# Extraction description and estimated costs

This form is intended for use by applicants, data controllers and Findata.

* The applicant can ask the controllers for advice on the **feasibility** and **cost** of the planned extraction, before submitting a data permit application or data request to Findata. Findata confirms the estimated collection costs from the data controllers during application processing.
* The controllers can give applicants **advice** in referred to in the Act on the Secondary Use of Health and Social Data on the **data content of registers** and their **suitability** for a purpose referred to.
* Pursuant to the Act, Findata requests a free estimate from the controllers on the implementation of the extraction and the maximum costs.

The information on the feasibility of extraction **reduces** the number of **additional extractions and corrections** that are required afterwards. Preliminary information on the costs related to the extraction facilitates the planning of project funding.

## **Instructions for data permit applicants**

**Fill in sections 1–4 of this form**. They are indicated with a blue background.

Before sending an application to Findata:

1. Fill in one form per controller.
2. Send the filled-in form to the controllers whose registers the extraction concerns.
3. If necessary, the controller provides advice on the content of the extraction and the selected variables and returns the form to the applicant with additional information.
4. Attach the filled-in form as a Word file to the data permit application you submit to Findata.

*Tip: if you send the form to several controllers, fill in sections 1 and 2 of the form first. Then make a copy of the form for each controller and fill in sections 3 and 4.*

## **Instructions for controllers**

**Fill in sections 5–7** and, if necessary, the diary number to section 1. They are indicated with an orange background.

When you receive a form filled in by an applicant or Findata:

1. Comment on whether it is possible to carry out the extraction described in the form. If not, advise the applicant to modify the extraction so that it is feasible.
2. Estimate the number of files that it will take to deliver the data resource and the size of the files.
3. Estimate the maximum cost of the extraction described in the form.
4. Send the filled-in form back as a Word file.
   * If you have received the form from the applicant, send the form back to the applicant.
   * If you have received the form from Findata, send the form back to Findata.

## **1 Short description of the project** *[applicant fills in]*

The purpose of this section is to provide the controller with information on the project and extraction as a whole.

The content of the table below **must always be identical** regardless of which controller the form is sent to.

|  |  |
| --- | --- |
| Required information | Fill in this column |
| Project name |  |
| Short description of the project | *Controllers need background information on the project to assess the feasibility of the extractions. Describe with a few sentences (max. about 70 words / 500 characters).* |
| Applicant / applicant organisation | *Permit applicant refers to the organisation or person named in the application who the data permit will be issued to or to whom the statistical data with the data request decision will be delivered.* |
| Applicant contact person | *Enter the contact person’s name, email address and phone number. Contact person refers to the person who responds to enquiries concerning the application and extraction.* |
| Invoice payer | *Please fill in name, VAT-ID, country*  *Whether or not VAT is added to the extraction costs depends on the final payer of the invoice. Regarding the costs of the controllers, Findata treats the invoice as a pass-through item in its accounting. Thus, the controller determines the VAT liability based on the final payer of the invoice.* |
| Intended use referred to in the Act on the Secondary Use of Health and Social Data | *Please choose from the following.* ***Data permits:*** *scientific research; statistics; education; planning and reporting duties of an authority* ***Data requests****: in addition to the above, development and innovation activities; knowledge management; guidance and supervision of a social and healthcare authority* |
| Recurrence of extraction | *Is the data extracted only once or for example annually until a certain year?* |
| Previous permits related to this extraction | *If there are previous/old permits related to the project or extractions, please list the requested information about the permits below. This way, the data controllers know what the extraction is related to.*   * Permit issuer, date, diary number * Permit issuer, date, diary number |
| The diary number of the controller |  |

## **2 Order of extraction** *[applicant fills in]*

In the table below, describe the stages of compiling the data if the process involves several extractions (e.g. if the target group, controls and data resource are extracted from different controllers).

Only fill in register-specific extractions into their separate tables in section 4.

**The order of the extraction must always be identical** regardless of which controller the form is sent to.

|  |  |
| --- | --- |
| Extraction step no. | Step description |
| 1 | The purpose of this extraction is to identify the persons in the target group. Describe which register is used for identifying the target group and the estimated size or magnitude of the target group.  If the target group does not need to be extracted separately because you already have it, describe here where it came from.  *E.g. Findata sends a data request concerning the target group to Finnish Institute for Health and Welfare (THL), and THL delivers the target group to Findata.* |
| 2 | Possible control group and/or relatives are extracted at this stage.  *E.g. Findata sends the personal identity codes of the target group to the Digital and Population Data Services Agency (DVV) for control group extraction. DVV delivers the control group to Findata.* |
| 3 | The final data is extracted at this stage. Describe which registers are used to extract data for the target group.  *E.g. Findata sends the target group and the personal identity codes of the control group to THL and Kela for data extraction. Kela and THL deliver the data resource to Findata.* |
| 4 |  |

## **3 Controller this form concerns** *[applicant fills in]*

|  |  |
| --- | --- |
| Required information | Fill in here |
| Controller | *E.g. Finnish Institute for Health and Welfare (THL)* |
| Registers and data resources / datasets | *Utilise the Data Catalogue. E.g.*   * *Care Register for Health Care: Register of Primary Healthcare Visits 2011 and Care Register for Social Welfare* * *National Infectious Diseases Register* |

## **4 Variables to be extracted** *[applicant fills in]*

Fill in the variables to be extracted in the table below as precisely as possible.

If you need variables from several different registers or datasets from the same controller, copy and fill in a separate table for each.

|  |  |
| --- | --- |
| **Extraction step (see order of extraction)** | *E.g. Step 1, target group extraction* |
| **Registers and data resource (dataset)** | *E.g. Care Register for Health Care: Social welfare care notification register* |
| **Limitations** | **Target group:**  The purpose of this extraction is to identify persons in the target group.  *Target group refers to the group of persons that is to be identified by this extraction and for which information is to be extracted either from this register or from other registers.*  **OR**  The purpose of this extraction is to extract data resources for an already identified target group.  *Target group refers to the persons for whom this extraction is used to extract data.*  *If necessary, note here if the data according to this table is extracted not only for the subjects but also e.g. for control subjects and/or relatives (enter them as separate target groups). If necessary, note here if there are multiple target groups and different limitations are used for extractions for different target groups.* |
| **Size of the target group(s):** *Enter here how large the target group is estimated to be. Overestimate the size rather than underestimate. If data is also extracted for relatives or a control group, include them in this amount as well.* |
| **The size of the target group has been estimated by:** *If the size of the target group has been estimated, indicate here by which body* |
| **Date range from which data is extracted:**  *Enter the time range for this extraction.* |
| **Date range delimiter variable:**  *Enter here the variable according to which the date range is limited.* |
| **Extraction limitations:**  *E.g. only persons over 18 years of age* |
| **Variables to be extracted** | *Describe the data resources to be extracted at the variable level.* |
| **Unstructured variables** | List here again the variables that you think will be extracted in an unstructured format.   * Explain briefly the need for the unstructured variables * Determine whether the controller may directly create structural data from the unstructured variables, if possible   *E.g. “Patient document texts or text excerpts extracted with word search. It is not possible to obtain this information in any other way. The texts and excerpts must be delivered as is.”* |
| **Checkpoint (proposed by the applicant)** | I propose that, in order to ensure the accuracy of the extraction, a checkpoint is kept for this extraction during the data extractions.  *Tick the box if you want to propose a checkpoint to the controller to verify the accuracy of the extraction.*  *In principle, the checkpoint is intended to ensure the accuracy of the target group in sequental/intertwined extractions, but it can also be used in conjunction with other complex non-sequental/intertwined picks.*  *The checkpoint can be implemented in different ways, e.g. by providing the applicant with an extended data description (size of the target group and codes used) or by a Teams meeting between the applicant and the controllers or a discussion by email.*  *Findata does not participate in the practical arrangements of the checkpoint* *which will be agreed between the applicant and the controller. Findata does not determine whether a checkpoint has been carried out when processing the target group. The applicant or the controller will inform Findata if any issues arise as a result of the checkpoint, which require Findata to act.* |
| **Other remarks** |  |

## **5** **Controller's comments on the proposed extraction** *[controller fills in]*

|  |
| --- |
| ***Controller****:*  *Enter in this box any possible comments on the extraction.*  **Checkpoint (proposed by controller):**  I propose that, in order to ensure the accuracy of the extraction, a checkpoint with the applicant is kept for this extraction during the data extractions.  *Tick the box if you want to propose a checkpoint to the applicant to verify the accuracy of the extraction.*  *Enter here comments about the need for a checkpoint if the applicant has ticked the proposal for a checkpoint in the table in section 4.*  *In principle, the checkpoint is intended to ensure the accuracy of the target group in sequental/intertwined extractions, but it can also be used in conjunction with other complex non-sequental/intertwined picks.*  *The checkpoint can be implemented in different ways, e.g. by providing the applicant with an extended data description (size of the target group and codes used) by a Teams meeting between the applicant and the controllers or a discussion by email. Findata does not participate in the practical arrangements of the checkpoint* *which will be agreed between the applicant and the controller. Findata does not determine whether a checkpoint has been carried out when processing the target group. The applicant or the controller will inform Findata if any issues arise as a result of the checkpoint, which require Findata to act.* |

## **6 Extraction estimates by the controller** *[controller fills in]*

In the table below, fill in the estimated number of files that will be retrieved. Quantity estimates help Findata to estimate the costs of data processing for the data permit applicant.

|  |  |
| --- | --- |
| Required information | Fill in this column |
| Estimate of the number of persons included in the requested data resource | Approx: |
| Estimate of the number of lines and files in the data resource that will be delivered in a structural format | Files with less than 5 million lines: XX pcs  Files with 5 to 20 million lines: XX pcs  Files with more than 20 million lines: XX pcs |
| Variables containing identification information | *List here*  *- all variables that may contain identifying information, and*  *- what kind of identification information is involved.* |
| Does the requested data resource contain unstructured data? | Yes / No |
| Is it possible for the controller to refine the open text into a structured format instead of providing the open text? | Yes: *which variables are concerned?*  *E.g.: smoking data can be provided in the form  smokes/quit smoking/never smoked*  No |
| Description of unstructured data, as detailed as possible | 1. Does the unstructured data contain individuals’ identification data (e.g. names, addresses, telephone numbers): 2. The file format of the unstructured data: 3. Estimate of how many files include unstructured data: 4. Estimate of the number of lines of unstructured text data per file: 5. Description of unstructured data (e.g. X-rays, text extracts with 100 characters or three sentences around a keyword): |
| In which format the register data to be extracted is delivered? | *The primary file format is CSV.*  *If some data resources to be extracted cannot be delivered as CSV files – for example imaging materials – indicate the file formats used for delivering the data resources.* |

## **7 Maximum cost estimate for the extraction** *[controller fills in]*

In the cost estimate table below, enter

* an estimate of the maximum total cost of the extraction described above
* a breakdown of the cost estimate, i.e. information on what the estimate of the total costs consists of
* VAT information

You can edit the information content in the table on a case-specific basis to correspond to the breakdown of costs for your extraction. The table should indicate prices with and without VAT.

If compiling the requested data requires multiple extractions, copy the table below and fill in separate tables with the maximum cost estimates for each extraction.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Required information | Task | Task unit | Number of units (pcs) | Unit price (EUR) | Total price for the task |
| Cost estimate breakdown *Examples in italics* | *E.g. extraction from register X* | *Work hour* | *35* | *115* | *4025* |
| *IT cost* | *Fixed cost* | *1* | *500* | *500* |
|  |  |  |  |  |
| Total price of the tasks exclusive of VAT | EUR XX  *E.g. EUR 4.525,00*  *Note! The cost estimate requested by the applicant may be given as an approximation. During the application process, Findata will verify the extraction costs with the controllers.* | | | | | |
| Value added tax percentage (if VAT is invoiced) | XX%  Note! *Findata treats the invoice as a pass-through item in its accounting. Thus, the controller determines the VAT liability based on the final payer of the invoice. The invoice payer is indicated in section 1.*  *For invoices addressed to Findata, add a reference to the data recipient or other final invoice payer indicated by the data recipient. This will allow the data recipient or other final invoice payer to deduct the VAT on the invoice in their own taxation.*  *Use the following text and complete with the payer's details: The final payer of the invoice is \*payer’s name (VAT-ID: )\*, according to which the VAT treatment of the invoice is determined. Findata treats the invoice as a pass-through invoicing and is therefore not liable for VAT on the invoice. (Guidance from Finnish Tax Administration)* | | | | | |
| Total estimate of maximum costs inclusive of VAT | EUR XX | | | | | |
| Contact details of the issuer of the cost estimate | Name and email | | | | | |
| Date and validity of the cost estimate | dd Month yyyy | | | | | |