Project N°: 262608

ACRONYM: Data without Boundaries

DELIVERABLE D5.2
(Report and Databank Documenting OS Microdata)

WORK PACKAGE 5
(Servicing European Researchers in the use of OS Microdata)

<table>
<thead>
<tr>
<th>REPORTING PERIOD:</th>
<th>From: Month 19</th>
<th>To: Month 36</th>
</tr>
</thead>
<tbody>
<tr>
<td>PROJECT START DATE:</td>
<td>1st May 2011</td>
<td>DURATION: 48 Months</td>
</tr>
<tr>
<td>DATE OF ISSUE OF DELIVERABLE:</td>
<td>08th October 2013</td>
<td></td>
</tr>
<tr>
<td>DOCUMENT PREPARED BY:</td>
<td>Partners 1 &amp; 5</td>
<td>CNRS-RQ &amp; GESIS</td>
</tr>
</tbody>
</table>

Combination of CP & CSA project funded by the European Community
Under the programme “FP7 - SP4 Capacities”
Priority 1.1.3: European Social Science Data Archives and remote access to Official Statistics

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 262608 (DwB - Data without Boundaries)
This document was prepared by Raphaëlle FLEUREUX, Cyril JAYET and Roxane SILBERMAN (CNRS-RQ) as well as Alexander MACK and Christof WOLF (GESIS).

The authors gratefully acknowledge contributions and feedback from all WP5 team partners. In particular: Sebastian KOCAR and Janez STEBE (UL-ADP); Atle ALVHEIM (NSD); Adrian DUS A and Irina CRISTESCU (RODA); Andreas PERRET (FORS); Steve BOND and Tim PATEMAN (ONS); Albert ESTEVE, Toni LOPEZ and Anna TURU (CED).
TABLE OF CONTENTS

Introduction: Addressing the obstacles for researchers through a single point of documentation ....... 3

1. Conducting the task ................................................................................................................... 5
   The schema and the fields........................................................................................................5
   The use of controlled vocabulary and official classification............................................. 8
   Development of a software...................................................................................................... 8
   WT5.2 in connection with other work packages................................................................. 9

2. The scope of the database: state of play, issues raised for each country ......................... 10

3. Enhancing IECM database.................................................................................................... 12

Conclusion: Accessing the metadata, maintaining the database and future of CIMES .......... 15

Appendix ...................................................................................................................................... 17
   Specific reports by country on the metadata collection process ........................................ 17
   WP5 Task 2 - Metadata Schema.......................................................................................... 24
   CIMES - Database Model....................................................................................................... 37
   CIMES - User Guide.............................................................................................................. 38
Introduction: Addressing the Obstacles for Researchers through a Single Point of Documentation

The objective of Work Task 5.2 (WT5.2) is to facilitate the use of the increasing number of official microdata. As raised by Grenet (2012), researchers are facing serious obstacles when trying to get information on restricted data and accreditation procedures. Conducting comparative research requires an in-depth search through the websites of each national statistical agency to discover which data is available while understanding how each producer document their own data. This is very often time-consuming and burdensome.

Making these data easier to find, more comprehensible and more usable, requires the improvement of the descriptive metadata on two key elements. Firstly, these metadata, which are currently scattered all over European NSIs’ websites or not even findable, should be centralised, so that researchers can have a single point of access and an overview of available data in Europe. Secondly, these metadata, which currently consist of unstructured and semi-structured documents, should be transformed into standardized documentation. Standardized documents are much more reader-friendly and have the advantage of being “machine readable”; thus facilitating the access, evaluation and indexation of information by search engines. Standardized documents also facilitate interoperability: they can easily be converted into other formats or used by other applications.

In social and economic sciences, the international standard which is now used by data archivists to produce structured metadata of microdata was designed under the name DDI (Data Documentation Infrastructure).

---

1 The name of the report was changed. The name given in the Description of Work for the D5.2 (p. 27) was the same as for the D5.4 and did not describe correctly the objectives of the task.

Documentation Initiative). It “accompanies and enables data conceptualization, collection, processing, distribution, discovery, analysis, repurposing, and archiving” 3. However, a survey conducted by the WP8 of DwB evidenced that most NSIs still disseminate their documentation via pdf or word documents or their macrodata with the SDMX standard, but most of the time they do not store the documentation of microdata in a database complying with international standards.

WT5.2 aims at contributing to overcome these obstacles by two means. Firstly, WP5 has built a database describing available official microdata in Europe. It has gathered the pieces of information scattered throughout Europe, either in NSIs or data archives, and has stored them in a structured database compatible with the DDI standard. This new database provides researchers with a comprehensive overview of official data available for research purposes in Europe as well as the procedures for requesting access to these data. It is a first step to build a single point of access to metadata in Europe. Secondly, it has enhanced the Integrated European Census Microdata database4 which integrates, harmonises, and disseminates European microdata and metadata from 24 countries and 55 censuses.

These two databases should eventually facilitate international comparisons of European countries in economics and social sciences. They should thus foster the development of a Knowledge-Based economy while promoting research-driven public policy evaluations at the European level.

3 http://www.ddialliance.org/what
4 www.iecm-project.org
1. Conducting the task

The schema and the fields

WP5 agreed on a documentation schema compatible with DDI standard and designed this schema in close collaboration with both WP8 and WP12. It defined how to structure the data documentation by distinguishing the documentation of series, studies, and datasets. Each object is documented by a list of fields and a controlled vocabulary to be applied in these fields. The formal and detailed description of the corresponding object model is provided in the WP8 report (D8.1 OS Object Model). We will present here a simplified description, the detailed one is provided in the appendix (see Appendix 2 - WP5, Task 2: Metadata Schema, pp. 24 - 36).

A series is a set of studies and represents a longitudinal and repeated cross-sectional data collection process. It also illustrates the continuous data collection process carried out by the NSIs, where each wave of collection is a study. The dataset refers to the set of files which are disseminated.

- Series: a group of studies, for example Labour Force Survey (LFS)
- Study: the actual collection of data, for example LFS 2007
- Dataset: a set of files disseminated, for example LFS Scientific Use File (SUF) 2007
In many cases, the relationship between these three objects can be conceived as a simple three-level hierarchy (as in graph 1 above): a dataset belongs to a study which belongs to a series. A study can have one or more associated datasets. However, this hierarchical representation does not cover all the possible relations between these three objects: some studies are not attached to series, and some datasets can be attached to different studies or correspond to one part of a study only.

This schema is based on the DDI Codebook and DDI Lifecycle; the use of DDI Lifecycle is limited, as documenting the whole life-cycle of these data is considerably beyond the scope of this WP. It is however consistent with DDI best practices to use some parts of DDI3 only.

The fields are listed in table 1.
### Table 1 - Fields

<table>
<thead>
<tr>
<th>Series level</th>
<th>Study level</th>
<th>Dataset level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>Title</td>
<td>Dataset Identification Number</td>
</tr>
<tr>
<td>Subtitle</td>
<td>Subtitle</td>
<td>Name</td>
</tr>
<tr>
<td>Alternative Title</td>
<td>Alternative Title</td>
<td>Access conditions</td>
</tr>
<tr>
<td>Abstract</td>
<td>Producer</td>
<td>Other Access Conditions</td>
</tr>
<tr>
<td>Keywords</td>
<td>Abstract</td>
<td>Description of Accreditation Procedures</td>
</tr>
<tr>
<td>Geographic Coverage</td>
<td>Keywords</td>
<td>Contact</td>
</tr>
<tr>
<td>Time Period Covered</td>
<td>Geographic Coverage</td>
<td>Notes</td>
</tr>
<tr>
<td>Time Method</td>
<td>Universe</td>
<td></td>
</tr>
<tr>
<td>Notes</td>
<td>Sampling Procedure</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Date of Collection</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Time Period Covered</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Kind of Data</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Documents</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sources</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Notes</td>
<td></td>
</tr>
</tbody>
</table>
**The use of controlled vocabulary and official classification**

Each NSI or Archive has usually their own classification and set of keywords for describing their data and there is no standard classification that would be shared across Europe. We decided to use the CESSDA portal classification as keywords; it provides a classification that has already been used to describe data across Europe and allows keeping the consistency with this portal, to which CIMES may be integrated in the future. The controlled vocabulary was chosen using different international standards (DDI, Nesstar). The list of keywords and the controlled vocabulary are provided in the metadata schema in the appendix (see *Appendix 2 - WP5, Task 2: Metadata Schema*, pp. 24 - 36).

**Development of a software**

WP5 tested different software solutions, which were very scarce at that moment (Nesstar, Colectica, etc.). Yet, they were not considered adequate for the purpose of the planned documentation. In particular, Nesstar does not allow for the documentation of hierarchically ordered element, whereas Colectica was insufficiently developed at that time and too expensive for the limited resources of this work package. It was therefore decided that CNRS-RQ would develop a specific tool: CIMES (Centralising and Integrating Metadata from European Statistics). It is a web-based application which allows different partners to produce metadata simultaneously and to store it in the same database (a MySQL relational database). The model used for the CIMES database is provided as an appendix (see *Appendix 3 - CIMES Database Model*, p. 37).

CIMES is likely to take the form of an open-source freeware. However, it is not fully decided when CNRS-RQ will make the source code available, as it requires - among other development tasks - some solid documentation before being publicised.
WT5.2 in connection with other work packages

The WP has contacts with many other work packages of DwB. The CIMES database can be considered to be a first response to the various problems that were discussed at the EDAF meeting organized by WP6 and which emphasized the difficulties for researchers to access data when they are making international comparisons.

As part of its WP3 (WT3.1), DwB has produced a database on the accreditation procedures and data access conditions at the national level. In order to avoid silos, there are plans to integrate both databases, at least partly, within the European Service Centre for Official statistics (ESCS-OS, see Conclusion, p. 15). For the time being, CIMES includes a field which will allow directing the users toward the WP3 database once it is made available online.

The WP8 and WP12 were fully involved in designing the metadata model that was used for CIMES. The objective was to ensure that the resource discovery portal\(^5\), to be implemented and aligned with the CESSDA portal functionality by WP12, would allow researchers to search and browse the documentation on official microdata created by WP5 or on the IECM database. Importantly enough, the metadata schema used for the documentation of national surveys under WT5.2 has been included in the portal metadata model (deliverable D8.2) along with a mapping to SDMX & DDI3.0.

\(^5\) Its specifications and implementation model were designed by WP8.

Due to budget and time constraints, documenting all national surveys available in Europe would not have been feasible. It was decided to document surveys on topics which are most commonly studied in comparative research projects, thus facilitating international comparisons across European countries through a harmonized database. The following topics were selected: Living conditions, Employment, Education, Social Mobility, Structure of Earnings, Working Conditions, Innovation.

The Scope of the database was EU28 + Switzerland + Norway + Iceland. Specific countries were allocated to each of the partners. As the documentation was often not available in English, partners had to translate the documentation by themselves. In terms of amount of work, the situation was very different from one country to another and lead to a documentation that can be different in size (number of survey documented) as well as in coverage (proportion of surveys documented). For four countries (Bulgaria Hungary, and Malta), no documentation of surveys could be obtained. For Cyprus, there is no readily available microdata.

The CIMES database contains over one thousand studies. The apportionment and final number of documented surveys is provided in the table 2 below and the appendix provides specific reports for each partner on the metadata collection process (see Appendix 1, pp. 17 - 23)
Table 2 - Metadata Collection Process - Sharing out of Documentation across Countries

<table>
<thead>
<tr>
<th>Country</th>
<th>Number of series</th>
<th>Number of studies</th>
<th>Number of datasets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Austria</td>
<td>5</td>
<td>29</td>
<td>51</td>
</tr>
<tr>
<td>Belgium</td>
<td>7</td>
<td>42</td>
<td>42</td>
</tr>
<tr>
<td>Bulgaria</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Croatia</td>
<td>4</td>
<td>37</td>
<td>91</td>
</tr>
<tr>
<td>Cyprus</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Czech Republic</td>
<td>5</td>
<td>35</td>
<td>125</td>
</tr>
<tr>
<td>Denmark</td>
<td>20</td>
<td>18</td>
<td>16</td>
</tr>
<tr>
<td>Estonia</td>
<td>3</td>
<td>34</td>
<td>40</td>
</tr>
<tr>
<td>Finland</td>
<td>22</td>
<td>17</td>
<td>11</td>
</tr>
<tr>
<td>France</td>
<td>18</td>
<td>134</td>
<td>156</td>
</tr>
<tr>
<td>Germany</td>
<td>9</td>
<td>87</td>
<td>131</td>
</tr>
<tr>
<td>Greece</td>
<td>4</td>
<td>17</td>
<td>30</td>
</tr>
<tr>
<td>Hungary</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Iceland</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Ireland</td>
<td>6</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>Italy</td>
<td>10</td>
<td>49</td>
<td>49</td>
</tr>
<tr>
<td>Latvia</td>
<td>6</td>
<td>29</td>
<td>27</td>
</tr>
<tr>
<td>Lithuania</td>
<td>4</td>
<td>10</td>
<td>6</td>
</tr>
<tr>
<td>Luxembourg</td>
<td>1</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>Malta</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Netherlands</td>
<td>13</td>
<td>171</td>
<td>0</td>
</tr>
<tr>
<td>Norway</td>
<td>8</td>
<td>137</td>
<td>231</td>
</tr>
<tr>
<td>Poland</td>
<td>4</td>
<td>14</td>
<td>13</td>
</tr>
<tr>
<td>Portugal</td>
<td>13</td>
<td>82</td>
<td>85</td>
</tr>
<tr>
<td>Romania</td>
<td>15</td>
<td>145</td>
<td>143</td>
</tr>
<tr>
<td>Slovakia</td>
<td>5</td>
<td>26</td>
<td>27</td>
</tr>
<tr>
<td>Slovenia</td>
<td>19</td>
<td>35</td>
<td>67</td>
</tr>
<tr>
<td>Spain</td>
<td>21</td>
<td>227</td>
<td>290</td>
</tr>
<tr>
<td>Sweden</td>
<td>2</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>Switzerland</td>
<td>15</td>
<td>151</td>
<td>160</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>8</td>
<td>24</td>
<td>24</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>248</strong></td>
<td><strong>1570</strong></td>
<td><strong>1821</strong></td>
</tr>
</tbody>
</table>

The main problem encountered was the documentation of register data, which are common in northern European countries. Most countries do not provide precise documentation on how they collected data that constitute registers. Such data require a different documentation which was not possible to conduct with the resource of the WP. We will go back over this issue when presenting the future of CIMES (see Conclusion, p. 15).
3. Enhancing IECM Database

Building the CIMES system was the first route to facilitate the use of microdata from European official statistics. The second route was to enhance the Integrated European Census Microdata database (www.iecm-project.org), which provides a centralised access to census microdata samples from 16 countries and corresponding metadata. Table 3 below shows the availability of census microdata samples across 24 European countries. The IECM website offers access to harmonized microdata from 55 censuses (in bold and underlined character in table 3).

Since the start of the Data without Boundaries project in May 2011, 14 new European samples have been added to the IECM database: France (06), Germany (70, 87), Ireland (71, 79, 81, 86, 91, 96, 02, 06) and Turkey (85, 90, 00). Before dissemination, each sample goes through a process of data formatting, harmonization and data documentation. Table 4 shows the distribution of harmonized variables by topic, which are currently available through the IECM website. It also provides information about the type of variables that have been harmonized during the last 2 years for the 14 new samples. In total, 278 variables at the household level and 440 at the person level were harmonized. The IECM website provides extended documentation on all of these variables, including universe, codes, comparability descriptions, among other aspects.

CED is currently working on the harmonization of 19 samples that have been entrusted to the project but not yet disseminated to the research community (bold in table 3 below). Regarding the 2010 census round, an agreement with 9 countries has been reached (one asterisk in table 3 below). For most of these countries, CED expects that the data will be entrusted by mid-2014. To date, Ireland has been the only country to entrust the 2011 census microdata sample to the project (two asterisks in table 3 below).
<table>
<thead>
<tr>
<th>Country</th>
<th>60s</th>
<th>70s</th>
<th>80s</th>
<th>90s</th>
<th>00s</th>
<th>10s</th>
</tr>
</thead>
<tbody>
<tr>
<td>Belarus</td>
<td></td>
<td></td>
<td></td>
<td>1989</td>
<td>1999</td>
<td>2009</td>
</tr>
<tr>
<td>Croatia</td>
<td></td>
<td></td>
<td>1991</td>
<td>2001</td>
<td>2011</td>
<td></td>
</tr>
<tr>
<td>France</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Germany</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Latvia</td>
<td></td>
<td></td>
<td>1989</td>
<td>2000</td>
<td>2011</td>
<td></td>
</tr>
<tr>
<td>Moldova</td>
<td></td>
<td></td>
<td>1989</td>
<td>2004</td>
<td>2011</td>
<td></td>
</tr>
<tr>
<td>Slovenia</td>
<td></td>
<td></td>
<td></td>
<td>2002</td>
<td></td>
<td>2011</td>
</tr>
<tr>
<td>Ukraine</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2001</td>
<td>2013</td>
</tr>
</tbody>
</table>

**Bold** Microdata entrusted; **Bold** Disseminated; * Promised when available; ** 2010 census round sample entrusted
The IECM website is based on the IPUMS data extraction system, which allows checking the documentation in an interactive way while requesting the data. Nevertheless, users can download complete sets of metadata based on DDI standards. This documentation includes information at the series, study and dataset levels, including documentation for each variable.

---

| Table 4 - IECM harmonized variables by topic |

<table>
<thead>
<tr>
<th></th>
<th>Integrated variables</th>
<th>Unharmonized variables</th>
<th>France 06</th>
<th>Germany 70-87</th>
<th>Ireland 71-79-81-86-91-96-02-06</th>
<th>Turkey 85-90-00</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Household</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Technical</td>
<td>8</td>
<td>226</td>
<td>5</td>
<td>10</td>
<td>32</td>
<td>11</td>
</tr>
<tr>
<td>Group Quarters</td>
<td>2</td>
<td>87</td>
<td>2</td>
<td>3</td>
<td>17</td>
<td>3</td>
</tr>
<tr>
<td>Geography</td>
<td>5</td>
<td>106</td>
<td>1</td>
<td>4</td>
<td>14</td>
<td>9</td>
</tr>
<tr>
<td>Economic</td>
<td>1</td>
<td>109</td>
<td>-</td>
<td>8</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>Utilities</td>
<td>7</td>
<td>215</td>
<td>4</td>
<td>9</td>
<td>14</td>
<td>1</td>
</tr>
<tr>
<td>Appliances, mechanicals, other amenities</td>
<td>5</td>
<td>74</td>
<td>5</td>
<td>1</td>
<td>9</td>
<td>-</td>
</tr>
<tr>
<td>Dwelling characteristics</td>
<td>13</td>
<td>341</td>
<td>11</td>
<td>14</td>
<td>24</td>
<td>5</td>
</tr>
<tr>
<td>Constructed households</td>
<td>6</td>
<td>179</td>
<td>4</td>
<td>31</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Other</td>
<td>86</td>
<td>5</td>
<td>8</td>
<td>-</td>
<td>-</td>
<td>6</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>47</strong></td>
<td><strong>1423</strong></td>
<td><strong>37</strong></td>
<td><strong>88</strong></td>
<td><strong>115</strong></td>
<td><strong>38</strong></td>
</tr>
<tr>
<td><strong>Persons</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Technical</td>
<td>3</td>
<td>105</td>
<td>1</td>
<td>4</td>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td>Constructed household interrelations hip</td>
<td>3</td>
<td>60</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Demographic</td>
<td>19</td>
<td>315</td>
<td>15</td>
<td>12</td>
<td>39</td>
<td>12</td>
</tr>
<tr>
<td>Fertility and mortality</td>
<td>6</td>
<td>49</td>
<td>-</td>
<td>7</td>
<td>1</td>
<td>20</td>
</tr>
<tr>
<td>Nativity and birthplace</td>
<td>18</td>
<td>179</td>
<td>13</td>
<td>3</td>
<td>12</td>
<td>6</td>
</tr>
<tr>
<td>Ethnicity and language</td>
<td>14</td>
<td>77</td>
<td>-</td>
<td>3</td>
<td>16</td>
<td>-</td>
</tr>
<tr>
<td>Education</td>
<td>21</td>
<td>219</td>
<td>4</td>
<td>12</td>
<td>25</td>
<td>6</td>
</tr>
<tr>
<td>Work</td>
<td>21</td>
<td>578</td>
<td>18</td>
<td>38</td>
<td>34</td>
<td>32</td>
</tr>
<tr>
<td>Income</td>
<td>2</td>
<td>30</td>
<td>-</td>
<td>3</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Migration</td>
<td>28</td>
<td>173</td>
<td>5</td>
<td>6</td>
<td>29</td>
<td>-</td>
</tr>
<tr>
<td>Disability</td>
<td>10</td>
<td>6</td>
<td>-</td>
<td>-</td>
<td>2</td>
<td>-</td>
</tr>
<tr>
<td>Other</td>
<td>172</td>
<td>5</td>
<td>16</td>
<td>24</td>
<td>24</td>
<td>5</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>145</strong></td>
<td><strong>1963</strong></td>
<td><strong>62</strong></td>
<td><strong>104</strong></td>
<td><strong>190</strong></td>
<td><strong>84</strong></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>239</strong></td>
<td><strong>4809</strong></td>
<td><strong>99</strong></td>
<td><strong>192</strong></td>
<td><strong>305</strong></td>
<td><strong>122</strong></td>
</tr>
</tbody>
</table>

---

Conclusion: Accessing the Metadata, Maintaining the Database and Future of CIMES

As incomplete as it may be, CIMES already provides an overview describing a very broad range of official microdata from all over Europe; thus greatly improving our knowledge of what is currently available for undertaking research in social sciences and international comparison. The tool and the schema can also provide NSIs with a relevant documentation norm and may hopefully contribute to standardizing the metadata production process across in Europe. It will however need to be maintained and updated. As of yet, CIMES is not publicly available yet, there are plans for a public release by the beginning of next year (2014), when the direct URL to the database will be displayed on the DwB website.

In the future, CIMES can stand alone as a centralized point of access to metadata. A content management system (CMS) could allow identified representatives of NSIs and data archives to update and complete the database for their respective institution/country. Because it is a structured database, it will be easy to reuse the information into any application that would be the base of such a portal. It could thus easily become part of the envisaged European Service Centre whose objective would be to “promote the scientific use of European official statistics microdata by working towards an improved data access infrastructure and by providing comprehensive metadata on integrated European statistics as well as national statistics from Europe” (see report D5.17) to be set up in conjunction with the CESSDA portal. This integration should be made in close cooperation with the development of the MISSY system by GESIS, which documents Eurostat official data; and the IECM database, which provides a centralised access to census microdata samples. Such an integration should also be facilitated by WP12, which is currently setting up a web-based application to harvest and ingest the metadata available across Europe through various sources (incl. CIMES and MISSY); to be later aligned with the CESSDA portal.

7 Report on concept for and components of European Service Centre for official statistics
CIMES should be improved so as to document and display isolated datasets. This would allow providing a better documentation of register data or datasets that are built on the basis of numerous different studies. It could also be improved to document the variable level of a survey. This would lead to a centralized and complete documentation of all European official microdata and would greatly facilitate the use of these data. Such developments could be taken over by the ESC-OS, together with the maintenance and update of the database either centrally or in a decentralised way (see CMS suggestion above), according to the so-called principle of subsidiarity. CIMES will be a decisive step toward the construction of a centralized description of official data in Europe. It will thus greatly improve the use of these data and the integration of European research with economics and social sciences.
Appendix 1 - Specific reports by country on the metadata collection process

**CNRS-RQ, Beneficiary number 1.** In charge of eight different countries

**Denmark:** The data were collected on Statistics Denmark website\(^8\). All statistics have an affiliated “Quality Declaration”, offering a comprehensive description of the sources and methods used. The declarations describe quality and content of data. Part of this documentation was stored into the CIMES database. Considering the significant amount of Danish data, the documentation provided through CIMES only covers a small part of the available datasets.

**Estonia:** the metadata were collected from statistics Estonia website\(^9\), and from their pdf codebook and methodological reports. These were often available in English, but not for every survey or for all years. Statistics Estonia provided further details and clarification when needed.

**Finland:** Similar to Statistics Denmark. The information comes from the website. There were no available documents describing surveys. Certainly because most data are register data and the website provides a precise documentation.

**France:** The documentation available at CNRS-RQ was used.

**Greece:** The metadata were collected through the website of the Greek statistical authority\(^10\) and with the pdf documentation of surveys.

**Latvia:** Almost all data are documented in English. Available pdf documents were used.

---

\(^8\) [http://www.dst.dk/en](http://www.dst.dk/en)


\(^10\) [http://www.statistics.gr/portal/page/portal/ESYE](http://www.statistics.gr/portal/page/portal/ESYE)
Lithuania: Almost all data are documented in English. Available pdf documents were used.

Poland: Almost all data are documented in English. Available pdf documents were used.

Slovakia: We were in contact with their NSI which provided some metadata documentation (also available publicly) and information on the access conditions. But the majority of structured metadata was prepared using Eurostat documentation (60%) and documentation publicly available on their website (40%). We had found 95% all of that documentation on their,

UL, Beneficiary number 4. In charge of four countries

Croatia: The NSI provided the information on the access conditions. The estimate is that 60% of all metadata was prepared using NSI metadata documentation, survey reports and methodological explanations, which are available on their website. The remaining 40% was prepared using Eurostat documentation. In the majority of survey cases, both sources were used to prepare metadata of sufficient quality. In the case of Croatia, the whole 100% of microdata, available to researchers, are covered.

Generally speaking, the main problem is the lack of metadata documentation. There is also very little documentation available for microdata from the 1990s. One of the common problems was the uncertainty linked to the duplication of some fields from previous-year surveys as there was no evidence that there had been no methodological change in survey production between different years.

Country specific problems: Covering all available microdata, 4 survey series can be currently accessed by Croatian registered researchers only, it might change in the future (2014-2015).

Czech Republic: The NSI provided approximately 40% of required metadata documentation. The rest of structured metadata was prepared using publicly available Eurostat documentation. Metadata have been prepared for approximately 40% of their microdata. Their NSI rated the covered microdata as the most important for researchers.

Slovakia: We were in contact with their NSI which provided some metadata documentation (also available publicly) and information on the access conditions. But the majority of structured metadata was prepared using Eurostat documentation (60%) and documentation publicly available on their website (40%). We had found 95% all of that documentation on their
Metadata have been prepared for approximately 60% of their microdata. Their NSI rated the covered microdata as the most important for researchers.

Slovenia: UL maintains close relations with the NSI; collecting metadata documentation to prepare structured metadata was therefore easier, although that available documentation remains limited. Rough estimate would be that 60% of metadata was prepared using NSI documentation publicly available (quality reports, methodological explanation, standard reports, research webpages etc.), 30% of metadata was prepared using Eurostat documentation (quality reports, metadata webpages etc.) and 10% of metadata was prepared using restricted-access NSI documentation and information provided by NSI employees.

No (semi)automatic harvesting of metadata from their systems/applications was possible. For some surveys there is no available NSI documentation.

Currently there are 6 completely covered survey series, 1 incompletely covered series and 12 series with only series level metadata available (due to lack of metadata documentation to prepare study level metadata and lack of PM). That means that almost 10% of survey series covered. We have to be aware of the fact that, at least in theory, registered researchers can access all NSI microdata, both survey and register.

The criteria were: (a) nr. of researchers accessing certain microdata, (b) importance of data for the research community, (c) availability of metadata documentation

GESIS, Beneficiary number 5. In charge of three countries

Austria: Statistik Austria homepage\textsuperscript{11}. ~70% of available surveys are covered; same criteria as in Germany were applied. However Statistik Austria appears to be the only provider of OS microdata in Austria. No problems encountered. All documents are easily made available on a well-structured website and Statistik Austria is very responsive to enquiries.

\textsuperscript{11} http://www.statistik.at/
Germany: The metadata come from DESTATIS FDZ\(^{12}\), IAB FDZ\(^{13}\) and GESIS German Microdata Lab homepage\(^{14}\). Older German data from DESTATIS tend to be poorly documented and it is hard to find any information on issues such as sampling. Beyond that documentation is sufficient to fill all fields of the metadata schema.

For Germany roughly 20\% of all available OS data were covered. The criteria employed were: relevance (used by research and relevant to social sciences), cross national comparability, duration (long running series preferred over one time surveys), broad topical coverage of database, focus on ‘official’ statistics (produced by government institutions).

Netherlands: Metadata were collected from the CBS homepage\(^{15}\), the DANS homepage\(^{16}\), and the SCP homepage\(^{17}\). Metadata were compiled from existing documentation available on the sites listed above. Most of this information had to be translated and was compiled in excel sheets which were then re-entered into CIMES. This two-step procedure for data entry was chosen in order to minimize errors in the documentation. CBS and DANS supported us in our efforts by providing access to additional documents and verifying our metadata. The task of documenting Dutch data was complicated by the fact that no English language documentation is available. The existing documentation is produced for internal use at CBS but tends to contain sufficient information for documentation. Another issue encountered was that the way in which register data are distributed by CBS (many small datasets within one series) does not fit with the initial idea behind the metadata schema. We choose to document series as they are defined by CBS as 1 series, individual registers were classified as studies as this fit in the metadata schema fairly well. The dataset level was used to highlight access conditions, which are identical for all datasets.

\(^{12}\) [http://www.forschungsdatenzentrum.de/](http://www.forschungsdatenzentrum.de/)

\(^{13}\) [http://fdz.iab.de/en.aspx](http://fdz.iab.de/en.aspx)


\(^{15}\) [http://www.cbs.nl/nl-NL/menu/home/default.htm](http://www.cbs.nl/nl-NL/menu/home/default.htm)

\(^{16}\) [http://www.dans.knaw.nl/en/content/access-cbs-data](http://www.dans.knaw.nl/en/content/access-cbs-data)

\(^{17}\) [http://www.scp.nl/Onderzoek/Bronnen/Beknopte_onderzoeksbeschrijvingen/Permanent_onderzoek_naar_de_leefsituatie_PO_LS](http://www.scp.nl/Onderzoek/Bronnen/Beknopte_onderzoeksbeschrijvingen/Permanent_onderzoek_naar_de_leefsituatie_PO_LS)
We covered ~10% of available series. However those covered are the most extensive and relevant to social science; as the objective of the documentation is to cover the largest series and those that could be compared internationally. The metadata in CIMES referring to CBS data were checked and validated by CBS.

**RODA, Beneficiary number 8. In charge of one country**

**Romania:** The Romanian Social Data Archive (RODA) collaborated with the National Institute of Statistics from Bucharest. The documentation was available in printed version (Catalogues) and online, on the NSI website. The NSI metadata base consists of structured methodologies - on which basis the statistical research from the Yearly Statistical Plan is developed by the NSI or other of its associated institution - definitions of the resulting indicators and the most useful statistical terms, as well as theoretical statistical information.

**FORS, Beneficiary number 17. In charge of three countries**

We used the Belgian and Luxembourg NSI’s websites, as well as direct discussions with the DwB’s contact person (as per institutional map). For the Swiss NSI we used our own available information (COMPASS), crosschecked with the persons in charge of the specific surveys at the NSI.

Coverage: 95% of the Swiss surveys considered as interesting for the researchers in social sciences (by the NSI), 80% of the equivalent Belgian and Luxemburg survey.

**ONS, Beneficiary number 24. In charge of two countries**

**Ireland:** Data for Ireland came from the Central Statistics Office website. This covers the majority of government social surveys in Ireland. A description of each survey and information on questionnaires and field practices is available, but other meta-data varied from survey to survey. The surveys themselves tend to be carried out by other organisations, such as the

---


Economic and Social Research Institute, and data is generally deposited with Irish Social Sciences Data Archive.

**UK:** There are two complications with surveys in the UK. Firstly, surveys cover different nations: some cover England and Wales, others Great Britain (England, Wales and Scotland), and others UK. The geographical coverage field contains notes on which parts of the UK are covered, but where multiple surveys exist within the UK only the survey covering the largest territory can be described in detail.

The other complication is the variety of bodies who own the surveys. The Office for National statistics owns many surveys for England and Wales, or carries them out on behalf of other government departments. However, the most relevant surveys for CIMES include those conducted by organisations other than national statistics institutes. The UK data archive contains details of all major government social surveys, and so metadata were taken from there as well as from the websites of survey owners (ONS and NATCEN).


**Irish archive:** [http://www.ucd.ie/issda/](http://www.ucd.ie/issda/)

**UKDA:** [http://discover.ukdataservice.ac.uk/](http://discover.ukdataservice.ac.uk/)


---

**CED. Beneficiary number 25. In charge of three countries**

**Italy:** The data were collected at *Istituto nazionale di statistica (ISTAT)*[^20] Microdata[^21]: Less than 25% of the documentation was available in English, so we had to translate from Italian into English. Around 60% of the available datasets have been covered. We gave priority to the population, labor force, health, living conditions and education surveys.

**Portugal:** The data were collected at INE, the Instituto Nacional de Estatística\(^{22}\). The microdata files are stored on the website of the General Statistics, Science and Education Directorate. Almost all the available information is in Portuguese. There are very few documents in English. Roughly, 50% of surveys have been covered. We have basically covered Population, labour force, living conditions and health surveys.

**Spain:** The data were collected at INE (the Instituto Nacional de Estadística\(^{23}\)), CIS (Centro de Investigaciones Sociológicas\(^{24}\)), and the Ministery of employment and Social Security\(^{25}\). There were few but still some documents that were only available in Spanish. The information about the series is often incomplete. The Centre d'Estudis Demogràfics had some documentation about the Spanish Labour Force Survey that was not available on the INE's website. 75% of available surveys have been covered. We have basically covered all microdata series related to Population, living conditions, labour force, health and education. We have also taken into account those studies with long series of data.

\(^{23}\) www.ine.es
\(^{25}\) [http://www.seg-social.es/Internet_1/Estadistica/Est/Muestra_Cotinua_de_Vidas_Laborales/index.htm](http://www.seg-social.es/Internet_1/Estadistica/Est/Muestra_Cotinua_de_Vidas_Laborales/index.htm)
Appendix 2 - WP5 Task 2: Metadata Schema

The documentation with DwB WP5 Task 2 shall follow a nested hierarchy which contains the following elements: **Series, Study, and Dataset**. The series level describes a data collection programme which is carried out over time (e.g. German Microcensus). The study level describes an individual instance of this series, usually a year in which a study was carried out (e.g. German Microcensus 2007).

The dataset level then describes different versions of this study which are issued by the data producer. For the purpose of WP5 Task 2 this will most often be files with varying degrees of anonymisation and accessibility such as Scientific or Public Use Files.

In many cases, the relations between these three objects can be conceived of as a simple three levels hierarchy: a dataset belongs to a study which belongs to a series. A study can have one or more datasets associated with it. Information is therefore inherited downward between these levels wherever possible. However, this hierarchical representation does not cover all the possible relations between those three objects. Some studies are not attached to series, and some datasets can be attached to different studies or correspond only to a part of a study.
1) **Series**

- **CIMES Series Identification Number:** Automatically filled by the application.
- **Original Title:** Provides the title of the series in the original language.
- **English Title:** Provides the English title of the series. If such a title is not provided by the data producer it should be translated by us and possibly harmonized within the project.
- **Original Subtitle:** Provides the subtitle of the series.
- **English Subtitle:** Provides the subtitle of the series in English. If such a title is not provided by the data producer it should be translated by us and possibly harmonized within the project.
- **Original Alternative Title:** As defined by the data producer.
- **English Alternative Title:** As defined by the data producer in English.
- **Abstract:** A short summary describing the purpose, nature, and scope of the series. The abstract should not exceed 500 words.
- **Keywords:** Thematic coverage of the series. We will use ELSST for controlled vocabulary. Several keywords can be used.
- **Geographic Coverage:** Information on the geographic coverage of the series. Includes the total geographic scope of the data.
  - A controlled vocabulary is provided.
  - Other: This field should only be used to add some information to the controlled vocabulary (for example, if the series covers only a region of the country).
- **Time Period Covered:** The time period to which the data refer. This item reflects the time period covered by the data: first and last years of the series.
  - Start Date
  - End Date
- **Time Method:** Describes the time dimension of the data collection.
  - A controlled vocabulary is provided
  - Other: This field should be filled only when “other” has been selected in the controlled vocabulary.
- **Notes:** Put any information here that you feel is important for researchers and that cannot be included in any other field.
2) Study

- **CIMES Study Identification Number**: Automatically filled by the application.
- **Series Name**: Automatically filled by the application.
- **Original Title**: Provides the original title of the study and its official date.
- **English Title**: Provides the title of the study in English and its official date. If such a title is not provided by the data producer it should be translated by us and possibly harmonized within the project.
- **Original Subtitle**: Provides the subtitle of the study.
- **English Subtitle**: Provides the subtitle of the study in English. If such a subtitle is not provided by the data producer it should be translated by us and possibly harmonized within the project.
- **Original Alternative Title**: As defined by the data producer.
- **English Alternative Title**: As defined by the data producer in English.
- **Producer**: The legal entity/entities responsible for carrying out the study (in English). This does not necessarily have to be the data publisher.
  - Name
  - Abbreviation
- **Abstract**: Provide here any information that is pertinent to the specific study but not the series. For example you should mention here any ad-hoc modules implemented in the current study or alterations in study design. This information should be complementary to the information provided in the abstract at the series level.
- **Keywords**: *The information will be inherited from the series level, you should only add the information specific to the study and not to the series.*
- **Geographic Coverage**: Information on the geographic coverage of the study. Includes the total geographic scope of the data, and any additional levels of geographic coding provided in the variables. *The information will be inherited from the series level, you should only add the information specific to the study and not to the series.*
  - A controlled vocabulary is provided.
  - Other: This field should only be used to add some information to the controlled vocabulary (for example, if the study covers only a region of the country).
- **Universe**: The group of persons or other elements that are the object of the research.

- **Sampling Procedure**: The sampling fraction of the study in reference to the Universe.

- **Date of Collection**: Details when data were collected.
  - Start Date (YYYY/MM/DD)
  - End Date (YYYY/MM/DD)
  - Cycle
  - Other

- **Time Period Covered**: The time period the data refer to. This can correspond to different points in time (if, for example, retrospective questions were asked).

- **Kind of Data**: Indicates the type(s) of data used within the study. Multiple data types should be able to be selected from a list of controlled vocabularies.

- **Documents**: Provides URL link to version of questionnaires, interviewer instructions, quality reports and other adjoining documents. These documents should be presented in the following structure (see below). A controlled vocabulary id provided for language. Links should lead to the institution hosting the file, however additionally these files could be stored. A broken link function could also be included so that users can report if links go down.

<table>
<thead>
<tr>
<th><strong>Document name</strong></th>
<th><strong>Document type</strong></th>
<th><strong>Language</strong></th>
<th><strong>Link</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Questionnaires</td>
<td></td>
<td>French</td>
<td>http</td>
</tr>
<tr>
<td>Interviewer instructions</td>
<td></td>
<td>German</td>
<td>http</td>
</tr>
<tr>
<td>Quality Report</td>
<td></td>
<td>English</td>
<td>http</td>
</tr>
</tbody>
</table>

- **Sources**: List all publications cited in the study description here. This is required due to copyright reasons and also provides a valuable resource for researchers.

- **Notes**: Put any information here that you feel is important for researchers and that cannot be included in any other field.
3) **Dataset**

- **CIMES Dataset Identification Number:** Automatically filled by the application.
- **Disseminating Identification Number:** Identification Number provided by the disseminator of the dataset.
- **Original Name:** The name of the dataset as provided by the data producer (e.g. Mikrozensus 2007 Scientific Use File). We could add here a complementary identification if the producer gave the same names to different datasets (because he used the title of the study).
- **English Name:** English name of the dataset. If such a name is not provided by the data producer it should be translated by us and possibly harmonized within the project.
- **Access Conditions:** Describes the **Type of Access** (File possession / Onsite Access / Remote Access or Execution) and the **Costs of Access** (Free or not) for each **Type of User** (Public / Student / Researchers):

<table>
<thead>
<tr>
<th>Type of Access</th>
<th>File possession</th>
<th>Onsite Access</th>
<th>Remote Access</th>
<th>Remote Execution</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Free</td>
<td>Not free</td>
<td>Free</td>
<td>Not free</td>
</tr>
<tr>
<td>Public</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Students</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PhD Students</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Researchers</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Foreign Researchers</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **Other Access Conditions:** Details the terms of access for the specific dataset as issued by the publisher.
- **General description of the NSI's accreditation procedures:** Link to the WP3 web pages describing the accreditation procedures for each NSI.
- **Contact:** Provides information on the institutions granting access to data and relevant contact information. Reference to specific persons should be avoided.
  - Institution Name
  - URL
  - Email
- **Notes:** Put here any dataset specific information not listed anywhere else.
4) **Controlled Vocabularies**

**Keywords**

Weblink: [Click here](#)

**ECONOMICS**
Consumption/consumer behaviour  
Economic conditions and indicators  
Economic policy  
Economic systems and development  
Income, property and investment/saving  
Rural economics

**TRADE, INDUSTRY AND MARKETS**
Agricultural, forestry and rural industry  
Business/industrial management and organisation

**LABOUR AND EMPLOYMENT**
Employment  
In-job training  
Labour relations/conflict  
Retirement  
Unemployment  
Working conditions

**POLITICS**
Conflict, security and peace  
Domestic political issues  
Elections  
Government, political systems and organisations  
International politics and organisations  
Mass political behaviour, attitudes/opinion  
Political ideology
LAW, CRIME AND LEGAL SYSTEMS

Crime
Law enforcement
Legal systems
Legislation
Rehabilitation/reintegration into society

EDUCATION

Basic skills education
Compulsory and pre-school education
Educational policy
Life-long/continuing education
Post-compulsory education
Teaching profession
Vocational education

INFORMATION AND COMMUNICATION

Advertising
Information society
Language and linguistics
Mass media

HEALTH

Accidents and injuries
Childbearing, family planning and abortion
Drug abuse, alcohol and smoking
General health
Health care and medical treatment
Health policy
Nutrition
Physical fitness and exercise
Specific diseases and medical conditions

**NATURAL ENVIRONMENT**
Environmental degradation/pollution and protection
Natural landscapes
Natural resources and energy
Plant and animal distribution

**HOUSING AND LAND USE PLANNING**
Housing
Land use and planning

**TRANSPORT, TRAVEL AND MOBILITY**

**SOCIAL STRATIFICATION AND GROUPINGS**
Children
Elderly
Elites and leadership
Equality and inequality
Family life and marriage
Gender and gender roles
Minorities
Social and occupational mobility
Social exclusion
Youth

**SOCIETY AND CULTURE**
Community, urban and rural life
Cultural activities and participation
Cultural and national identity
Leisure, tourism and sport
Religion and values
Social behaviour and attitudes
Social change
Social conditions and indicators
Time use

**DEMOGRAPHY AND POPULATION**
Censuses
Fertility
Migration
Morbidity and mortality

**SOCIAL WELFARE POLICY AND SYSTEMS**
Social welfare policy
Social welfare systems/structures
Specific social services: use and provision

**SCIENCE AND TECHNOLOGY**
Biotechnology
Information technology

**PSYCHOLOGY**

**HISTORY**

**REFERENCE AND INSTRUCTIONAL RESOURCES**
Computer and simulation programs
Reference sources
Teaching packages and test datasets

**GEOGRAPHIC COVERAGE**
Weblink: [Click here](#)

<table>
<thead>
<tr>
<th>Austria</th>
<th>Czech Republic</th>
<th>France</th>
</tr>
</thead>
<tbody>
<tr>
<td>Belgium</td>
<td>Denmark</td>
<td>Germany</td>
</tr>
<tr>
<td>Bulgaria</td>
<td>Estonia</td>
<td>Greece</td>
</tr>
<tr>
<td>Cyprus</td>
<td>Finland</td>
<td>Hungary</td>
</tr>
</tbody>
</table>
Iceland  Malta  Slovakia
Ireland  Netherlands  Slovenia
Italy  Norway  Spain
Latvia  Poland  Sweden
Lithuania  Portugal  Switzerland
Luxembourg  Romania  United Kingdom

**Kind of Data**

Survey data
Time budget diaries
Clinical data
Psychological test
Observation data/ratings
Census/enumeration data
Aggregate data
Administrative records data
Process-produced data
Event/transaction data
Voting data
Other

**Time Method**

Weblink: [Click here](#)

**Longitudinal:** Data collected repeatedly over time to study change in a population.

**Longitudinal: Cohort/Event-based:** Data collected over time about a group of individuals that are connected in some way or have shared some significant experience within a given period. Examples: birth, disease, education, employment, family formation, participation in an event.

---

26 These controlled vocabularies were taken from “Guide for study documentation in NESSTAR” by NSD.
Longitudinal: Trend/Repeated cross-section: Studies different samples/different groups of people from the same population at several points in time, using the same set of questions/variables. Conclusions are drawn for the population. Examples: public opinion polls, elections studies, etc.

Longitudinal: Panel: Data collected over time from, or about, the same sample of respondents.

Longitudinal: Panel: Continuous: Reports from the panel are collected on a regular basis.

Longitudinal: Panel: Interval: Measurements are taken only when information is needed.

Time Series: Data collected repeatedly over time to study change in observations. These are typically "objective" measurements of phenomena that can be observed externally, as opposed to attitudes/opinions or feelings. Examples may include economic/financial indicators, natural/meteorological phenomena, vital statistics, etc.

TimeSeries: Continuous: Phenomena are measured at every instant in time. Examples: lie detectors, electrocardiograms, etc.

TimeSeries: Discrete: Measurements are taken at (usually regularly) spaced intervals. Examples: macroeconomics (weekly share prices, monthly profits, sales); meteorology (hourly temperature); measurements of individuals (blood pressure, weight, height); sociology (crime figures, employment figures), etc.

Cross-section: Data about a population are obtained only once.

Cross-section ad-hoc follow-up: Data collected at one point in time to complete information collected in a previous cross-sectional study; the decision to collect follow-up data is not included in the study design.

Other: Use if the time method is known, but not found in the list.

Language

Weblink: [Click here](#)

Spanish       Danish       Estonian
Czech         German       Greek
<table>
<thead>
<tr>
<th>English</th>
<th>Romanian</th>
<th>Macedonian</th>
</tr>
</thead>
<tbody>
<tr>
<td>French</td>
<td>Slovak</td>
<td>Turkish</td>
</tr>
<tr>
<td>Irish</td>
<td>Slovenian</td>
<td>Norwegian</td>
</tr>
<tr>
<td>Italian</td>
<td>Finnish</td>
<td>Albanian</td>
</tr>
<tr>
<td>Latvian</td>
<td>Swedish</td>
<td>Belarusian</td>
</tr>
<tr>
<td>Lithuanian</td>
<td>Basque</td>
<td>Bosnian</td>
</tr>
<tr>
<td>Hungarian</td>
<td>Galician</td>
<td>Russian</td>
</tr>
<tr>
<td>Maltese</td>
<td>Luxembourgish</td>
<td>Serbian</td>
</tr>
<tr>
<td>Dutch</td>
<td>Welsh</td>
<td>Ukrainian</td>
</tr>
<tr>
<td>Polish</td>
<td>Croatian</td>
<td>Arabic</td>
</tr>
<tr>
<td>Portuguese</td>
<td>Icelandic</td>
<td></td>
</tr>
</tbody>
</table>
# Appendix 4 - CIMES User Guide

## TABLE OF CONTENTS

| 1. Log | .......................................................................................................................... 39 |
| 2. Functionalities | ....................................................................................................................... 39 |
| 3. Fields in “add a Series” | .................................................................................................................. 41 |
| 4. Fields in Add a Study | .................................................................................................................. 42 |
| 5. Fields in Add a Dataset | .................................................................................................................. 44 |
1) **Log**

Enter your email and the password and click on "Login".

To log out, click on the cross on the top of the right corner of the page.

2) **Functionalities**

**Add a series/study/dataset**

To create a new series, study or dataset, you just have to click on the tabs at the top of the page (blue tab for adding a Series, orange tab for adding a Study, and red tab for adding a Dataset) and fill in the form.

**Modify a series/study/dataset**

You have two possibilities to modify a series, study or dataset:
- After the validation of the information entered, there is a menu with a button "Modify".
- On the home page: there is a button "Modify" at the bottom of the displayed information for one level.

**Duplicate a series/study/dataset**

- After the validation of the information entered, there is a menu with a button "Duplicate".
- On the home page: at the bottom of the displayed information for one level, there is a button "Duplicate".

Once you’ve clicked on duplicate, you can choose the fields to duplicate among all the fields of a series/study/dataset (you can choose all the fields by checking the box "Select all checkboxes").

For the study level, you can also choose if you want to keep the duplicated study in the same series. If not, you can choose the series to which this new study belongs in the filling page. Idem for the dataset level.
Delete a level
- On the home page: there is a button "Delete" at the bottom of the displayed information.

Add a study to a series

Two ways:
- After the validation of the information entered for a series, there is a menu with a button "Add a study to this Series".
- On the home page: on the bottom of the information for one level, there is a button "Add a study to this Series".

Then you can choose which fields to inherit from a Series to a Study.
- Original Title
- English Title
- Original Subtitle
- English Subtitle
- Original Alternative Title
- English Alternative Title
- Keywords
- Geographic Coverage
- Geographic Coverage Other

Add a dataset to a study

Two ways:
- After the validation of the information entered for a series, there is a menu with a button "Add a dataset to this study ".
- On the home page: on the bottom of the information for one level, there is a button "Add a dataset to this study ".

Visualize the information entered
You can click on the button "Visualize" in the menu which appears after the validation of the information entered for a series/study/dataset.

Import from DDI
You can import a DDI XML file to fill in the information at the study level.
Click on the yellow tab named DDI. Choose the file you want to import and then click on the button "Upload". The conversion can take a few seconds. Then click on "Review before import". You can choose to display the XML tree by checking the box "Show XML Tree". The information imported appears and you can choose the fields to import in the study level or cancel the import.

Once the import is completed, the filling page appears and you still have to fill in the mandatory fields which haven't been imported.

The filling pages
- Fields with a * are mandatory, if you don’t fill them, the validation won’t work
- If you want some explanation about a field, there is an interrogation mark next to each field on which you can click.

- Once you have filled all the mandatory fields, you can validate the creation of the series/study/dataset. If you don't want to save the data entered, click on "Back to the welcome menu".

3) Fields in “add a Series”
- **Original Title**: Title of the series in the original language.
  ➔ Mandatory field.

- **English Title**: English title of the series. If such a title is not provided by the data producer it should be translated by us and possibly harmonized within the project.
  ➔ Mandatory field.

- **Original Subtitle**: Subtitle of the series.

- **English Subtitle**: Subtitle of the series in English. If such a title is not provided by the data producer it should be translated by us and possibly harmonized within the project.

- **Original Alternative Title**: Alternative title of the series in original language.
- **English Alternative Title**: Alternative title of the series in English. If such a title is not provided by the data producer it should be translated and possibly harmonized within the project.

- **Abstract**: A short summary describing the purpose, nature, and scope of the series.
  ➔ Mandatory field.

- **Keywords**: Thematic coverage of the series.
  ➔ If you choose one keyword in the second level, the corresponding keyword of the upper level will be automatically checked.
  ➔ Mandatory field. Multiple field.

- **Geographic Coverage**: Information on the geographic coverage of the series. Other: This field should only be used to add some information to the controlled vocabulary (for example, if the series covers only a region of the country).
  ➔ Mandatory field. Multiple field.

- **Time Period Covered**: The time period covered by the data: first and last years of the series.
  ➔ Mandatory field.

- **Time Method**: Describes the time dimension of the data collection. Other: This field should be filled only when “other” has been selected in the controlled vocabulary.
  ➔ Multiple field.

- **Notes**: Put any information here that you feel is important for researchers and that cannot be included in any other field.

4) **Fields in Add a Study**

- **Series Name**: Series of the study.

- **Original Title**: Title of the study in the original language and its official date.
  ➔ Mandatory field.

- **English Title**: English title of the study and its official date. If such a title is not provided by the data producer it should be translated by us and possibly harmonized within the project.
  ➔ Mandatory field.
- **Original Subtitle**: Subtitle of the study.

- **English Subtitle**: Subtitle of the study in English. If such a subtitle is not provided by the data producer it should be translated by us and possibly harmonized within the project.

- **Original Alternative Title**: Alternative title of the study in original language.

- **English Alternative Title**: Alternative title of the study in English. If such a title is not provided by the data producer it should be translated and possibly harmonized within the project.

- **Producer**: The legal entity/entities responsible for carrying out the study (in English). This does not necessarily have to be the data publisher.
  ➔ Click on "Add a producer", then choose a producer in the list. If it doesn't exist you can add a new one by clicking on "New producer...". Add the producer name and abbreviation.
  ➔ Mandatory field. Multiple field.

**Abstract**: Provide here any information that is pertinent to the specific study but not the series. For example you should mention here any ad-hoc modules implemented in the current study or alterations in study design.

This information should be complementary to the information provided in the abstract at the series level.
  ➔ Mandatory field.

- **Keywords**: Thematic coverage of the study.
  ➔ Mandatory field.

- **Geographic Coverage**: Information on the geographic coverage of the study. Other: This field should only be used to add some information to the controlled vocabulary (for example, if the study covers only a region of the country).
  ➔ Mandatory field. Multiple field.

- **Universe**: The group of persons or other elements that are the object of the research.

- **Sampling Procedure**: The sampling fraction of the study in reference to the Universe.

- **Date of Collection**: Details when data were collected.
  ➔ Multiple field.
- **Time Period Covered:** The time period the data refer to. This can correspond to different points in time (if, for example, retrospective questions were asked).

- **Kind of Data:** Indicates the type(s) of data used within the study. Multiple data types should be able to be selected from a list of controlled vocabularies.
  ➔ Mandatory field. Multiple field.

- **Documents:** Provides URL link to version of questionnaires, interviewer instructions, quality reports and other adjoining documents.
  ➔ Multiple field.

- **Sources:** List all publications cited in the study description here. This is required due to copyright reasons and also provides a valuable resource for researchers.

- **Notes:** Put any information here that you feel is important for researchers and that cannot be included in any other field.

**5) Fields in Add a Dataset**

- **Disseminating Identification Number:** Identification Number provided by the disseminator of the dataset.

- **Original Name:** The name of the dataset as provided by the data producer (e.g. Mikrozensus 2007 Scientific Use File). We could add here a complementary identification if the producer gave the same names to different datasets (because he used the title of the study).
  ➔ Mandatory field.

- **English Name:** English name of the dataset. If such a name is not provided by the data producer it should be translated by us and possibly harmonized within the project.
  ➔ Mandatory field.

- **Access Conditions:** Describes the **Type of Access** (File possession / Onsite Access / Remote Access or Execution) and the **Costs of Access** (Free or not) for each **Type of User** (Public / Student / Researchers).
  ➔ For each type of user, choose if the data are available and if it is for free.

- **Other Access Conditions:** Details the terms of access for the specific dataset as issued by the publisher.
- **General description of the NSI's accreditation procedures**: Link to the WP3 web pages describing the accreditation procedures for each NSI.
  ➤ Don't add anything to this field. We will fill it once the page will be created.

- **Contact**: Provides information on the institutions granting access to data and relevant contact information. Reference to specific persons should be avoided.
  ➤ Choose a contact in the list. If it doesn't exist you can add a new one by clicking on "New...". Add the institution name, an URL and an email.
  ➤ Mandatory field

- **Notes**: Put here any dataset specific information not listed anywhere else.