico's op het gebied van privacy en gegevensbescherming bijeenbracht.
4. In kaart brengen van mogelijke risico's. Hierin zijn de resultaten van de diepte-interviews en de literatuurstudie verwerkt, en is een theoretisch kader voor risicoanalyse ontwikkeld.
5. Bundelen van de vier tussenstappen tot een samenhangend afwegingskader en een voorstel voor een scenario-instrument.

De resultaten zijn gebundeld in het hier gepresenteerde finale eindrapport. Het rapport representeert de visie van de onafhankelijke onderzoekscommissie. Het CBS is niet verantwoordelijk voor de juistheid, volledigheid en actualiteit van de inhoud.

Een breed samengestelde begeleidingscommissie heeft tijdens het onderzoek mede beoordeeld of de diverse belangen die spelen bij toegang tot databestanden in het onderzoek voldoende belicht worden, zoals gebruiksvriendelijkheid van de faciliteit, maatschappelijke waarde van de onderzoeken, de positie van verschillende categorieën gebruikers en bescherming van privacy. De begeleidingscommissie meent dat het finale rapport een nuttig en werkbaar kader geeft om tot beleidskeuzes te komen.

Conclusie en vervolg

De commissie concludeert dat de remote access-voorziening hoge wetenschappelijke en maatschappelijke relevantie heeft en dat deze voldoet aan de huidige veiligheids- en privacy-eisen.



Om dat ook voor de komende jaren veilig te stellen is blijvende aandacht voor beveiliging en privacy nodig. De commissie heeft geadviseerd om de beleidskaders explicieter en transparanter vast te stellen. In aanvulling op de aanbevelingen van de commissie heeft het CBS gekeken naar betere aansluiting van het beleid bij de CBS-wet en de AVG.

Dit heeft geleid tot enkele aanscherpingen in het beleid rond de toegang tot en het gebruik van CBS-databestanden. Zo zijn de voorwaarden voor instellingen en projecten verduidelijkt, om te verzekeren dat het gebruik van de remote access-voorziening wordt ingezet voor statistisch of wetenschappelijk onderzoek dat voldoet aan normen als zorgvuldigheid, controleerbaarheid en onafhankelijkheid. Verder is beter omschreven wat wordt verstaan onder het openbaar maken van de resultaten van het onderzoek. Ook zijn de toegangsvoorwaarden aangescherpt om de privacy nog beter te borgen. Het CBS dient te voldoen aan de AVG en wil zijn data alleen beschikbaar stellen aan organisaties uit landen die aantoonbaar een passend beveiligingsniveau kennen. Om die reden verleent het CBS alleen nog toegang aan instellingen, diensten en organisaties die gevestigd zijn in landen die onder de reikwijdte van de AVG vallen (de Europese Economische Ruimte) of waarvoor een adequaatheidsbesluit van de Europese Commissie van toepassing is.

Voor het gebruik van remote access is een instellingsmachtiging nodig. De Beleidsregel voor het verkrijgen van die machtiging wordt voor 1 augustus gepubliceerd in de Staatscourant en is zo straks kenbaar voor eenieder. Eventuele toekomstige aanpassingen aan het beleid rond projectmachtigen of procedures, publiceert het CBS deze op zijn eigen website. Ook worden gebruikers van de remote access-voorziening voor ingang van de wijzigingen actief op de hoogte gesteld.



Universiteit
Leiden

Faculty Governance and Global Affairs

Centraal Bureau voor Statistiek

t.a.v. de heer A.H. Kroese

Henri Faasdreef 312

2492 JP Den Haag

Reference

Your reference

Subject

Milestone #5 Onderzoekscommissie

Remote Access to Microdata

Date

30 Oktober 2020

Doorkiesnr.

Contactpersoon Prof.dr. B. van den Berg,

Geachte heer Kroese,

In maart van dit jaar startte de Onderzoekscommissie Remote Access to Microdata in opdracht van het CBS met een onderzoek naar potentiële privacy- en security risico's rondom het aanbieden van toegang tot microdata door uw organisatie. Hierbij bieden we u de laatste milestone aan van dit onderzoek: een rapport met de titel '**Remote Access to Microdata: Final report**'. Dit rapport is de output van het vijfde werkpakket in dit onderzoeksproject, maar brengt bovendien alle voorgaande werkpakketten samen.

Met hartelijke groet namens de Onderzoekscommissie,

Prof.dr. Bibi van den Berg

Universiteit Leiden



Remote Access to Microdata Final report

Milestone #5: **Work package 5**
Version: **Final**
Delivery date: **30 October 2020**

Authors:

Name	University	Expertise
Prof.dr. Bibi van den Berg (vz)	Universiteit Leiden	Cybersecurity governance
Prof.dr. Bram Klievink	Universiteit Leiden	Technology, policy and government
Dr. Tommy van Steen	Universiteit Leiden	Organisational change & cybersecurity
Dr. Esther Keymolen	Universiteit van Tilburg	Privacy & data protection
Dr. Bart van der Sloot	Universiteit van Tilburg	Law & ICT
Dr. Zeki Erkin	TU Delft	Data security & privacy



Table of contents

1.	Introduction	4
2.	Outline	8
3.	Remote access to microdata at CBS	10
	3.1 The remote access to microdata environment at CBS: Key elements	10
	3.2 Using the remote access environment: Findings from interviews	12
	3.3 The risks regarding remote access and microdata: A literature review	15
	3.4 An analysis of the main risks of the current remote access to microdata environment	20
4.	Value-based risk analysis	24
5.	Parameters of (remote) access to microdata at CBS	27
	5.1 End users	28
	5.2 Use	32
	5.3 Data sets	35
	5.4 Processes	39
	5.5 Combining all parameters	42
	5.6 Residual risk	45
6.	Scenarios for access to microdata at CBS	47
	6.1 A stringent scenario	47
	6.2 High accessibility	50
	6.3 Moderate in all	52
	6.4 Stringent on end users, lenient on use, data sets and processes	54
	6.5 Stringent on data sets, lenient on end users, use and processes	57
7.	Conclusions and recommendations	61
	7.1 Values	61
	7.2 Risks	62
	7.3 Recommendations	63



1. Introduction

The Central Bureau of Statistics (CBS) in the Netherlands operates a Remote Access environment, through which end users can get access to microdata for (statistical) research purposes. Offering this service has great societal value: end users, for example researchers at universities and knowledge institutes or government organisations and planning bureaus, can use microdata to shed light on academic, societal and policy challenges. Using factual knowledge contributes to a thorough understanding of the challenges at hand, and of the potential effects of government interventions. Moreover, research that builds on such data helps inform the public of societal developments and, especially in times of disinformation and fake news, may contribute to a more nuanced, fact-driven debate on these developments. While recognizing the high societal value of the Remote Access to Microdata environment, in the past years CBS has become increasingly concerned over the security and privacy aspects of providing access to such data. Much time and effort has been invested in ensuring that the technical environment in which data are made available is as secure as possible, and that a number of checks and balances are in place to ensure that the environment is only accessible to carefully vetted end users.

CBS sought to establish whether its current implementation of the Remote Access to Microdata environment is ‘future-proof’, and whether it makes use of the latest insights into the protection of privacy and into security. To this end, CBS enlisted the help of an Evaluation Committee, consisting of specialists in the field of cybersecurity, privacy, governance and law to answer the following **research questions**:

- What potential cybersecurity and privacy risks may be involved in offering the Remote Access to Microdata service in its current form? and
- What measures could CBS take, or what (policy) choices could it make to find the right balance between on the one hand serving the public interest by providing access to the data it gathers, while on the other hand protecting the private and collective interest of citizens by warranting security and privacy?

The Evaluation Committee started its work in March of 2020 and has delivered four intermediary reports, which detail aspects of the evaluation, so far. After an **information gathering** phase (milestones 1-3) the Committee delivered an analysis of the **main risks** to this environment, as well as a **framework** that can help CBS weigh competing values with



Blad 5/65

respect to access to microdata services (milestone 4). This deliverable is the final one in Evaluation Committee's investigation of the current Remote Access to Microdata environment. It summarizes the key findings from the previous reports and presents a **methodology**, as well as a set of scenarios that CBS may use to chart the future of its Remote Access environment.

The fact that CBS seeks to maintain some form of functionality of access to microdata to outsiders seems evident. Embracing the values of **public responsibility** for societal and economic advancement, **transparency** and **accessibility** underpin this choice. End users, CBS itself and the Evaluation Committee all agree that this service has **high societal and economic relevance** and should therefore not be suspended. At the same time, in light of other values that CBS seeks to safeguard, most importantly **security** and **privacy**, questions arise as to how to structure the access to microdata environment, now and in the future. In an ideal world, this service does justice to both the values of making data accessible for others to use and public value on the one hand, while safeguarding security and privacy on the other.

This is no easy feat. When thinking about the ways in which a remote access environment to microdata has been designed in the past, and could potentially be redesigned in the future, there are several **parameters** one can focus on, viz.:

1. **End users:** *who* gets (remote) access to microdata, and under which conditions?
2. **Use:** what do end users *use microdata for*, what can they do with microdata, and under which conditions?
3. **Data sets:** which *data sets* does CBS own and operate, and which data sets that contain microdata are made accessible (remotely)?
4. **Processes and procedures:** which *processes* and *procedures* are in place for (remote) access to microdata, and to whom, for which purposes, and under which conditions do these apply?

In essence, each of these parameters of (remote) access to microdata can be adjusted by CBS, almost like a '**dial**' that can be turned left or right, from **more stringent** to **more lenient**, as shown in Figure 1¹. For instance, with respect to end users, CBS might choose to diversify the categories of users that are allowed access, which would increase the values of accessibility, public value and transparency. At the same time, this might decrease security

¹ The Evaluation Committee has chosen to use the colors red, orange and green to signify a transition from stringent (red) to lenient (green). The colors in no way are intended to signal 'bad' or 'good' practices or choices.

Blad 6/65

and raise privacy concerns. CBS could also choose narrow the categories of users that have access to microdata, with the opposite impact on values as a result.

The parameters listed here are those that the Committee identified in its work that could be considered adjusting. As such it does not represent an overview of all potential policy parameters that govern the current situation and is not a suitable instrument to assess or get an overview of the (policy) configuration of the microdata environment.



Figure 1: The parameters in this report can be used as a dial.

In this deliverable, the Committee will show that each of the aforementioned **parameters** can be **adjusted** (applied in a more stringent or a more lenient fashion) to decide the future of (remote) access to microdata. The choices made in doing so express a leaning towards certain values rather than others, both per parameter and with respect to the totality of choices made. Collectively, this leads to an outcome for a future implementation of (remote) access to microdata, which CBS may legitimize with reference to the value choices that underpin each parameter.

Since each parameter can be set to different outcomes, the number of possible instantiations of the actual (remote) access to microdata environment is large. In this deliverable, the Committee presents **five scenarios** to show what different combinations of parameters would lead to. These scenarios are intended as outlooks into potential futures only. They are in no way intended to contain ‘the path’ that CBS should follow or to provide guidance of where to take (remote) access to microdata from here. The Committee advises CBS to use the scenarios as an inspiration in its thinking of the desired setting for each of the four parameters.



Blad 7/65

When all parameters are set, and a choice has been made for a future implementation of (remote) access to microdata, one challenge still remains: **residual risk**. As long as CBS offers access to microdata, no matter how stringently this is done, there are risks to security and privacy. The two most important ones are:

- **Identification** of individuals or companies in the data sets made available via the access to microdata functionality; and
- The **copying** of data(sets) by end users through e.g. the use of video or screen capture technologies.

These two risks are unsolvable, no matter how hard CBS works to minimize the risks. While CBS should be aware that these risks are inherent to providing access to microdata to outsiders, and should do all it can to minimize these risks, the only way to truly eliminate them is to no longer provide access. As we have argued above, this is in nobody's interest, and would be nobody's recommendation. The societal and economics value of access to microdata is such that a certain level of risk has to be accepted as part of offering this service. The goal of using the 'dials' is to find an optimal balance between competing values and factor in an acceptable level of residual risk.



2. Outline

In section 3, this deliverable will start with a **summary** of the **main findings** of Committee's research into remote access to microdata at CBS, which was conducted between 1 March and 31 July 2020. It describes:

- **What remote access is** and how the current remote access to microdata environment is designed and **implemented** at CBS (section 3.1);
- The results from **interviews** with a variety of **end users** (section 3.2);
- The results of a large **literature review** on the **key risks** relating to remote access of microdata (section 3.3); and
- An **analysis** of the **main risks** in the **current remote access to microdata environment** at CBS (section 3.4).

Traditionally, risks are mapped and analysed using risk management tools and methods. One of the key elements of this type of approach is the quantification of risks through the calculation of the likelihood of the materialisation of particular risks (turning them into incidents) and the consequences this materialisation would have. This Committee has chosen a different approach to establish risks, viz. one that is based on **tensions between particular values**. Section 4 will describe a motivation for this choice and an explanation of the **value-based risk analysis** developed in this research.

One of the main benefits of using a **value-based risk analysis** is that it facilitates thinking about risks as a set of **dials** that can be set to different levels of **stringency** or **leniency**, translating into a different tension between particular values. The remote access to microdata environment at CBS consists of four key **parameters**:

1. **End users:** *who* gets (remote) access to microdata, and under which conditions?
2. **Use:** what do end users *use microdata for*, what can they do with microdata, and under which conditions?
3. **Data sets:** which *data sets* does CBS own and operate, which data sets, containing microdata, are made accessible (remotely)?
4. **Processes and procedures:** which *processes* and *procedures* are in place for (remote) access to microdata, and to whom, for which purposes, and under which conditions do these apply?



Blad 9/65

Each of these parameters can be adjusted by CBS in future implementations of the remote access environment using the value-based risks analysis; each parameter can be set to a more stringent setting, or to one that is more lenient. Choosing for the one or the other, or somewhere in between, will express different values. In section 5 we will discuss each parameter and the different choices that can be made with respect to them. In the same section we will discuss the category of **residual risks**: those risks that are inherent to offering access to microdata and that cannot be resolved, other than by ending that service entirely.

While the four parameters can be adjusted individually, it is of course the **collective outcome** of the choices made vis-à-vis each parameter that decides the actual shape and form that access to microdata might take at CBS in the future. In order to show how this works, the Committee provides CBS with **5 scenarios** in section 6:

1. A **stringent** scenario
2. High **accessibility**
3. **Moderate** in all
4. Stringent on **end user**, lenient on use, data sets and processes
5. Stringent on **data sets**, lenient on end users, use and processes

Many other variations and scenarios may be developed – the ones discussed here are simply intended to provide a starting point for CBS's decisions with respect to the future of remote access.

The document ends with a set of **conclusions** of the entire evaluation as conducted by this Committee, and a set of **general recommendations** to take into consideration for the future of remote access to microdata by CBS in the Netherlands (section 7).



3. Remote access to microdata at CBS

This section provides a summary of the four previous reports that the Evaluation Committee has written during its research period (March 2020 - September 2020).

3.1 *The remote access to microdata environment at CBS: Key elements*

On the basis of the provisions in Article 41 of the Statistics Act Netherlands, CBS offers a so-called **remote access service to microdata**. CBS started offering access to confidentialised microdata in 1994 and since 2006 this access is provided through a remote access service. The remote access service enables institutions from a range of categories to gain access to particular data sets of CBS via an internet connection. A special environment has been created to facilitate this in a technical sense. CBS has invested significant effort into practical and technical procedures to ensure that the remote access service can be used in a secure way, and that the privacy of citizens/consumers and companies is protected when data sets containing microdata are accessed by individuals. Moreover, CBS has also sought to ensure that its data are shared and stored in a secure way.

This access to microdata is available to, and used by, **five different types of users**:

1. Dutch universities;
2. Dutch planning agencies such as Planbureau voor de Leefomgeving (PBL) and Centraal Plan Bureau (CPB);
3. Research institutes established by law (TNO);
4. Statistics bureaus in other European countries and Eurostat;
5. Other institutes.

Since CBS's remote access service could be considered a 'channel to the outside world', in which subsets of the CBS data become accessible to outsiders, CBS sought to investigate:

- What potential **cybersecurity and privacy risks** may be involved in offering this service in its current form;
- What **measures** CBS could take, or what (policy) choices it could make to serve the public interest by providing access to the data it gathers and facilitate researchers, while protecting the private and collective interest of citizens and companies by warranting security and privacy.



Blad 11/65

Accessing the Remote Access environment is facilitated by setting up a Citrix client in combination with a Virtual Private Network (VPN) connection. Only Mac and Windows systems are allowed to connect to CBS. In addition, a physical token, a pin code, and a mobile phone as well as a username/password are needed. In order to be able to access a specific project, a researcher will also need a (project specific) username and password and TAN-code, which will be sent to the smartphone of the user. As a safety measure, CBS will disconnect other network connections of the end-point (the researcher) during the time they spend within the environment to prevent the recording of data. Before gaining first-time access to the Remote Access environment, researchers and their manager must sign a non-disclosure agreement and researchers will have to answer a number of security questions to test their knowledge on what is and is not allowed when working with the microdata. During later log-on attempts, a single security question is asked which is to ensure that researchers remain aware of the necessary security measures they have to take.

When allowed access, researchers will have a protected work environment at their disposal in which they can store intermediate files, syntaxes and output. Various software is available for analyses, such as SPSS, Stata, R, Python, and packages for existing software can be made available upon request. During the project, the researcher can request CBS to make various adjustments: add or remove researchers, add files (if linked to the research question), temporarily suspend the project, etc.

The remote access environment is a **separated environment** of CBS and all **microdata remain within CBS**. Only the microdata necessary to answer the research question posed by the user will be made accessible. On their website, CBS maintains a list with information reports on available microdata, distinguishing different themes (from ‘labour and social security’ to ‘leisure and culture’). Aside from these data sets, it is also possible to request for a tailor-made microdata set. Moreover, end users are allowed to connect their **own data set** with the microdata, “*provided that investigators are legally entitled to use these data and (as far as it concerns personal data) the applicable privacy and data protection legislation is respected.*”² CBS will replace directly identifiable variables (such as Social Service Number, Chamber of Commerce registration, combination day of birth, and address, etc.) with a **Record Identification Number**. This is a form of **pseudonymisation**. The data available to

² Instructions for file format and upload process (Date: 2018-10-09), p.4.



Blad 12/65 the researcher is therefore still considered personal data under the GDPR. Privacy-sensitive information will be protected from unauthorised access throughout the entire project.

If one wants to move findings outside the secure environment, this can only be done via the **export folder**. Prior to the export of (interim) results from the secured environment, CBS will check whether the results do not contain any disclosure risks. In order to ensure that there are no direct identifiers, CBS provides ‘rule-of-thumb’ solutions which are constructed in such a way that an export file which is in accordance with these rules can be considered almost 100% ‘safe’. These solutions contain amongst others the possibility to: aggregate data, insert blank cells, blur results, add noise. Thus, it is not a quality check of the research results, but a **confidentiality check of the output**.

3.2 Using the remote access environment: Findings from interviews

The next phase of the evaluation conducted by the Committee involved gaining in-depth insight into the **use of the remote access system to microdata by ‘customers’**, or end users, using this CBS system. The goal was to get a better understanding of the various ways in which the remote access environment is used and perceived by end users. More specifically, we sought to answer three general questions:

1. What do end users **use the remote access environment for** and what is the perceived usability?
2. What are the end users’ **experiences** in terms with respect to **security** and **data protection**? How do security measures influence usability and vice versa? And
3. What would it mean for end users if CBS would decide to **expand** or **scale down** the remote access functionality?

We envisioned face-to-face interviews to collect our interview data, preferably with all parties. However, due to the COVID-19 restrictions, all interviews were carried out via video calling services. We conducted semi-structured interviews with **11 end users** from a range of organizations: three universities, three companies, two foundations, one planning agency, one ministry and one municipality. These interviewees were chosen by looking at the list of authorized institutions that may use the remote access to microdata system, and the list of projects with microdata sets. The Committee chose six interviewees on the basis of the following criteria:

- Their experience with the remote access system;
- The type of projects they run; and



Blad 13/65

- The type of organization they work for (i.e. universities, local and national government and commercial organizations).

A further two parties were suggested by the Advisory Board, and one party was selected via the researchers' own network. Furthermore, CBS was requested contact details of parties that were **denied access** to the Remote Access environment, and as a result two such parties were interviewed. Below we outline the main findings of these interviews.

What do researchers use their access to microdata for?

The interviews show that end users make use of microdata for a wide variety of research. They feel privileged to have access to these data and express a high degree of caution with respect to the proper use of the system and the data. For some end users existing procedures and processes for access to microdata fit well; for others these processes are sometimes too slow. CBS might solve this by **diversifying** more between different types of end users, and to see whether **services may be adjusted to their needs and wishes more**, rather than offering a one-size-fits-all solution for all end users, with exceptions on request of the end user, as appears to be the case now.

Research practices prior to microdata access

End users all have **valid reasons** for wanting/having access to microdata at CBS. Some already conducted similar research, but at much higher cost and effort; others could only do their research at a smaller scale, and some could not do their current research without CBS data at all. The high societal and economic value of access to microdata is underwritten by all interviewees.

Combining CBS data with end users' own data

In the interviews, many end users explain that they use the CBS service to **combine** microdata with their **own data sets**. The end users we interviewed claim that they have the potential to **uniquely identify** individuals or companies in the combined data set, when the data set is not too big or the data too homogenous.³ This takes time and effort and is against the rules, so end users state they would not do this, even though they could. This is also possible without combining CBS data with data collected by users, but the identification becomes much easier

³ Note: this risk is not strictly linked to the fact that end users can combine CBS data with their own. In principle, the same risk could materialize by combining CBS data from multiple data sets as well. Whenever data sets are small and the data are homogenous, the risk of identification exists. However, this risk has a higher likelihood when end users combine CBS data with their own data, because (1) they are generally more familiar with the data in their own data sets, and know better to which uniquely identifiable human being or organizations they may refer, and (2) their own data sets may not use the same levels of pseudonymization and may be much richer in detail per record, which facilitates identification.



Blad 14/65 in combined data sets. For the future, CBS must be aware that **not all end users are guaranteed to have similarly benign intentions**. Their approach is at least partially based on trust. Should a researcher not have a long-term interest to safeguard access to microdata, and be interested in using it for finding information about a specific individual or company (for example to engage in fraud or extortion), then the **current practice of combining data sets at CBS poses a risk**.

End users' experiences with the remote access environment

In the interviews, end users expressed that their **experiences** with the current remote access system to microdata at CBS are **positive** overall. According to end users, improvements could be made in the following areas:

- Increasing the speed of the network.
- Improving user feedback when the system is unresponsive or slow.
- Increasing the number of licenses for software in the remote access environment.
- Strictly monitoring that end users do not access the remote access environment with (their own personal) devices if such devices do not meet the security standards of CBS.

The tension between security and usability in the remote access system

The usability of the remote access environment is considered good by end users, especially in light of the tension with security measures. One point of concern they have is the fact that **CBS does not appear to engage in regular checks of end users and institutions after they have been granted access to the remote access system**. Most importantly, CBS does not check regularly whether or not individual employees with remote access functionality are still employed by the same organization⁴, and it does not check often enough whether institutions' access has expired. Both may lead to security issues.

Potential security risks surrounding the use of microdata

The security risks of the remote access system are not at the front of end users' minds, as expressed in the interviews. The risks they mention most are:

- The risk of extracting data through photos, videos etc.
- The risk of identifying individuals or companies in the data.

⁴ In the contract that organizations sign with CBS, one of the stipulated obligations for organizations is that they should inform CBS in a timely fashion whenever an employee changes jobs. In practice, however, organizations seem to fail to do so on a regular basis.



Blad 15/65

Expanding and improving the remote access facilities

The most important addition to the current remote access service offered by CBS, according to interviewees lies in providing **more technical assistance for researchers** who struggle with statistical tools. Especially for young and early-career end users it would help if they could get more support for their use of the remote access system.

3.3 The risks regarding remote access and microdata: A literature review

In order to provide input for the risk analysis of WP4, in WP3 the Evaluation Committee conducted a large **literature review** on the academic research relating to privacy, trust, and cybersecurity issues in the context of remote access. In addition, WP3 produced a number of best practices for CBS to take into account. WP3 consisted of three sections: a section on technical security, one section on trust, and one section containing a legal analysis.

Technical security

In the section on technical security, four themes that are relevant for CBS's remote access to microdata environment were discussed: (1) common approaches to protecting data confidentiality, (2) key security measures, (3) anonymization techniques, and (4) differential privacy. Each will be briefly summarized below.

There are two common approaches to protecting **data confidentiality**:

1. Preventing unauthorized access to sensitive data by means of security measures and cryptographic tools; and/or
2. Limiting the amount of information revealed on the data and its owner, by deploying techniques such as pseudonymization, anonymization and differential privacy.

The state-of-the-art on **key security measures** for remote access to (personal, sensitive or otherwise critical) data includes:

1. Secure rooms: physical locations in which access to confidential data is provided. A secure room requires, inter alia, physical security, access control, identity management, screening, monitoring, audient and appropriate security software.
2. Secure transmission of data: To send data from the server to the client, end-to-end encryption is required. End-to-end encryption requires two important features, namely



Blad 16/65

authentication, and integrity of the data. Both features can be provided by well-known techniques such as using SSL, TLS IPSec, and approaches such as VPN.

3. Secure processing of data: approach requires more advanced cryptographic tools such as homomorphic encryption, garbled circuits and secret sharing schemes.

With regard to the topic of **anonymization**, it is important to note that there is a difference between anonymization and **pseudonymization**. The latter refers to the act of replacing identifiers with one or more artificial identifiers, also called pseudonyms. This approach provides a way to re-identify the individuals in a data set since the identifiers are not completely lost as is the case in anonymization. **Anonymization**, by contrast, is a common technique used to **process data without risk to individuals**. To achieve this goal, the identifying information from data points in the data set is removed, preventing malicious attackers from performing inferences from the data. Linkage attacks try to link one individual to a record or to a value in a given table or to establish the presence or absence in the table itself. There are different models of privacy:

1. K-Anonymity: K-anonymity seeks to reduce the granularity in a data set in such a way that for each quasi-identifier group in the table, there should be at least $k-1$ other records with the same quasi-identifiers.
2. L-Diversity: A table is said to be l -diverse if every quasi-identifier block in the table contains at least l 'well-represented' values for the sensitive attribute.
3. T-closeness: A table is said to achieve t -closeness if for every quasi-identifier group in the table, the distribution of a sensitive value in the group is within t of the distribution values in the whole population.

There is no single method that can be deployed which is effective against all attacks. There is a trade-off between utility and privacy: more anonymization means less utility. This observation is especially relevant in the case of CBS: anonymizing data might lead to (significantly) decreased usability because it becomes difficult, if not impossible, to combine data sets (either those of end users or those of CBS) when there is no pseudonymized, common identifier in records. While anonymization would strengthen privacy, therefore, at the same time the cost to usability, in this case, might be too high. Research on anonymization techniques is burgeoning, and the Committee recommends that CBS follows the developments closely to see whether an optimal form of anonymization with less cost on usability can be found in the near future.



Blad 17/65

With respect to **differential privacy**, Dwork has proposed to add noise with a certain distribution to data sets to increase privacy. The idea is as follows: given two data sets with a difference of a single record, based on the output of a statistical query, an attacker's ability to distinguish whether the query is performed on the first or the second data set is given with a probability that is bounded by a parameter ϵ , $\exp(\epsilon)$, which is determined by the data publisher. This mechanism is not about changing the data set but adding noise to the outcome of a certain query. Thus, differential privacy is about the mechanism (algorithm that produces the outcome); it is unaffected by the auxiliary information and it is independent of adversaries' computational power.

Trust

Aside from the technical measures described above, social and organizational strategies are also key to a resilient and effective remote access policy. Trust, as a social strategy, plays an important role within organizations as well as in interpersonal interactions to deal with uncertainty. When actors have trust, they accept that they cannot completely control the consequences of a certain transaction or interaction and that they are dependent on others. The difference between **control-based** and **trust-based strategies** is that the former are focused on limiting uncertainty and preventing negative outcomes, while the latter are focused on enabling human actors to interact in and deal with an uncertain environment. As a consequence, the process and outcome in trust-based interactions is generally much more *open* than in control-based interactions, leaving more room for creativity and innovation.

On a very basic level, trust is a **three-partite relation**, consisting of a **trustor** (x), who trusts a **trustee** (y), to perform a certain **action** (z) that is of importance to x . This trust relation is embedded in a specific socio-technical context, which influences the trust relation. For instance, I might be more inclined to trust someone if I am in familiar place where I feel safe than when I am in an unfamiliar place. Or, to relate it to the CBS case, CBS (x) might be more inclined to trust researchers (y) when they are working *on* CBS premises (z) rather than *off* CBS premises. While trust in essence is about 'not knowing', it is not completely blind either. Several trust cues are considered to enable trust:

1. **Reputation** is one of these. The way in which actors have behaved in the past and how they are judged by others gives trustors some assurance that trustees will live up to their commitments.
2. **Third-party-trust** is also a robust mechanism that can help to foster trust. If the trustor and trustee, for instance, do not know each other well enough to engage in a transaction,



Blad 18/65

a third party can be added to the interaction who mediates the trust relation between the trustor and the trustee.

3. **Reciprocity** can also be a strong trust cue. If both actors mutually depend on each other (so each of them is simultaneously a trustor and a trustee) there is a strong incentive to not betray trust as this would most likely harm their own interests.

A risk could be that while CBS assumes that reputation and repeated access to the remote access service is of importance to the researcher, a malicious actor who aims at one-time access will be less convinced by that. Moreover, what also needs to be considered is, to what extent the trust cues that are currently used still fit with the socio-technical context in which CBS and its remote access service operate. In the trust relation of CBS and researcher, the following examples are **trustworthiness conditions**:

1. **Assurances**: signing a contract, come to CBS to meet remote access employees.
2. **Competence**: the affiliation of a researcher with research institute, his or her prior work, and/or the fact that (s)he successfully took the security test.
3. **Commitment**: this is less tangible since it refers to a personal attitude. There is always the possibility that researchers pretend to be committed to follow the rules, but in reality, they don't. One way of encouraging this personal positive attitude is to foster "community trust". By developing and promoting a community of remote access workers who have shared norms and values and interests, trustworthiness could be strengthened.

Legal analysis

The CBS Law stipulates five categories of institutions to which access to microdata may be granted, the fifth category being: Other institutions, most notably "*research departments of ministries and other departments, organisations and institutions*"⁵. There is no clear legal guidance on the basis of which criteria it must be decided whether to grant 'other institutions' access to microdata or not. Legal analysis provided the following findings:

1. From the early 1980s onwards, expectations with respect to the use of CBS microdata by other organizations were high. In reality, however, the use of this possibility in practice has remained fairly limited for a long time.
2. Over time, more restrictions have been imposed on sharing of personal/individual data.
3. There has always been controversy about the category of 'other institutions'.

⁵ Statistics Netherlands Act (effective from 1 January 2017), Section 41, sub 2.



Blad 19/65

4. Although there was a legal requirement to have a governmental decree making clear which institutions this category concerned in the 1988 CBS Law (which did not enter into force), this requirement has been removed.
5. Although there was a legal requirement to have a supervisory authority/board (CCS) assent to a decision of the CBS on this point in the CBS Laws from 1996 and 2003, this requirement was removed when the supervisory board was abolished as of 2017.
6. There was a suggestion in 2003 to remove the requirement that other organizations should request access to microdata, instead granting the DG the authority to decide on this matter on his own initiative, but this proposal was rejected.

Compliance with the GDPR

An analysis of CBS **policies** and **practices** with respect to the GDPR was performed and resulted in the following conclusions:

1. **Personal data:** CBS states that microdata are to be considered personal data, so that the GDPR applies. It is also clear that publications, either by CBS or by other organizations having access to the data, cannot regard or reveal individual cases. CBS uses a minimum of 10 persons/units per category about which results are published (though CBS has indicated that this number is a rule of thumb and can in very specific instances be adjusted based on the research project at hand). It is unsure whether the minimum of 10 units will still hold as a sufficient threshold for ensuring that these results cannot be led back to the 10 relevant units. CBS could also decide, in line with what has been discussed, to treat all data – aggregated, statistical and non-personal data – under the GDPR-regime or a GDPR-light regime.
2. **Controllershship:** Some scholars have suggested that strict categorizations and separations between e.g. data controller and data processor will not hold in the current data-driven environment where such roles are often unclear or hybrid and can change per day. Although on some accounts CBS already seems to embrace such a broad approach to responsibility and accountability, in particular when it comes to the parties it gives access to its microdata, a strict separation between roles and responsibilities still is the basis of the data policy by CBS.
3. **Legitimate processing ground:** Where private organizations want to have access to CBS microdata for other interests than public interests, it would be advisable for CBS to do a careful assessment per case of the legitimacy of those organizations' interests, whether the processing of these data are really necessary in light of that interest and whether these



Blad 20/65

interests should be deemed higher or more important than the interests of the data subjects concerned.

4. **Sanctions:** Although there are many strict technical and organizational security measures, the sanctions are mild and depend on the belief that parties having access to microdata now would want to have access to those data in the future. Although this might currently function as a deterrent, it is doubtful whether this will hold in the future. More severe consequences, such as damages being requested via tort or contract law or criminal prosecution have not occurred so far. Whether there are contractually agreed sanctions and fines and to what extent they can be successfully imposed on non-EU parties is also unclear.
5. **DPIA:** It may be recommendable to do pre-Data Protection Impact Assessments (pre-DPIA's), also for the processing operations of third parties, and where necessary full fledged Data Protection Impact Assessments (DPIAs). Recent literature suggests that such an impact assessment not only needs to assess GDPR-compliance, but compliance with all fundamental rights guaranteed within the EU and broader ethical and societal concerns.

3.4 An analysis of the main risks of the current remote access to microdata environment

WP4 provided an **overview of the main risks** the Committee has found surrounding the **current remote access to microdata environment** as it has been implemented at CBS. The identified risks are clustered in four categories:

1. End users
2. Use
3. Data sets and
4. Processes and procedures.

Aside from the identified risks, a **framework** has been developed that can be used to **evaluate the competing values** that CBS faces when making decisions about (the future of) remote access to microdata. These values have also been clustered. Three categories are identified:

1. **Driving** values
2. **Underpinning** values and
3. **Process-based** values.



Based on the identified risks surrounding the remote access case and the identified values, WP 4 maps **which risks impact which values**, as summarized in the Table 1 below.

Risks/values	Driving values research, accessibility & user-friendliness	Underpinning values trust, security & privacy	Process-based values responsibility, compliance & leading by example
End users			
1. Untrustworthiness of users	No direct impact. In the longer term, this might lead to lower accessibility.	Can have a negative impact on privacy and security if it leads to abuse of access.	May signal insufficient CBS procedures and have a negative impact in relation to compliance and leading by example.
2. Underspecification category 'other users'	May impact research since it may not always be clear if all end users actually have purely a research goal.	May negatively impact data protection principles.	May impact compliance of CBS, if there are no clear policies to identify actors as stipulated in the CBS law.
Use			
1. Risk relating to the use of microdata	No direct impact. In the longer term, this will probably lead to lower accessibility.	May have a serious impact on trust in CBS. Privacy and security may be compromised.	Is in violation of CBS' responsibility, both in terms of compliance and of leading by example.
2. Risks relating to the use of the remote access environment	No direct impact. In the longer term, this will probably lead to lower accessibility.	May have a negative impact on trust in CBS.	May have a negative impact on leading by example.
Data sets			
1. No risk assessment data sets	Supports research, accessibility and user-friendliness as data are made accessible indiscriminately for all researchers and all types of research.	May lower the level of security and have a negative impact on privacy and data protection.	Doing risk assessments on data sets could be seen as part of CBS' responsibility.
2. No anonymization, only pseudonymization	Supports accessibility and research	May have a negative impact on privacy and security	Pseudonymization seems to be the middle ground between CBS' responsibility towards its open data and its privacy obligations.
Processes and procedures			
1. Quality of internal control	No direct impact. In the longer term, research may benefit from low internal controls because end users' activities are not hampered by monitoring and control	Low quality of internal controls might have a negative impact on both privacy and security, and ultimately trust.	Having low procedural safeguards in place in this respect may undermine CBS' obligation of responsibility.
2. Limited checks after entry	Limited checks after entry might further accessibility and user-friendliness, and therewith the overarching aim of research.	Limited checks after entry might have a negative impact on both privacy and security, and ultimately trust.	Having low procedural safeguards in place in this respect undermines CBS' obligation of responsibility, and may affect compliance and leading by example.

Table 1: Value-based risk analysis for the remote access to microdata environment at CBS



Blad 22/65

End users

As summarised in the table, untrustworthiness of users may have impact on CBS with respect to compliance and leading by example. In case of misuse and abuse, there will likely be consequences on security and privacy of the data sets. The current situation where the end user category “other users” is underspecified creates a concern since there is no clear identification of actors as such. This can raise compliance related problems and might also damage data protection principles as the purpose of end users labelled as “other users” cannot be identified clearly. A possible consequence is mission creep, and eroding support for the programme by end user who feel disadvantaged.

Use

The risks of providing access to microdata to researchers can vary: the data can be transferred to other parties, copied, or used to identify individuals and organisation by using one of the services of CBS: bringing in own data. While it could be time intensive to do so, actors with malicious intent can decide it is worth the effort to be able to store copies of (part of) micro data sets. Any misuse or abuse will have a negative impact on trust in CBS, and consequences with respect to compliance and leading by example.

Data sets

CBS has variety of data sets open to researchers. It is known that these data sets are valuable, however, no privacy risk assessment are carried out on these data sets. It might be the case that some data sets may have more information that can be connected with other (internal) data sets, leading to identification of individuals and organizations. To mitigate potential risks, pseudonymization is being used at the moment. However, as explained above, it is worthwhile to investigate which, if any, state-of-the-art anonymization techniques could be used to increase privacy and security, while at the same time not unnecessarily hampering usability too much, and thus safeguarding the extensive societal and economic value of (combining) microdata.

Processes and procedures

Last but not the least, the quality of internal control should be constantly monitored. Continuity of the staff is important; thus, regular training should be provided since the



Blad 23/65 reputation of CBS can be degraded significantly in case of a human error in a low-quality control case. Furthermore, the same level of attention should be applied to researchers for monitoring them after entry, noting that researchers and their drives can change over time.



4. Value-based risk analysis

When organisations seek to understand, and act upon, the risks they face – be they financial, legal, organisational, technical – one of the commonly used approaches is to take a **risk management approach**. This approach was originally developed for environments in which safety incidents may lead to severe consequences, and risks should, therefore, be minimized. Think for instance of aviation or traffic risks, or environmental risks, or occupational or epidemiological risks. Risk management generally consists of a number of steps: identifying risks, analysing them, assessing what their likelihood and impact is, developing means and methods to mitigate these risks, and monitoring whether these interventions are effective at reducing risks after their implementation (Berg 2010). The third step, risk assessment, is the step for which risk management is most widely known. It enables organisations to take a **quantitative** approach to risk. Generally, this is done using a formula to calculate risk, such as

$$Risk = Likelihood \times Impact$$

(or a more complex version thereof). This, in turn, may help decision makers to make rational judgements when they compare different (types of) risks, or technologies, or solutions.

While risk management originates in various subdomains of safety science, such as aviation and automobile engineering, its high practical relevance has led to a rapid spread of this approach to a wide variety of other fields, first solely focused on safety issues, but later also on security issues. As a matter of fact, in recent decades risk management has become the dominant paradigm for thinking about, and mitigating risk. It is not surprising that this is the case. Risk management approaches have proven to be highly effective in making airplanes and cars safer, in protecting our food against pathogens, or in reducing workplace accidents. Generally, one can argue that a risk management approach works best (1) in contained environments, (2) with limited complexity, and (3) where lots of information about past incidents and near misses is available. The bigger the complexity of a system, or the more networked or interconnected it is, the more **uncertainties** and **noise** appear in identifying, quantifying and weighing risks. This entails that the quantifications of risks produced under such conditions **provide less certainty**. Knowledge and experience also play a crucial role: the older a system or a technology is, the more we know about potential risks, for example on



Blad 25/65 the basis of past accidents or near misses. In relatively new domains, using a risk management approach leaves a much higher level of uncertainty.

It is especially this latter argument that is relevant for CBS. Cybersecurity risks in general are a relatively new phenomenon. They are very diverse and evolve very quickly. Our knowledge of cybersecurity risks is limited in the sense that we do not have large data sets to base our calculations on (De Bruijne and Van Eeten 2007). While organizations worldwide can learn from past incidents in cybersecurity, so far, due to the variety of incidents and the novelty and dynamism of the phenomenon, such learning is often more qualitative than quantitative in nature. For CBS, too, this means that **quantifying the risks of the remote access to microdata environment is an approach with limited utility** – or at least one **that offers limited certainty only**.

In order to come to a meaningful risk assessment and provide CBS with a future-proof approach – in the sense that CBS can also put it to use when this research project ends –, the Committee has chosen to use a **value-based** risk analysis as its main method. The starting point for a value-based analysis of risks is that to judge a certain event or action to be risky, it has to potentially affect something that is **valuable** to an actor. If CBS did not think of research or security as being important, then risks relating to these domains would immediately become less pressing. Risks need to be **contextualized** first, in order to prioritize them properly. Moreover, risks often bring forth a **tension** and/or competition between different values which an organization – often implicitly – both seeks to embrace. Think, for instance of the tension between security and user-friendliness. As shown in the table on page 21 of this document, limited checks after entry can be a risk for the value ‘security’, while simultaneously contributing to the value ‘user-friendliness’. In order to come to a grounded analysis of the identified risks, it is therefore essential that an organisation **first explicates which values it wants to hold high**. By making explicit which values an organization holds dear, it subsequently becomes possible to investigate what behaviors, processes, procedures and ways of working each value produces.

In order to assist CBS in explicating its **value profile**, this research identified the key values related to the remote access case and mapped them into three categories: **driving values**, **underpinning values**, and **process-based values** in WP4. One advantage of clustering values is that it is a **future-proof** approach. The underlying idea is that while the specific values may change over time, the **clusters themselves will remain stable**. In other words, by regularly



Blad 26/65 checking and updating the three clusters, an organization such as CBS can monitor in which way their values or, to put it differently, their goals and interests are impacted by the risks involved.

Once a value profile has been established, a first, general prioritization of risks can be carried out, based on the table on page 21 of this document, which provides an overview of how the identified risks impact the key values. However, the strong point of the value-based approach to risks – namely that it allows for a contextualized understanding of risks – also brings along one important challenge. The high number of possible value-profiles in relation to the various risks, makes it rather undoable to spell out all possibilities in detail. Therefore, it is key to complement a value-based risk approach with **scenario-building**. By making use of **scenarios**, it becomes possible to analyze how different behaviors, processes, procedures and ways of working may express a certain value profile.

In section 6 of this deliverable, **five scenarios** will be fleshed out as a way of illustrating how to move forward with this approach. While the choice for these three specific scenarios is not meant as a straightforward recommendation for CBS to opt for one of these, we also wanted to be pragmatic and choose scenarios of which we foresee that they might be on the mind of CBS.

Finally, it is important to emphasize that a value-based risk analysis should be understood as an **iterative process**. It might well be that, when investigating the behaviours, processes, and procedures needed to mitigate the risks impacting the value profile, one finds that these options are for instance too **costly** or **complex**. Consequently, an organisation will have to reconsider which values (and to what extent) it wants to hold high. In turn, this may result in changes in behaviours, processes, procedures and ways of working, leading organisations towards a more unified or coherent value expression.

5. Parameters of (remote) access to microdata at CBS

When viewing the remote access to microdata environment at CBS – and the risks that may be present in it – from a slightly higher level of analysis, one can decompose this system into a number of different, interconnected elements or **parameters**:

1. **End users:** *who* gets (remote) access to microdata, and under which conditions?
2. **Use:** what do end users *use microdata for*, what can they do with microdata, and under which conditions?
3. **Data sets:** which **data sets** does CBS own and operate, which data sets, containing microdata, are made accessible (remotely)?
4. **Processes and procedures:** which *processes* and *procedures* are in place for (remote) access to microdata, and to whom, for which purposes, and under which conditions do these apply?

These parameters can be considered as ‘**dials**’ in relation to specific risks, which CBS can turn to a more stringent or a more lenient position, or somewhere in between. For instance, with respect to end users, one of the risks is that access to microdata is granted to a diverse set of end users, ranging from researchers at universities and other research institutes in the Netherlands, to government organisations and commercial parties – after a rigorous entry process. The ‘dial’ that CBS can turn vis-à-vis this risk is to choose for a stricter versus a looser end user access policy, which would lead to a choice for less diversity versus more diversity in end users. The dial settings CBS can choose from are depicted in Figure 2 below.

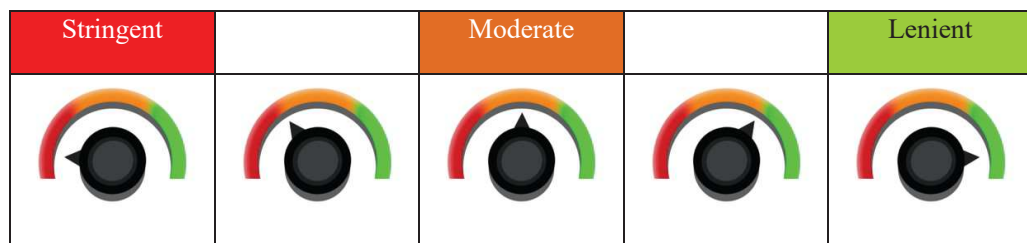


Figure 2: The spectrum of 'dials' that CBS can turn to more stringent or more lenient settings for each parameter.

When turning the dials, the choice for a particular setting of the dials for each parameter is an expression of an attachment to one or more **values**, for instance to more security or privacy (the more stringent side of the spectrum of dials), or to be available to more actors or use cases, or generally offer more possibilities for research on microdata (the more lenient side of the spectrum). For each individual risk, CBS will have to make a **value-based decision** on



Blad 28/65

where to set the dial. We note here that the more stringent options may severely affect the uses of microdata, and thereby the societal value derived from that use. To give an example from some of the settings mentioned below; limiting access to secure rooms would make the system much less easily accessible, but it would retain most of the functionality. However, removing access entirely would void certain use and thus come at a cost to the societal value derived from microdata. Collectively, the dial settings of the totality of choices CBS makes for all risks related to (remote) access to microdata express its **value proposition** with respect to that environment.

In this section we will describe each parameter in more detail, and summarize the main risks *and* potential settings for each dial in more detail. After that we will combine the dials and their resulting design choices in a single table (see Table 10 on page 44).

5.1 End users

Who are the end users?

End users are all users that populate the remote access environment at CBS. As explained in section 3, in the current constellation, there are five different types of end users:

1. (Dutch) universities;
2. Institutions for scientific research established by law, such as TNO in the Netherlands;
3. Planning offices and organisations for policy advice or policy analysis, established by or by virtue of the law, such as CPB or SCP in the Netherlands;
4. The statistical agency of the European Union (Eurostat) and national statistical agencies of the member states of the European Union and
5. Other institutions, most notably “*research departments of ministries and other departments, organisations and institutions*”⁶. On the website of CBS this latter category has the added qualifier that this pertains to institutions that are “...*authorised to work with the microdata.*”⁷

CBS has made a deliberate choice to offer remote access to a diverse population of researchers, not just at universities, but also at (specific) government organisations or (specific) private companies. The underlying idea is that the data that CBS collects, stores and

⁶ Statistics Netherlands Act (effective from 1 January 2017), Section 41, sub 2.

⁷ See <https://www.cbs.nl/en-gb/our-services/customised-services-microdata/microdata-conducting-your-own-research> (last accessed on 28 September 2020).



Blad 29/65

hosts, should be made **maximally useful for society**; high quality, detailed and reliable data are an essential ingredient for economic, political and societal innovation, and hence should be made available to this end to a variety of different parties. Microdata can help academics, decision-makers, and the public at large as a source of factual, objective information that can be used in public debates on a wide variety of challenging economic and societal questions. **Accessibility** is one of the key values that underpins the (remote) access to microdata approach at CBS.

What are the main risks related to end users?

At the same time, the Committee concludes that there are **two key risks** relating to the end users that currently populate the remote access environment:

1. First of all, there are risks to the end users presence on the system in a **remote** fashion. Because end users are not on premise at CBS when accessing data, or do not use a strictly controlled so-called safe room, it is difficult to prevent end users from abusing the system and, for instance, stealing or copying data. Due to the fact that the environment facilitates access **over distance**, from an end user's home or office space, or anywhere else, CBS has little control over, or insight into, the activities of end users, some of whom may have malicious intentions.

To reduce this risk, CBS has already implemented a great number of **technical** and **legal/procedural measures**. End users, for instance, can never download data from the environment, or copy data directly into documents or other files. All data that are made available via remote access stay in the CBS-environment and cannot be changed, moved, or copied. End users also sign a contract that stipulates that copying data is forbidden. Despite these measures, there is no way to prevent end users with bad intentions from taking **screen captures** or **making pictures** of their screens, or even copying the data by hand, thus obtaining these data for further use outside the remote access environment after all. Note that image processing techniques are advanced at the moment; extracting data from image files is straightforward.

2. Second of all, there are risks related to the **diversity of end users** that is currently allowed access to microdata. Here, the risk is not so much related to the remoteness of the connection, but to the **microdata itself**. In the current setup, there are five different categories of end users, and after proper screening, all of them get access to microdata. While the screening itself is thorough and solid, one could wonder about the demarcation



Blad 30/65

of the five different groups of end users, and especially about the composition of the group of **'other users'**. Over the years, the number of applicants that falls in this category has grown, and the diversity of organisations in this group has increased as well. It becomes ever more difficult to truly find clear grounds for acceptance or rejection for new applications in that light. This means that over time, the population of end users in the remote access environment may grow too big. This, in itself, is a risk because it may put constant pressure on CBS to expand capacity for the service, both in a technical sense and on an organisational level. To make matters worse, the diversification of end users in the system may make it harder for CBS to provide the right service levels for a wider variety of 'customers', and it also leads to ever more complex security and privacy risks.

How can the risks related to end users be reduced?

For each of the two risk categories in relation to end users, there are **interventions** available to address the challenge faced, with different levels of stringency:

1. To reduce the risks surrounding **malicious end users** that take advantage of the vulnerabilities in the system facilitated by **remote** access, the Committee proposes the creation of a system of **security level clearances**. In this system, end users will be subdivided into one of three different security level categories:

- Level 1: **basic** access level
- Level 2: **intermediate** access level
- Level 3: **advanced** access level

Security levels could be based on a variety of quality controls. CBS could factor in how long an end user has been a trusted user of the system, whether or not the end user works within the Netherlands (or the EU) only, or to which degree the organization (s)he works for has implemented proper privacy and/or cybersecurity controls (e.g. compliance with GDPR but also ISO27001/2 or COBIT or NIST or CIS controls etc).

Using these categories, CBS could distinguish between **different types of access** for end users, for instance by distinguishing between remote access, access via a safe room at the end users' workplace, or access via safe rooms at CBS locations only. One way of implementing this is depicted in Table 2 below.

Blad 31/65

Stringent		Moderate		Lenient
				
End users will be sub-divided into categories based on their security level clearance. End users with level 3 clearance will have remote access to data; all other end users do not get access to microdata at all.	End users will be sub-divided into categories based on their security level clearance. End users with level 3 clearance will have remote access to data; end users with level 2 clearance will have access to data via secure rooms ; all other end users do not get access to microdata.	End users will be sub-divided into categories based on their security level clearance. End users with level 3 clearance will have remote access to data; end users with level 2 clearance will have access to data via secure rooms ; end users with level 1 clearance must come to a CBS location to access microdata; end users without clearance do not get access to microdata.	End users will be sub-divided into categories based on their security level clearance. End users with level 2 and 3 clearance will have remote access to data; end users with level 1 clearance will have access to data via secure rooms ; end users without clearance may come to a CBS location to access microdata.	No sub-division will be made between end users in terms of security level clearance. All end users may access microdata remotely , via secure rooms or they may come to a CBS location to access microdata.

Table 2: Diversifying access.

2. To reduce the risk of **too many end users**, or a **too diverse population** in the remote access environment at CBS, the Committee proposes that CBS primarily rethinks the category '**other users**', since this category is the most important source of diversification and risk. One solution would be to rethink access to microdata in the first place, and perhaps reserve it for researchers at Dutch universities and research institutes only (stringent), or, if a more lenient approach is desirable, to at least develop more specified, detailed procedures to verify which government and/or commercial parties will be granted access. In this way, the category 'other users' becomes more circumscribed, and may be replaced by more clearly designated (categories of) end users, and in the process, access will become stricter as well. One way of implementing this is depicted in Table 3 below.

Blad 32/65

Stringent		Moderate		Lenient
				
End users do not have (remote) access to microdata.	Only Dutch universities and research institutes may have (remote) access to microdata.	Only Dutch universities, research institutes, and verified government organizations may have (remote) access to microdata.	Only Dutch universities, research institutes, verified government organizations and verified commercial parties may have (remote) access to microdata.	After a fundamental rewrite of the CBS law, anyone might get (remote) access to microdata.

Table 3: End user access to microdata.

5.2 Use

What use is made of microdata using remote access?

CBS has a large variety of data sets offered to end users who may utilize the data for the following reasons:

1. Research purposes,
2. (Statistical) analysis for decision making by governmental organizations and,
3. Commercial parties offering statistical analyses for the benefit of society.⁸

CBS has intake procedures in which the purpose of remote access is investigated.

What are the main risks related to the use of microdata via remote access?

The Committee considers the following risk categories in relation to the **use of microdata** at CBS:

1. One area of particular concern on the use of microdata is the ability to **uniquely identify individuals or organizations** within data sets. While CBS removes key identifiers and pseudonymizes data before making data sets accessible to end users, it is possible to use **correlation techniques** between different data sets to identify and track individuals and/or organisations in the microdata. Identification is made easier when CBS data are combined with data from end users' own data sets, or with publicly known information,

⁸ Note: using the remote access to microdata environment for all other commercial purposes is not allowed under the CBS law.



Blad 33/65

e.g. online social media. One of the reasons why malicious end users might wish to engage in this kind of behaviour is to engage in **fraud or extortion**. A more benign reason might be **curiosity**, for instance looking up well-known individuals or large organisations. The risk of identification is aggravated when **data sets are small**, e.g. the **sample size** in the study an end user is conducting is **low**. In these cases, especially when end users combine the data sets with their own data, it is feasible to uniquely identify individuals and organizations and potentially also enrich one's own data sets with data found in the data sets at CBS by copying it manually.

2. A second risk relating to the use of microdata pertains to the CBS rule that end users must **publish the results** of research they have conducted on or with microdata. While there is a requirement to publish the results from research on microdata, checking whether this is actually the case places a heavy administrative burden on CBS, especially with a growing user base. Moreover, people with malicious intent could get access under the guise of a genuine research project, and publish the findings as agreed upon when granted access, but they could still run a wide variety of other analyses that suits their own purpose alongside the genuine research project. Additionally, end users could forget, or misinterpret, the publication requirement, and accidentally choose to publish only a subset of their findings, while using the rest at a later point in time.

How can the risks related to the use of microdata via remote access be reduced?

For each of the two risk categories in relation to end users, there are **interventions** available to address the challenge faced, with different levels of stringency:

1. To reduce the risks surrounding the **identification of individuals or organisations** within data sets, the Committee recognizes that complete elimination of this risk is not feasible, other than by ending access to microdata for all end users completely – which is not a realistic or desirable option in light of the high societal value of the microdata environment. What CBS can do, however, is seek to reduce the risk of identification. One of the most straightforward ways of doing so is to reconsider the aggregation level that end users are allowed to report on. Currently, the minimum aggregation level is $n = 10$. This means that when users export the findings of their analysis, each cell or reported statistic (such as for example a mean score, average increase, or a standard deviation) needs to report on 10 or more cases (such as 10 or more households, individuals, SMEs

Blad 34/65

etc.) in order for the user to be able to export the findings to locations outside the microdata environment. By raising this threshold, it would become harder to uniquely (re)identify individuals and/or organisations. For studies in which a high level of granularity is desired, this is problematic, but it would increase privacy and security protection to take this measure. One way of implementing this suggestion is depicted in Table 4 below.

Stringent		Moderate		Lenient
				
For any study the aggregation level n has to be bigger than 100.	For any study the aggregation level n has to be bigger than 50.	For any study the aggregation level n has to be bigger than 25.	For any study the aggregation level n has to be bigger than 10.	For any study the aggregation level n has to be bigger than 5.

Table 4: Reducing the risk of identification.

2. To address this risk of avoiding the rule that results should be published CBS could rethink its publication verification strategy. It could make requirements more stringent, for instance by actually checking the content of all publications that end users generate with their research on microdata, or by doing random checks on the content of a subset of these publications. It could also make requirements more lenient, for instance by asking for a publication plan only, without verifying actual publications (or bibliographical information on publications) at all. One way of implementing this is depicted in Table 5 below.

Blad 35/65

Stringent		Moderate		Lenient
				
CBS requires proof of all publications containing research on microdata. CBS checks these publications to see whether the output data are correctly presented. If this is not the case, end users' credentials to the remote access system are revoked.	CBS requires proof of all publications containing research on microdata. CBS randomly selects a sub-set of all publications by different end users and checks these publications to see whether the output data are correctly presented. If this is not the case, end users' credentials to the remote access system are revoked.	CBS regularly asks random researchers to provide proof of the publications containing research on microdata. CBS checks these publications to see whether the output data are correctly presented. If this is not the case, end users receive a one-time warning ; if a second check leads to the same findings, the end user's credentials to the remote access system are revoked.	CBS does not require proof of all publications that use findings on the basis of microdata. End users must submit the title and an abstract of each publication , along with publication details such as the publication outlet. CBS checks these details to ensure that its publication requirements have been met, but does not check the content of publications.	CBS does not require proof of any publications that use findings on the basis of microdata. End users must have a publication plan , but the actual publications themselves are not checked .

Table 5: Publication verification strategy.

5.3 Data sets

Which data sets are used in the remote access to microdata environment?

Microdata are stored in, and made available remotely, via data sets. In the remote access to microdata environment, two different types of data sets can be distinguished:

1. Data sets that are added after an end user has requested access. These data sets are then made available to other users as well; and
2. Data sets that end users bring themselves.

Data sets are prepared especially for the remote access environment, so that no identifiable data regarding natural persons or organizations is in there.

What are the main risks regarding data sets?

The Committee considers the following risk categories in relation to the **data sets** at CBS:

1. In the current remote access to microdata environment all data sets are ‘treated equally’ with respect to the privacy and security risks of the data they hold. CBS currently **does**



Blad 36/65

not perform risk assessments on the data sets (but it does perform high quality risk assessments on the environment itself). No distinction is made between data sets that might contain more sensitive data compared to data sets that contain less sensitive data. Because all data are treated equally, there is a risk that more sensitive data is made available on an equal footing, indiscriminately, with non-sensitive or less-sensitive data. This may lead to insufficient insights into what the risks of providing access to each specific data set actually are, and also to what risks combinations of data sets may give rise.

2. CBS currently uses **pseudonymization** to remove identifiers from data sets and protect individuals' privacy, but not **anonymization**. This is acceptable, since the assumption is that the data sets are to be used by trusted researchers, and since using current anonymization techniques would run the risk of hampering usability to an unacceptable degree, thus undermining the societal and economic value of remote access to microdata. At the same time, there is a chance to use **quasi-identifiers** on these data sets to uniquely identify individuals. Particularly, with the information available online, deanonymization is not as difficult as it is believed.

How can the risks relating to data sets be reduced?

The Committee suggests the followings to address the risks relating to **data sets** that were identified:

1. A thorough risk assessment of data sets should be conducted. Based on this assessment, a **data classification** could be made, in which **data sets can be labeled** according to the **sensitivity** of the data they contain. Four levels of sensitivity could be distinguished:
 - Highly sensitive
 - Sensitive
 - Confidential
 - Internal⁹

⁹ In data classifications, the terms 'public', 'confidential', 'sensitive' and 'highly sensitive' are commonly used. The Committee has chosen to replace the term 'public' with 'internal' to ensure that there is no misunderstanding over the fact that microdata can always only be accessed by end users that have been granted access to the remote access environment, not by all citizens. Should the use of this term lead to confusion, CBS could also choose to label data using another categorization, even as simple as 'level 1' (no restrictions), 'level 2' (some restrictions), etc.

Blad 37/65

Note that the term 'internal' here means that data with this label may be made available to all end users that have been accepted into the remote access environment, without further requirements.

CBS could then choose to make data sets with different classifications available to different types of end users, or at different locations (remotely, in safe rooms, on premise at CBS only). It could even make data sets available using the parameter of time, whereby researchers only get access to data sets of a more stringent class after they have proven to be trustworthy end users for a longer period of time. One way of implementing this suggestion is depicted in Table 6 below.

Stringent		Moderate		Lenient
				
Only data sets that CBS has published will be made available via remote access.	Data sets will be sub-divided into categories based on their level of sensitivity, using the labels 'highly sensitive', 'sensitive', 'confidential', and 'internal'. Data sets with the labels 'highly sensitive', 'sensitive' and 'confidential' will not be made available through the remote access system; data sets with the label 'internal' will be made available via remote access.	Data sets will be sub-divided into categories based on their level of sensitivity, using the labels 'highly sensitive', 'sensitive', 'confidential', and 'internal'. Data sets with the labels 'highly sensitive' and 'sensitive' will not be made available through the remote access system; data sets with the labels 'confidential' and 'internal' will be made available via remote access.	Data sets will be sub-divided into categories based on their level of sensitivity, using the labels 'highly sensitive', 'sensitive', 'confidential', and 'internal'. Data sets with the label 'highly sensitive' will not be made available through the remote access system; data sets with the labels 'sensitive', 'confidential' and 'internal' will be made available via remote access.	All data sets will be made available via remote access.

Table 6: A data classification for data sets.

- To address the risks relating to pseudonymization, the Committee recommends that CBS consider additional methods to provide increase privacy protection in data sets, for instance by applying **anonymization techniques** where possible (without too much negative impact on usability), by using **differential privacy** or by using advanced **cryptographic tools**. These techniques may not be necessary for all data sets; they are



Blad 38/65

especially relevant for data sets that contain (highly) sensitive data. Using the data classification discussed above, CBS could make a distinction between the level of sensitivity of data sets and apply more stringent techniques of anonymization, differential privacy and/or cryptography to those data sets that are most sensitive, leaving other data sets in their current (pseudonymized) state. One way of implementing this idea is depicted in Table 7 below.

Stringent		Moderate		Lenient
				
CBS will anonymize all data sets .	Data sets will be sub-divided into categories based on their level of sensitivity, using the labels 'highly sensitive', 'sensitive', 'confidential', and 'internal'. Data sets with the labels ' highly sensitive ', ' sensitive ' and ' confidential ' will be anonymized ; data sets with the label 'internal' will be pseudonymized.	Data sets will be sub-divided into categories based on their level of sensitivity, using the labels 'highly sensitive', 'sensitive', 'confidential', and 'internal'. Data sets with the labels ' highly sensitive ' and ' sensitive ' will be anonymized; data sets with the labels 'confidential' and 'internal' will be pseudonymized.	Data sets will be sub-divided into categories based on their level of sensitivity, using the labels 'highly sensitive', 'sensitive', 'confidential', and 'internal'. Data sets with the label ' highly sensitive ' will be anonymized; data sets with the labels 'sensitive', 'confidential' and 'internal' will be pseudonymized.	CBS will pseudonymize all data sets.

Table 7: Anonymizing data sets.



Which processes and procedures surround the use microdata via remote access?

There are various processes and procedures that are relevant to microdata services. The most important ones for this document are those related to:

1. Getting access to microdata via a remote connection, at the levels of individual users and their institutions;
2. Setting up a specific research project, including requesting specific data sets and potentially bringing in own data;
3. Logging into and using the remote access environment;
4. Checks on the output; and
5. Monitoring and acting on violations.

These processes and procedures have been extensively described in the other deliverables.

Once parties have access, there are **rules** that institutions and individual users must abide by. There are **sanctions** for not following them, and measures in case of undesirable events, security incidents and data breaches.

What are the main risks related to the processes for remote access?

CBS performs a significant number of **checks** before end users are **allowed access** into the microdata environment. They have procedural, technical and other means in place to also ensure that no data leave the remote access environment, and that there is no risk to privacy in the findings that end users generate with their research by performing output controls. CBS monitors what researchers do at the beginning and the end of their research. Whilst these processes and procedures are well designed, the Committee identifies several potential risks associated with them:

1. First, the processes are **resource intensive** in their current form. Preparing data sets for research projects, technical support of researchers, and checking output are time-consuming and require the involvement of experts at CBS. Not only is this costly, it also requires sufficient professionals with the right skills. The increase in number and diversity in users means there is higher demand on the (limited) capacity towards these key processes. A number of the options also presented in this document (e.g. increased monitoring, wider and more frequent checks on data sets) would further increase the workload and thereby need for capable professionals. These professionals cannot be



Blad 40/65

found overnight and a high workload might increase risk of errors. Increasing the staff size would also lead to increased costs.

2. Second, there are assurances that the institution requesting access is truly research focused and will publish their findings as per CBS standard. Moreover, the institution itself has to get a license, as do the individual researchers for a project, and projects themselves need to be approved as well. These processes are thus heavy on granting access, but **once institutions and users are on the system, the number of checks go down**. Institutional access is also for quite substantial time (usually 5 years), and the diversity of users (see before) makes entry and extensions sometimes complicated as research roles can be complicated within institutions. Once access is granted, however, there is no active review of the user groups; it is mostly left to how the institutions organise this. All of this introduces risks that users may be able to continue to have access to projects that they no longer work on at their institution. Beyond access the **monitoring of the activities of individual users is limited**, relying mostly on trust in their institutions and on the threat of measures should users violate the terms of use. This may not be effective to safeguard against users that do have a valid project but hidden malicious intent and/or are displaying atypical behaviour during use.

How can the risks related to the processes for remote access be reduced?

Reducing the risks that are associated with processes and procedures involves making a **trade-off** with how **effective** measures are expected to be and respecting that procedural measures are usually **resource-intensive**, especially on specialized staff. The committee sees the following solutions that might be actionable:

1. To address the issue of **workload** for an ever-busier remote access environment, CBS could introduce a form of **protective monitoring**, that continuously **scans the remote access system and users**, (automatically) looking for unusual patterns of behaviour. A milder form of this is to **log all user behaviour**¹⁰. By outsourcing the responsibility for surveillance and logging to technical systems, no further pressure is put on the workload of staff within CBS responsible for maintaining and servicing the remote access to microdata environment, while security is increased. Obviously, however, monitoring and recording end user behaviour may lead to ethical issues as well. End users may not

¹⁰ Some user behavior is currently already logged.

Blad 41/65

appreciate that their every move on the system is observed. CBS could therefore develop a **classification of activities** that are deemed more or less ‘risky’ with respect to the privacy or security threats they may induce. For instance, using statistical software on a CBS data set could be considered a low-risk activity, whereas combining a CBS database with the end user’s own data set could be considered a higher risk. CBS could classify all activities that end users can undertake within the remote access environment on three levels:

- Level 1: low/no risk to privacy and/or security
- Level 2: intermediary risk to privacy and/or security
- Level 3: high risk to privacy and/or security.

Next, CBS could choose to undertake **different actions** for activities of different levels. For instance, it could **surveil and log all activities of level 2 and 3**, while only **logging activities of level 1**. One way of implementing this suggestion is depicted in Table 8.

Stringent		Moderate		Lenient
				
CBS will automate surveillance. It will surveil and log all activities in the remote access environment.	Activities in the remote access environment will be sub-divided into categories based on their security and privacy risk level. Level 2 and 3 activities will be surveilled and logged ; level 1 activities will only be logged .	Activities in the remote access environment will be sub-divided into categories based on their security and privacy risk level. Level 3 activities will be surveilled and logged ; level 2 activities will be logged ; level 1 activities will neither be surveilled nor logged.	Activities in the remote access environment will be sub-divided into categories based on their security and privacy risk level. Level 3 activities will be surveilled and logged ; level 1 and 2 activities will neither be surveilled nor logged.	CBS will not conduct intermediary testing or monitoring beyond access logs.

Table 8: Surveilling and logging behavior in the remote access to microdata environment.

2. To address the risk of a **lack of controls after end users have gained access to the system**, CBS could take a number of different measures, partially intensifying processes and procedures internally, but partially also outsourcing procedures to end users themselves or the organizations they work for. Currently, the processes within CBS rely heavily on trust in how the institutions organize their internal processes. This could be

Blad 42/65

more hands-on, by informing or **soliciting feedback from the institutions** in a way that empowers them to **play the gatekeeper function** that they effectively have. An easy measure would be to send the institutions a list of active users regularly, based on the monitoring CBS already has in place. This would allow the institutions to organize the internal checks needed to maintain restrictions on the user base and to reduce the risk that user accounts are longer open than needed. Stricter measures that improve control on users after the initial (and strict) entry procedures, could include **site visits, audits, shorter institutional access periods** or introducing **mid-term reviews**. This could also be related to differentiation in user categories. A second form of increased procedural control could include the introduction of **mystery shoppers** that explore the procedures or an **‘easter eggs’ approach** to introduce known mistakes. These can be used to test if staff correctly flags problems and to see processes and procedures from a user’s perspective in an attempt to improve them. A third option would be to **introduce blameless post-mortems** by both **users** and **staff**, incentivising reporting of incidents or vulnerabilities, and allowing CBS to learn from them without entering the blame game. One way of implementing these suggestions is depicted in Table 9 below.

Stringent		Moderate		Lenient
				
Controls on institutions are intensified, including site visits, active user management, mystery shoppers and blameless post-mortems .	Measures within CBS put in place to test quality of internal procedures and staff handling them. Reducing institutional access periods or introducing mid-term reviews .	Current procedures stay in place but institutions are better supported in their role as gatekeeper .	Current processes and procedures emphasize checks on entry and trust after .	Checks and barriers on entry are lowered , consequently relying on controls on data sets and on output for mitigating risks.

Table 9: Controls before and after access.

5.5 Combining all parameters

In the previous pages, the Committee has presented a variety of **solutions** to the risks it had reported in WP4. Using the ‘dials’ to respond to each risk entails that CBS has options to choose a more stringent or a more lenient approach for each specific risk. Collectively, these



Blad 43/65 choices lead to an **overall risk profile for the remote access to microdata environment** at CBS. On the next page, we summarize all the presented measures in a single table.



		Stringent		Moderate		Lenient
Parameters:						
End users	1. End users may steal or copy data, or behave in other untrustworthy ways when accessing microdata remotely.	End users will be sub-divided into categories based on their security level clearance. End users with level 3 clearance will have remote access to data; all other end users do not get access to microdata.	End users will be sub-divided into categories based on their security level clearance. End users with level 3 clearance will have remote access to data; end users with level 2 clearance will have access to data via secure rooms; all other end users do not get access to microdata.	End users will be sub-divided into categories based on their security level clearance. End users with level 3 clearance will have remote access to data; end users with level 2 clearance will have access to data via secure rooms; end users with level 1 clearance must come to a CBS location to access microdata; end users without clearance do not get access to microdata.	End users will be sub-divided into categories based on their security level clearance. End users with level 2 and 3 clearance will have remote access to data; end users with level 1 clearance will have access to data via secure rooms; end users without clearance may come to a CBS location to access microdata.	No sub-division will be made between end users in terms of security level clearance. All end users may access microdata remotely, via secure rooms or they may come to a CBS location to access microdata.
	2. Access to microdata is granted to all parties equally; there is no diversified access to microdata.	End users do not have (remote) access to microdata.	Only Dutch universities and research institutes may have (remote) remote access to microdata.	Only Dutch universities, research institutes, and verified government organisations may have (remote) access to microdata.	Only Dutch universities, research institutes, verified government organisations and verified commercial parties may have (remote) access to microdata.	Anyone may have (remote) access to microdata.
Use	1. Microdata can be used to identify individuals or organisations uniquely, for instance in small data sets, but also through correlation-based attacks.	For any study the aggregation level n has to be larger than 100.	For any study the aggregation level n has to be larger than 50.	For any study the aggregation level n has to be larger than 25.	For any study the aggregation level n has to be larger than 10.	For any study the aggregation level n has to be larger than 5.
	2. End users may violate the publication requirements of CBS in several ways.	CBS requires proof of all publications containing research for which microdata have been used. CBS checks these publications to see whether the output data are correctly presented. If this is not the case, end users' credentials to the remote access system are revoked.	CBS requires proof of all publications containing research for which microdata have been used. CBS randomly selects a sub-set of all publications by different end users and checks these publications to see whether the output data are correctly presented. If this is not the case, end users' credentials to the remote access system are revoked.	CBS regularly asks random researchers to provide proof of the publications containing research for which microdata have been used. CBS checks these publications to see whether the output data are correctly presented. If this is not the case, end users receive a one-time warning; if a second check leads to the same findings, the end user's credentials to the remote access system are revoked.	CBS does not require proof of all publications that use findings on the basis of microdata. End users must submit the title and an abstract of each publication, along with publication details such as the publication outlet. CBS checks these details to ensure that its publication requirements have been met, but does not check the content of publications.	CBS does not require proof of any publications that use findings on the basis of microdata. End users must have a publication plan, but the actual publications themselves are not checked.
Data sets	1. Data sets may have different levels of sensitivity; not all datasets should be treated equally or be made equally accessible.	Only data sets that CBS has published will be made available via remote access.	Data sets will be sub-divided into categories based on their level of sensitivity, using the labels 'highly sensitive', 'sensitive', 'confidential', and 'public'. Data sets with the labels 'highly sensitive', 'sensitive' and 'confidential' will not be made available through the remote access system; data sets with the label 'internal' will be made available via remote access.	Data sets will be sub-divided into categories based on their level of sensitivity, using the labels 'highly sensitive', 'sensitive', 'confidential', and 'public'. Data sets with the labels 'highly sensitive' and 'sensitive' will not be made available through the remote access system; data sets with the labels 'confidential' and 'internal' will be made available via remote access.	Data sets will be sub-divided into categories based on their level of sensitivity, using the labels 'highly sensitive', 'sensitive', 'confidential', and 'public'. Data sets with the label 'highly sensitive' will not be made available through the remote access system; data sets with the labels 'sensitive', 'confidential' and 'internal' will be made available via remote access.	All data sets will be made available via remote access.
	2. CBS uses pseudonymization techniques but not anonymization techniques. This leads to privacy risks.	CBS will anonymize all data sets.	Data sets will be sub-divided into categories based on their level of sensitivity, using the labels 'highly sensitive', 'sensitive', 'confidential', and 'public'. Data sets with the labels 'highly sensitive', 'sensitive' and 'confidential' will be anonymized; data sets with the label 'internal' will be pseudonymized.	Data sets will be sub-divided into categories based on their level of sensitivity, using the labels 'highly sensitive', 'sensitive', 'confidential', and 'public'. Data sets with the labels 'highly sensitive' and 'sensitive' will be anonymized; data sets with the labels 'confidential' and 'internal' will be pseudonymized.	Data sets will be sub-divided into categories based on their level of sensitivity, using the labels 'highly sensitive', 'sensitive', 'confidential', and 'public'. Data sets with the label 'highly sensitive' will be anonymized; data sets with the labels 'sensitive', 'confidential' and 'internal' will be pseudonymized.	CBS will pseudonymize all data sets.
Processes & procedures	1. There is high pressure on CBS in terms of workload to manage the use of the remote access environment in a safe way, especially considering its rapid growth.	CBS will automate surveillance. It will surveil and log all activities in the remote access environment.	Activities in the remote access environment will be sub-divided into categories based on their security and privacy risk level. Level 2 and 3 activities will be surveilled and logged; level 1 activities will only be logged.	Activities in the remote access environment will be sub-divided into categories based on their security and privacy risk level. Level 3 activities will be surveilled and logged; level 2 activities will be logged; level 1 activities will neither be surveilled nor logged.	Activities in the remote access environment will be sub-divided into categories based on their security and privacy risk level. Level 3 activities will be surveilled and logged; level 1 and 2 activities will neither be surveilled nor logged.	CBS will not conduct intermediary testing or monitoring beyond access logs.
	2. CBS performs limited internal checks once users are in the system (no surveillance, limited logging).	Controls on institutions are intensified, including site visits, active user management, mystery shoppers and blameless post-mortems.	Measures within CBS put in place to test quality of internal procedures and staff handling them. Reducing institutional access periods or introducing mid-term reviews.	Current procedures stay in place but institutions are better supported in their role as gatekeeper.	Current processes and procedures emphasize checks on entry and trust after.	Checks and barriers on entry are lowered, consequently relying on controls on data sets and on output for mitigating risks.

Table 10: The 'dials' CBS can turn for each parameter combined.



5.6 Residual risk

One point to note is that CBS can use the parameters discussed in this chapter to reduce a variety of risks, but that it is **impossible to eliminate many, if not all, risks entirely**. In some cases, eliminating risk is impossible per se, while in others it would require unreasonably large investments to be made, both in terms of money and effort. Sometimes, also, eliminating a risk from a security or privacy perspective could lead to unwanted side-effects, as was exemplified in this report, for example, when discussing the tension between anonymization of data (sets) and usability, i.e. the public value of access to microdata.

To find the right optimum between treating risks and keeping an eye on the costs this brings along (in the broadest sense of the word), CBS ought to establish how, and to which degree, those risks that it deems most important can be reduced to so-called ‘**acceptable risk levels**’ (Aven 2014). Once an acceptable risk level has been reached, what remains is **residual risk**, which needs to be accepted as part and parcel of (in this case) offering (remote) access to microdata to parties outside CBS. The Committee lists the following risks as **residual**:

1. When working with data sets, there is no way to exclude in full that an employee gains unlawful access to the data or uses the data for other purposes than (s)he is authorised/cleared for. Mitigating measures can never block employees with malicious intent from behaving in illegal ways.
2. Every database can be hacked, however strong encryption measures and security standards may be. Mitigation measures can only make it harder to hack into the system or make it harder to use the data in a meaningful way.
3. Anonymised data can almost always be de-anonymised, if enough resources, time and efforts are spent. Mitigation measures can only make it harder to de-anonymise data.
4. Remote access always entails a danger of third parties trying to (manually or automatically) copy data, e.g. through screen-casting. Completely excluding this possibility would require such control over the environment in which the system is used that it might negate the point of having a remote access system in place.
5. Especially once a user knows specific details about a person (e.g. through bringing in their own data), this person could be identified if the researcher has access to a relevant set of data. Awareness campaigns and background checks may help but, technological measures and screenings can never exclude this possibility in full.



Blad 46/65

Taking these residual risks as a given, as long as CBS seeks to offer access to microdata in any shape or form, using combinations of the dials offered in this document CBS can choose a variety of different risk profiles. In the next section we present a number of scenarios to showcase some of the choices available.



6. Scenarios for access to microdata at CBS

In the previous section we explained how CBS can use a system of dials, ranging from stringent to lenient, to address each specific risk that the Committee has identified in WP4. Collectively, these dials lead to a chosen **risk profile** for the remote access to microdata environment, and they express a specific **value proposition** for the organisation. In this section, we present a number of different **scenarios** to show what the **combined** choices on risk settings would lead to. Of course, using the dial system can lead to a very large number of possible scenarios emerges. Not all combinations of dial settings are equally likely to be chosen, but there is room for nuance in choosing them. The presented scenarios below, therefore, should be considered **examples** only. They are not suggestions on the part of the Committee on the preferability of choices to be made. The discussion of which dials to turn to which setting (i.e. more stringent/lenient than, or the same as is currently the case), is an internal matter to be resolved by the CBS. In this section, the Committee wants to illustrate some of the possibilities by discussing five possible scenarios:

1. A stringent scenario,
2. A lenient scenario and
3. Three diversification scenarios (moderate in all, stringent on access, stringent on data sets).

Below, each of these scenarios is discussed, with an overview table per scenario at the end of every section. Factors that are not considered in the dials table, but that are of relevance, are the **costs**, **capacity** and **time** required to implement and maintain each scenario. The Committee has left these factors outside their considerations, since the research conducted did not provide it with sufficient knowledge and insight into the organisation to warrant solid deliberation with regard to these areas.

6.1 A stringent scenario

In the stringent scenario, the dials on all factors discussed in section 5 are pointing at the **two most left columns** in Table 10 ('The 'dials' CBS can turn for each parameter combined.' on page 44). In this scenario, the values of **privacy**, **data protection** and **security** are chosen over the values of **accessibility**, **user-friendliness** and **trust**.



Blad 48/65

End users

In the stringent scenario, there would be strict rules and regulations as to who can access the micro-data. For instance, the ‘**other**’ category of users is clearly defined and only organisations that fit exactly into that category are allowed access. Similarly, there are strict norms regarding the **security and privacy clearance** that the end users have. Only end users with a high level of clearance may have **remote** access to the data, and end users a medium level clearance may only access data via secure rooms or on the CBS premises. Furthermore, only users directly affiliated with **Dutch organisations** are allowed access to the data. **Commercial** parties do not have access to the data, as making sure that these organisations do not use the data for other purposes is difficult. Once set, access levels are stable over time.

Use

In this scenario, end users can never export findings that report on groups with a **sample size $n < 50$** and the publication requirements are stringent: CBS checks the **content of all publications** based on research using microdata, or at least checks the content via randomly selected publications.

Data sets

In this scenario, for each data set, a **risk assessment** is carried out and a **data classification system** is developed. Depending on the sensitivity of the data set, there are different locations in which data can be viewed and used, so that the most **sensitive** data can only be accessed at one of the **CBS locations**; **confidential** data can be accessed in a **secure** room within the user’s organisation, and only the least sensitive data, labelled ‘**internal**’ can be accessed **remotely**, in a manner similar to the current remote access procedures. For all these categories, the rule is that users can only access the data while being physically in the Netherlands.

Processes and procedures

To ensure that the data is handled and accessed correctly, CBS carries out announced and unannounced **site visits** or **audits**, and it **monitors** and **logs** user **behaviour** in the remote access environment. Stringent policy requirements both before being granted access, as well as once access has been granted are implemented.

This scenario focuses on reducing the risks of the remote access to micro-data to the greatest extent possible. It does so by putting more checks in place to ensure the correct use of the



Blad 49/65

microdata, as well as by reducing the variety of organisations who are allowed access to the data, and reducing the number of data sets that users would have access to. While this improves the security of the micro-data, and reduces unwanted access to microdata, or data falling in the wrong hands, there are also clear downsides. The related **costs are relatively high**, as users might need to set up secure rooms, CBS needs to create workspaces for users requiring access to the most sensitive type of data, and site visits and online monitoring can also be time intensive. Additionally, some organisations that in the current situation would have access to the micro-data environment are **denied access** as the ‘other’ group of users is more stringently defined and only users in The Netherlands are allowed access in the first place. Hence the conclusion is that this scenario benefits the values of privacy and security, but hollows out those of accessibility and trust.

Parameter	Choice	
End users	User categories:	Strict specification of the category ‘other’; no one outside that specification is allowed access. Commercial parties are not granted access. Users must be affiliated with a Dutch organization.
Use	Use categories:	In all data sets the aggregation level for exporting findings outside the microdata environment is $n > 50$, and CBS checks the content of (almost) all publications using microdata.
Data sets	Data categories:	Strict specification of the categories of data sets: ‘internal’, ‘confidential’, ‘sensitive’ and ‘highly sensitive’. Remote access is allowed only for the category ‘internal’. The category ‘confidential’ may be accessed only via a safe room in an end user’s organisation. The categories ‘sensitive’ and ‘highly sensitive’ may be accessed only at one of the CBS locations. AND users may access data in any category only while on Dutch territory.
Processes	Site visits & surveillance:	End users will receive announced and unannounced site visits to check on their use of the microdata environment. There is a strict policy before entering the remote access system, but also after admission. Surveillance and logging of all activities.



6.2 High accessibility

The radical opposite of the stringent scenario would be to set all dials to a **lenient** setting. In this scenario the dials would be placed in the two columns on the right of Table 10 ('The 'dials' CBS can turn for each parameter combined.' on page 44).¹¹ It would favour **accessibility, user-friendliness and trust** over **privacy, data protection and security**.

End users

All users are **treated equally in terms of access**, and the 'other' category of users is interpreted **broadly** so that a wide range of organisations can get access to the microdata. This includes opportunities for users to conduct statistical research for commercial projects, but also allows citizens to access the microdata as well as organisations outside the EU. In the long term, CBS could opt to release all (anonymised) microdata for interested individuals and parties. Remote access for all microdata is possible, without restrictions on sensitivity once researchers have been accepted into the environment.

Use

The aggregation levels of data before they can be released to users are **as low as possible**, but not lower than $n > 5$, to facilitate small-scale research. There is no hard publication requirement. End users only need to deliver a publication plan, but the execution of this plan is not verified by CBS.

Data sets

No diversification of access based on data set **sensitivity** is required, this is both true for general access, as well a remote access to these data sets. **No data classification** needs to be made. CBS only ensures that data sets are pseudonymized before release and verifies that no personally identifiable data is in these data sets.

Processes and procedures

Onboarding and offboarding processes stay as they are now, and CBS does not conduct site visits or monitor online activity, unless there are **concrete complaints** or other information suggesting that this is required.

¹¹ Note: implementing this scenario in full would lead to a requirement to adjust the legal framework under which CBS, and the remote access to microdata environment with it, currently operate.



Blad 51/65

This scenario focuses on making the data **as useful as possible** for societal and commercial benefit. The **direct costs** for this scenario are low, as there are no proposed implementations of new policy, rules, regulations or guidelines that are time intensive. As a matter of fact, some of the existing procedures are abandoned, such as the time-intensive publication check. This scenario would improve the use of the microdata compared to the current situation, with more parties gaining access, and the yield of microdata would be improved by reduced limitations on the output that is allowed to leave the remote access environment. However, this scenario would also lead to increased risk of **identifying individuals**, and parties with **malicious intent** being able to access the microdata for their own purposes. Privacy of citizens and organizations would be reduced, and the potential for cyberthreats would rise.

Parameter	Choice	
End users	User categories:	A similar regime applies to all user types. A broad interpretation the ‘other institutions’ is opted. For example, statistical research for commercial purposes would be allowed. The requests of citizens who want to perform statistical research may also be accepted. Moreover, organizations from outside the EU may also be allowed access. Ultimately, CBS could choose to move towards an open access model for their (anonymized) microdata. Remote access is possible and all parties have access to the microdata directly.
Use	Use categories:	All types of use are allowed. Technical safeguards provide a basic level of security, but do not hamper remote access and usability. Most importantly, the aggregation level of results that need to be exported to outside the microdata environment is as low as $n > 5$.
Data sets	Data categories:	No data classification is made and all data are considered equally (non-)sensitive. All data are pseudonymized and made accessible remotely.
Processes	Site visits & surveillance:	CBS will not surveil or log end user behaviour, and will not conduct audits, unless there are concrete complaints or signs of actual abuse.



6.3 Moderate in all

The preceding two scenarios could be considered extreme options, based on a focus on either privacy and trust, or on accessibility and usefulness of the data that is being shared. However, a range of less extreme scenarios are also possible. Those scenarios involve the use of **diversification of access, data sets and processes**, that can be **tailored to balance the values that CBS decides to follow**. As an example, we first present a scenario where all dials are set to **'moderate'**. In this scenario the dials would be placed in the middle columns of Table 10 ('The 'dials' CBS can turn for each parameter combined.' on page 44).

After this scenario we will then proceed with two additional scenarios, one in which access is stringent, and one in which data sets are stringent. As said before, this scenario and the two to follow are illustrations of possible value propositions and risk profiles for CBS. For each factor, a (slightly) more lenient or stringent approach might be warranted in practice.

End users

Under a moderate scenario, **access** may be **strongly diversified**, so that **different types of users get different levels of access and permissions**. Organisations can be categorised based on different types of criteria, for instance on their main aim (e.g. research, policy advice, commercial), their years of experience with the remote access environment, or their geographical location (based in the Netherlands, based in the EU, based outside of the EU). Of course, a mixture of these criteria (or others) could also be used. On the basis of the selected criteria, a **clearance level** may be established, and access is granted to parts of the (remote access to) microdata environment that fit with this clearance level. The higher the level of clearance, the more activities users can perform in their own physical environment through remote access. For all users, access is limited initially to a relatively stringent level, and more access can be gained over time, as **trust** between CBS and the user organisation increases. Access can further be **diversified within an organisation**, so that long-term users within the organisation might be able to access more data compared to first time users.

Use

Diversified access could also be applied with respect to the aggregation level of data made available in data sets. For example, end users with a higher clearance level could aggregate at smaller levels, either at an organisational or an individual level. Publication requirements could also be adjusted to the clearance level, or, for instance, to the number of years an end user has worked with CBS microdata without incidents.



Blad 53/65

Data sets

In this scenario, the data sets that CBS makes available are categorized into levels of sensitivity, and the levels of sensitivity are directly linked to the **clearance levels** of the end users. Technically, it would even be feasible to diversify **within a data set**, whereby certain variables, or lower aggregation levels become available after end users are provided with a higher clearance level. In terms of location, depending on the data set, and the clearance level of the user, data can be **accessed remotely**, in a secure room on the premises of the user, or at CBS.

Processes and procedures

As in the stringent scenario, site visits and other checks could be put in place to ensure ongoing acceptable levels of security within the users' organisations. However, as the trust relationship between CBS and the user organisation builds, these could be scaled back in terms of frequency and/or duration, if repeated visits and checks in the past do not flag any urgent matters. In essence, this scenario builds on treating end users' organisations as a **partner** in **safeguarding** the remote access to microdata environment and its use. CBS should therefore provide end users' organisations with the information and support to identify bottlenecks and potential risks, such as overly long duration of user access.

This scenario finds a balance between protecting security and privacy on the one hand, and fostering accessibility and user-friendliness on the other. However, by trying to find the middle ground between these competing values, it does not embrace any one of them fully. Moreover, the moderate scenario would entail a **range of costs**, both **financially** as well as in terms of **manpower**, as the diversification processes cost time, energy and setting up a series of new procedures (e.g. risk assessments of data sets, site visits etc.). It is up to the CBS to decide to what extent they would want to diversify, and whether this should take place on all factors simultaneously, or whether to start, for example, by first defining the 'other' category more precisely, and starting an initial risk assessment of the most often used data sets.



Blad 54/65

Parameter	Choice	
End users	User categories:	An end user categorization will be made with different levels of clearance for each category. Categories can be built on different criteria, e.g. main aim, geographical location and/or years of experience with remote access to microdata.
Use	Use categories:	Depending on the user categorization and their assigned security level clearances, different aggregation levels of data may be made available and different requirements may apply for publications.
Data sets	Data categories:	Depending on the user categorisation and their assigned security level clearances, specific categories of data can be accessed.
Processes	Site visits & surveillance:	Depending on the user, use, and data category, different processes will be organized by CBS. There will be site visits (End users will receive announced and unannounced site visits to check on their use of the microdata environment); a strict policy before entering the remote access system, but also after admission; Random checks of what end users are doing in the remote environment (not just output checks). Support user institutions through information exchange and account management.

6.4 Stringent on end users, lenient on use, data sets and processes

In this scenario, the **dials** for **access granted to end users** are set to **stringent**, so that the **other dials can be set to more lenient standards**. The idea of this scenario is that if the access is guarded very strictly, this opens up the possibility to be more open to the users who gain access to the remote access to microdata facilities. In this scenario the dial for end users would be placed in to the left two columns of Table 10 (see page 44), while the rest of the dials is spread out over the middle column and the two columns on the right in the same table.

End users

Access for end users is set up in the same way as in the stringent scenario: there would be strict rules and regulations as to who can access the micro-data. The ‘other’ category is strictly defined and grey areas are limited, so that every organisation either fully falls within this category, or does not. Moreover, there is a user categorisation that reveals which **security and**



Blad 55/65

privacy clearance end users have. Remote access is only granted to end users with the highest levels of clearance. All other end users only get access via secure rooms in their own organisations or on-site at CBS. Furthermore, only end users directly affiliated with **Dutch organisations** are allowed access to the data, and **commercial** parties do not have access to the data.

Use

Since the access controls are very stringent in the scenario, once end users have been granted access to the system, **other controls can be set to a more lenient setting**. For instance, the **level of aggregation** for exporting the results of the statistical analysis to another location outside the microdata environment could be set to $n > 5$ (very lenient), and publication requirements could be set to moderate, i.e. CBS regularly asks random end users to provide proof of the **publications** containing research for which microdata have been used. It then checks these publications to see whether the output data are correctly presented. If this is not the case, end users receive a one-time warning; if a second check leads to the same findings, the end user's credentials to the remote access system are revoked.

Data sets

The **data set dial** can also be to more lenient settings since access controls are stringent. Only trustworthy users would have access to the remote access environment. This means that no, or **very little diversification** relating to the **sensitivity** of data in each data set is required. A data classification could be made with just a few degrees of sensitivity, or no classification could be made at all.

Processes and procedures

As in the high accessibility scenario, the onboarding and offboarding processes could remain the same as they are now, and CBS would not have to not conduct site visits or monitor online activity, unless there are concrete complaints or other information suggesting that this is required. Again, this is based on the notion that the access procedures, policies and guidelines are of a stringent level, so that trustworthy partners can be allowed more freedom in their use of the microdata.

In this scenario, the CBS applies a (very) **stringent form of gatekeeping**, so that only trusted parties can get access to the microdata services the CBS offers. One of the advantages of this



Blad 56/65

scenario is that end users know that once they receive access, they are not withheld information, or access to subsets of the data available depending on other factors, but have access to all the data they require. This could incentivize end users and their organizations to actively work towards **incorporating security and privacy frameworks** within their organization, due to the obvious payout in terms of data access. In terms of costs this scenario is moderate. It places a burden on CBS to ensure that access controls for end users are of the highest security level, which means (ongoing) investments in technologies for access control and (ongoing) investment in staff evaluating and monitoring the end user base. Having said that, costs should be manageable, since new access procedures would entail a one-time set up of new policies and guidelines, with some extra checks for each new user. Once a user is allowed access, no additional recurring costs are foreseen. One of the downsides of this scenario, is that **if gatekeeping at the access level fails** for a specific organization, these organizations then have **full access to data without any future monitoring**. In terms of security, then, this scenario contains a so-called **single point of failure**, which is a serious weakness.



Blad 57/65

Parameter	Choice	
End users	User categories:	Strict specification of the category ‘other’; no one outside that specification is allowed access. Commercial parties are not granted access. Users must be affiliated with a Dutch organization.
Use	Use categories:	Aggregation levels for exporting findings may be set as low as $n > 5$, and publication requirements may be lenient.
Data sets	Data categories:	No need for data classification per se; if a data classification is made, this could be a relatively simple classification, e.g. with only 3 categories of sensitivity. (Almost) all data could be accessed remotely.
Processes	Site visits & surveillance	CBS will not take action or conduct audits, unless there are concrete complaints or signs of actual abuse.

6.5 Stringent on data sets, lenient on end users, use and processes

The previous scenario showed how setting access for end users to a stringent setting opens up more lenient settings for the other categories. Another way of setting the dials would be to be **stringent on data sets**, so that who has access, what end users may do, and which processes and procedures surround remote access to microdata can be more lenient. In this scenario the dial for data sets would be placed to the left two columns of Table 10 (see page 44), while the rest of the dials is spread out over the middle column and the two columns on the right in the same table.

End users

Due to stringent settings for the data sets, the access can be set to more lenient settings, even going so far as the **high accessibility scenario** above does. In this case, the ‘other’ category would be interpreted broadly, so that a wide range of organisations could get access to the microdata. This includes opportunities for end users to conduct statistical research for commercial projects, but might also allow citizens to access the microdata as well as organisations outside the EU.

Use

Since the classification of data would be very stringent under this scenario, the dial of use could also be set to a more moderate setting. For instance, a **flexible approach** could be taken to the **aggregation level** of data sets made available to end users. For data bases that are labelled ‘internal’ the aggregation level needed to be able to export data to another location



Blad 58/65 than the microdata environment could be set to $n > 5$ (very lenient), whereas for data sets that are labelled 'confidential' a higher aggregation level could be chosen. Similarly, **different requirements** could be made with respect to **publications**, depending on the level of sensitivity of the data used for a study. For instance, when end users use a data set that is labelled 'highly sensitive', CBS could check the content of all publications generated on the basis of that study (very stringent). By contrast, should a data set labelled 'internal' be used, CBS could choose to only ask for a publication plan (very lenient).

Data sets

In this scenario, the dial for data sets is put to the most stringent option. This means that a **risk analysis** of all data sets is required, and that a strict **data classification** needs to be developed. For each data set the level of sensitivity is established. Depending on the sensitivity of the data set, there are different **locations** in which data may be accessed. Highly sensitive and sensitive data can only be accessed at one of the CBS locations. Confidential data can only be accessed in a secure room within the end user's organisation, and only internal data can be accessed remotely, in a manner similar to the current remote access procedures. One option to open up the (highly) sensitive and/or confidential data to a larger audience, so that end users could gain insights from analysing that data, would be to allow end users to only see **metadata** remotely so they can instruct **researchers at CBS** to **run the analysis** for them. This would come at an extra cost, but would facilitate research on even the most sensitive data, without opening up direct access. For all categories, the rule is that users can only access the data while being physically in the Netherlands.

Processes and procedures

The processes and procedures to follow would be made **flexible** to align with the level of sensitivity of the data that the end users would want to work with. For data that would be labelled internal, no additional procedures would need to be set up, while for more sensitive data, processes would be needed to assess the quality of e.g. safe rooms at the end user's organisation for instance. Similarly, under this scenario surveillance and logging would be made flexible: activities involving the most sensitive data at the CBS location would be surveilled and logged minutely, while activities via remote access on internally available data would not be surveilled or logged at all, or only marginally.

This scenario would allow a wide range of users to gain access to some CBS data relatively easily. However, through diversification of the data sets, there will be **differences between**



Blad 59/65

users in terms of access, as not all users will have the option to build safe rooms, or travel to CBS for data analysis. One added benefit is this: while the previous scenario (strict on access by end users, not on other parameters) had a single point of failure that leads to security risks, in this scenario the measures taken in terms of **use** and **procedures** are **diversified** along with the level of sensitivity of the data sets. Thus, gaining access to data sets that are labelled 'highly sensitive' does not enable attackers to exploit the system without this being noticed by the organisation. In terms of **costs**, this scenario would be relatively expensive, as not only an ongoing risk assessment would need to be carried out on the data sets, but also on-site access needs to be set up, preferably at multiple locations, that is safe, yet scalable. Moreover, conducting audits for safe rooms outside CBS is also a costly enterprise.



Blad 60/65

Parameter	Choice	
End users	User categories:	A similar regime applies to all user types. A broad interpretation the ‘other institutions’ is opted. For example, statistical research for commercial purposes would be allowed. The requests of citizens who want to perform statistical research may also be accepted. Moreover, organizations from outside the EU may also be allowed access. Ultimately, CBS could choose to move towards an open access model for their (anonymized) microdata. Remote access is possible and all parties have access to the microdata directly.
Use	Use categories:	Aggregation levels of data sets would depend on the level of sensitivity of the data in each data set. Publication requirements, too, would depend on the sensitivity of the data set(s) used.
Data sets	Data categories:	Strict specification of the categories of data sets: ‘internal’, ‘confidential’, ‘sensitive’ and ‘highly sensitive’. Remote access is allowed only for the category ‘internal’. The category ‘confidential’ may be accessed only via a safe room in an end user’s organisation. The categories ‘sensitive’ and ‘highly sensitive’ may be accessed only at one of the CBS locations. AND users may access data in any category only while on Dutch territory.
Processes	Site visits & surveillance:	CBS will surveil and log all activities relating to (highly) sensitive data; it will not, or only marginally log activities on internally available data. CBS will check the content of all publications that use data from data sets that are (highly) sensitive; for publications on internally available data it will only require a publication plan.



7. Conclusions and recommendations

In the period between March and October 2020, an interdisciplinary research group consisting of six researchers collectively endeavored to answer two main questions:

- What potential cybersecurity and privacy risks may be involved in CBS offering the remote access to microdata service in its current form?
- What measures could CBS take, or what (policy) choices could it make to serve the public interest by providing access to the data it gathers and facilitate researchers, while protecting the private and collective interest of citizens and companies by warranting security and privacy?

In order to answer these questions, the committee has conducted a literature review, a legal, ethical and technical analysis, and several interviews with multiple stakeholders. This research has been described in work packages 1, 2 and 3. Based on the first three work packages, the committee developed four specific products:

1. An **evaluative tool** to help CBS assess different, potentially competing values with respect to decisions regarding access to microdata in relation to privacy/security. This has been developed in work package 4.
2. A **risk analysis** with the main risks of the current implementation of the remote access system. This has also been developed in work package 4.
3. A **set of dials** to enable a tailored-made approach to mitigating risks. This has been developed in work package 5.
4. A set of **scenarios** for the future of remote access services for microdata. This has also been developed in work package 5.

7.1 Values

With these products the Committee believes CBS has the necessary tools at its disposal to come to a grounded decision on how it wants to (re)organize its remote access service. From the start, this Committee has argued that it would not be sufficient to use a classical risk management approach to chart and address the risks surrounding this environment. CBS, and in particular its remote access to microdata service, are driven by distinct, yet often **implicit values**. These values needed to be **explicated** first, in order for CBS to be able to set out a clear direction for its remote access services. This research established that the remote access service is essentially driven by the values of enabling **research, accessibility, and user-**



Blad 62/65 **friendliness**. Simultaneously, there are also values underpinning this service, namely: **trust**, **privacy** and **security**. These underpinning values limit the range of the driving values. Finally, as a professional organization, CBS also is steered by process-based values, such as **responsibility**, **compliance** and the aim to **lead by example**.

This Committee advises CBS to first develop a clear **value profile**, based on the framework developed in this research. Based on the interviews we held with stakeholders, we would recommend taking a **co-creating approach** to drafting such a profile. There is very valuable knowledge to be found with CBS employees involved throughout the whole remote access life cycle, as well as with end users making use of the service. This value-based mission statement will provide guidance to CBS in the decision-making process on (re)organizing the remote access services. Depending on the value-based mission statement, a clear picture of the risks and risk control strategies emerges.

7.2 Risks

Since 2006, the year the remote access service was introduced, numerous technological, legal, and societal developments have occurred which can negatively impact the values listed above. The **increasing number of actors** that engage in statistical research, sometimes with the desire to **reuse data for commercial purposes**, the possibility to **combine data sets and re-identify individuals**, and the increasingly **sophisticated programs that can be used to capture images** are just a few of the risks this committee has identified.

The committee found that **CBS has already invested extensively in security measures** – both technically and procedurally– to mitigate privacy risks and prevent abuse of data. However, these measures are predominantly focused on the **enrollment** phase for (potential) end users and are less concerned with the use phase. Moreover, current measures are predominantly directed to mitigate unintended privacy violations and security failures, rather than concentrate on dealing with **intended abuse of data**. This is understandable as in the beginning of the remote access service, actors who received access were perceived as trustworthy actors. CBS employees knew these individual researchers, their institutions, and their work. However, as statistical research is no longer an activity solely executed by a few research institutes and increasingly also companies as well as foreign actors want to make use of the remote access services, these trust-based measures may no longer suffice. Moreover,



Blad 63/65 the growing diversity of actors as well as of types of statistical research demand a more **tailor-made set of measures to mitigate risks** than the current one-size-fits all procedures.

The Committee therefore advises CBS to take a more **diversified approach to mitigating risks**. For all four parameters the Committee has identified (end users, use, data sets, and processes/procedures), it has developed dials. Based on the to be established value-profile, CBS can decide to be more lenient or more restrictive in providing access to its microdata. For instance, CBS could distinguish between different types of access to data, for instance by distinguishing between remote access, access via a safe room at the end users' workplace, or access via safe rooms at CBS locations only. By taking such a diversified approach, it becomes possible to **safeguard key values** such as accessibility and doing research, while also taking into account **privacy and security requirements**. The scenarios that have been developed in this document illustrate how CBS could approach such a diversified approach.

7.3 Recommendations

The committee advises CBS to conduct the following steps each year:

1. **Reassess its value-based mission statement.** Are these values still reflecting CBS's approach and vision for the future? Do political or societal developments necessitate amendments?
2. **Reassess its diversified approach.** How is the approach evaluated in practice by the organisations having access to the microdata? Have there been any signs of risks or needs to curb or expand access rights?
3. **Reassess its technical standards.** Have the technical/security standards shown any sign of weakness, where are patchworks needed and where are new technical applications needed altogether? What current and future developments in terms of technical security does CBS need to invest in to remain a trusted partner?
4. **Reassess its legal standards.** Have new laws, bylaws, case law or guidelines been published that not only need implementation on concrete points, but also in the general approach taken by CBS? For example, do new developments in privacy law or open access/re-use of PSI legislation necessity a new value-based approach or adding new modifications to the diversified approach?
5. **Reassess its organisational and procedural standards.** Are the procedures for vetting employees and organisations having access to microdata functioning well? Are the ways for creating transparency on the use of microdata functioning well?



Blad 64/65

The committee advises CBS to include the following parties in that process:

1. CBS's microdata team
2. CBS's privacy and IT officer, supported by a biannual external evaluation conducted by an independent and interdisciplinary team of researchers
3. CBS's User Council, supported by an annual survey among current and active end users of CBS microdata



References

- Aven, Terje. 2014. "What is safety science?" *Safety Science* 67 (0925): 15-20.
<https://doi.org/10.1016/j.ssci.2013.07.026>.
- Berg, Heinz-Peter. 2010. "Risk management: procedures, methods and experiences."
Reliability: Theory & Application 1 (17): 79-95.
- De Bruijne, Mark, and Michel Van Eeten. 2007. "Systems that should have failed: Critical infrastructure protection in an institutionally fragmented environment." *Journal of Contingencies and Crisis Management* 15 (1): 18-29.
- Zuiderwijk, A. and M. Janssen. "The negative effects of open government data-investigating the dark side of open data" in *Proceedings of the 15th Annual International Conference on Digital Government Research*. 2014, p147-152.