FIRST REPORTING PERIOD
STRATEGIC REPORT
JUNE 2015
ANNEXES

FP7 - Grant agreement n°313082
# TABLE OF CONTENTS

## PART 1: OPENING ACTIVITIES

<table>
<thead>
<tr>
<th>WP6 – COORDINATED PREPARATION TO THE OPENING OF DATASETS</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 – Objectives of WP as of Description of Work</td>
<td>4</td>
</tr>
<tr>
<td>2 – Work Programme Description</td>
<td>4</td>
</tr>
<tr>
<td>3 – Activities engaged and participants</td>
<td>6</td>
</tr>
<tr>
<td>4 – Outputs</td>
<td>8</td>
</tr>
<tr>
<td>5 – Challenges</td>
<td>9</td>
</tr>
<tr>
<td>6 – Program for next year and milestones</td>
<td>9</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>WP7 – PLATFORM OPENING</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 – CORTEXT progress report</td>
<td>10</td>
</tr>
<tr>
<td>2 – SMS progress report</td>
<td>12</td>
</tr>
<tr>
<td>2.1. SMS platform: a non-technical description of the idea and of progress</td>
<td>12</td>
</tr>
<tr>
<td>2.2. Work done in the first period</td>
<td>13</td>
</tr>
<tr>
<td>2.3. A non-technical description of the platform</td>
<td>14</td>
</tr>
<tr>
<td>2.4. SMS, a wish list for the next period</td>
<td>23</td>
</tr>
</tbody>
</table>

## PART 2: ACTIONS ACCOMPANYING DATASET OPENING

<table>
<thead>
<tr>
<th>1 – TRAINING ACTIVITIES</th>
<th>25</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 – Objectives of WP as of Description of Work</td>
<td>25</td>
</tr>
<tr>
<td>2 – Work programme Description</td>
<td>27</td>
</tr>
<tr>
<td>3 – Activities engaged and participants</td>
<td>28</td>
</tr>
<tr>
<td>4 – Outputs</td>
<td>29</td>
</tr>
<tr>
<td>5 – Challenges</td>
<td>32</td>
</tr>
<tr>
<td>6 – Program for next year and milestones</td>
<td>32</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>2 – GEOGRAPHY BASED HARMONISATION</th>
<th>32</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 – Objectives</td>
<td>32</td>
</tr>
<tr>
<td>2 – Work programme</td>
<td>33</td>
</tr>
<tr>
<td>3 – Activities engaged (and participants)</td>
<td>33</td>
</tr>
<tr>
<td>4 – Outputs</td>
<td>35</td>
</tr>
<tr>
<td>5 – Challenges</td>
<td>35</td>
</tr>
<tr>
<td>6 – Programme for next months</td>
<td>35</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>3 – WP8 REPORT ON THE ETER REGISTRY</th>
<th>36</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1. Overview of the facility</td>
<td>36</td>
</tr>
<tr>
<td>1.2. Context and aims</td>
<td>37</td>
</tr>
<tr>
<td>1.3. Legal name of the operating organization</td>
<td>38</td>
</tr>
<tr>
<td>1.4. Database location and type of access</td>
<td>38</td>
</tr>
<tr>
<td>2 – Data content of ETER</td>
<td>38</td>
</tr>
<tr>
<td>2.1. Definition and description of observations</td>
<td>38</td>
</tr>
<tr>
<td>2.2. Number of units and demography</td>
<td>39</td>
</tr>
<tr>
<td>2.3. Data acquisition and processing (e.g. data cleaning)</td>
<td>40</td>
</tr>
<tr>
<td>2.4. Cleaning of the EUMIDA datasets</td>
<td>40</td>
</tr>
<tr>
<td>2.5. Information all variables/indicators</td>
<td>42</td>
</tr>
<tr>
<td>2.6. Temporal, geographical and sectoral coverage</td>
<td>44</td>
</tr>
<tr>
<td>2.7. Classifications used in the database</td>
<td>45</td>
</tr>
<tr>
<td>3 – Quality and accuracy of data</td>
<td>48</td>
</tr>
<tr>
<td>3.1. Overall data quality approach</td>
<td>48</td>
</tr>
<tr>
<td>3.2. Accuracy and consistency</td>
<td>49</td>
</tr>
</tbody>
</table>
PART 3: NEW DATASET DEVELOPMENT ................................................................. 59

1 – INNOVATION DYNAMICS OF FIRMS (WP20) .................................................. 59
   1 – Objectives ............................................................................................................. 59
   2 – Work programme Description ............................................................................. 59
   3 – Activities engaged and participants ................................................................. 60
   4 – Substantive analysis of what has been done and where do the WP stands ........ 60
   5 – Outputs .................................................................................................................. 62
   4 – Challenges ............................................................................................................. 63
   5 – Programme for next months (with milestones anticipated) ............................... 63

2 – TYPES AND DYNAMICS OF PUBLIC RESEARCH ORGANISATIONS (WP23) .... 64

3 – FOSTERING EUROPEAN INTEGRATION (WP22) .............................................. 66
   1 – Objectives of WP as of Description of Work ...................................................... 66
   2 – Work programme Description .......................................................................... 67
   3 – Activities engaged and participants ................................................................. 68
   4 – Outputs .................................................................................................................. 68
   5 – Challenges ............................................................................................................. 70
   6 – Programme for next year and milestones ......................................................... 71

Annex – The dynamics of the Europeanisation of research activities: New insights from RISIS research infrastructures …… 71

4 – INTEGRATING FRAMEWORK AND DATASET FOR ANALYSING RESEARCHER CAREERS (WP24) ................................................................. 73
   1 – Objectives of WP as of Description of Work ...................................................... 73
   2 – Work programme Description .......................................................................... 74
   3 – Activities engaged and participants ................................................................. 75
   4 – Outputs .................................................................................................................. 75
   5 – Challenges ............................................................................................................. 79
   6 – Programme for next year and milestones ......................................................... 79

5 – Science and innovation policy evaluations repository (SIPER, WP6, WP21) ........ 79
  Annex to WP7 - Demonstrating advanced data integration with the SMS platform: Grant selection bias as case ……….. 81
  Introduction ................................................................................................................. 81
  SMS platform .............................................................................................................. 81
  The case ....................................................................................................................... 82
  The SMS platform demonstrator ............................................................................... 83
  The workflow ............................................................................................................. 85
  The results expected from the demonstrator ............................................................ 88
WP6 – COORDINATED PREPARATION TO THE OPENING OF DATASETS

1 – Objectives of WP as of Description of Work

WP6 constitutes the core work package of RISIS with respect to pushing forward the structured and systemic opening of RISIS data infrastructures. The overall objective is to coordinate and organize the preparation for the opening of 10 data infrastructures (see Section 3 for a list of the datasets under consideration). These infrastructures are characterized by

- mostly local use,
- very different technical choices which were selected for research developments rather than for favouring access and interconnectivity,
- ad-hoc solutions for solving disambiguation issues and for categorizations,
- low levels of or even no documentation (on their contents, but even more on their sphere of validity and on relevant conditions of use)

To address the overall objective, several subtasks have to be conducted, discussed in some detail in the context of the WP6 work program in Section 2. The subtasks are addressing different dimensions – among them substantive, technical and legal dimensions – that are necessary to be addressed in the opening process.

The opening process will actually take place over the whole project period of 48 months, starting in month 3 of the RISIS project. The coordination process should be organized in a way so that local technical choices made are compatible between datasets, classifications are as much as possible harmonized and when this is not yet possible, solutions adopted can be replaced by common solutions developed within other WPs (especially for organizational and geographical harmonization’s, see WP8 and WP9). Moreover, WP6 builds the legal basis for ensuring a smooth opening up of the datasets. This has entailed designing and implementing a) an authorization procedure before datasets are opened and b) an accreditation system for organizations and individuals wishing to use the opened facilities (see Deliverable 6.1a).

2 – Work Programme Description

In what follows, a more detailed description of the WP6 work programme is provided. Figure 1 provides an overview on the main tasks to be conducted during the course of the WP, also highlighting the respective workflow. The work program of WP6 consists of six inter-related subtasks, with two tasks, Task 1 and Task 6 constituting documentations of the starting point and the results, respectively, and Task 2 – Task 5 constituting respective actions necessary for opening along four different dimensions.

Task 1 of WP6, finalized in the first reporting period (see Deliverable D6.1), was intended to produce reports on the content and technical structure of each of the 10 datasets under consideration. The main essence has been to produce harmonized reports that lay the basis for a systematic screening of the characteristics of each dataset. This has been essential to gather systematic information for the subsequent tasks. The reports were intended to feature a harmonized structure, providing detailed information on each dataset, including systematic
information on the substantive content of the dataset, legal issues encountered and access conditions, and the technical structure of the dataset.

The reports on the content and technical structure of the 10 datasets have constituted the basis for the following tasks. **Task 2** aimed to coordinate the work on the different categorizations used in the datasets. In this context, a collective discussion has been developed for all categorizations used in order to insure harmonization. Furthermore, potential expansions of datasets concerning specific categorizations have been discussed. Task 2 has basically started after finalization of the reports on each dataset from Task 1 in autumn 2014, and has been finalized in the end of the first reporting period (see Deliverable D6.2). All categorizations of the datasets have been screened, and potential suggestions for harmonization have been derived. These suggestions have been discussed at the annual RISIS week in January 2015, which then lead to respective harmonization’s decisions and choices for each dataset.

**Task 3** aimed to stimulate an exchange with all databases holders to insure robustness and potential for connectivity of technical aspects (see checklist Deliverable D6.1a). In this sense, technical aspects listed in the reports (see Deliverable D6.1) on each dataset have been screened and technical solutions for potential connections will be derived. Here, a close interaction with WP8, working on the establishment of an organization registry is crucial. The registry may constitute the backbone for connectivity. Task 3 has also lead to the establishment of a website to collect metadata on the datasets in a harmonized way which will facilitate connectivity (http://sms.risis.eu/).

**Task 4** aimed to coordinate work on principles and conditions of legal access of the datasets as well as obligations for users. Also here the reports produced in Task 1 have provided the basis to screen in a first step the legal access conditions for each dataset. However, the mains essence was the development of an accreditation system for access to the datasets. The task has been finalized in the first reporting period (see Deliverable D6.3)
**Task 5** was a smaller step, dealing with issues to potentially rebuilt access to two datasets (SIPER and ETER) that will be accessible at a distance. Access conditions for the online datasets have been clarified and documented (see Task 1 reports on ETER and SIPER).

**Task 1 – Task 5** have been completed during the first reporting period. This has led to the production of Deliverables D6.1, D6.1a, D6.2 and D6.3 (Deliverables are attached as amendments to this report).

**Task 6** will conclude WP6 with a second report for each dataset containing full documentation taking into account the changes introduced through the collective work, including all updates, new access conditions, harmonization’s, etc. Furthermore, the second reports will in some more detail explain conditions of ‘relevant’ use. Task 6 has started in the first reporting period and will be finished in the second period with updated dataset reports.

### 3 – Activities engaged and participants

#### 3.1. Task 1: Reports on datasets (activities 6.1.1 – 6.1.9)

**Partners engaged:** AIT (coordination) UPEMLV, VU A, UNIMAN, CNR, NIFU, UL, POLIMI, IFQ, USI

Task 1 has been the first major step in the work of WP6 (see Figure 1), providing the basis for all subsequent opening activities, enabling the screening of content, substantive issues, categorizations as well as technical peculiarities and legal conditions for opening. The production of the Task 1 reports has involved 9 partner organizations that are owners of the 10 datasets under consideration. The partners engaged in this activity, along with their respective dataset are as follows:

- **AIT:** ‘EUPRO’ (research projects funded by the EU Framework Programmes and all participating organisations)
- **CNR:** ‘JOREP – Joint and Open Programmes’ (European trans-national joint cooperation programmes)
- **IFQ:** ‘Profile’ (long term monitoring of researcher careers in Germany)
- **NIFU:** ‘MORE’ (largest survey of career path and international mobility in EU27)
- **POLIMI:** ‘VICO’ (dataset focused on new high tech companies in Europe)
- **UL:** ‘Leiden Ranking of Universities’ (data from the Leiden ranking plus a more detailed level)
- **UNIMAN:** ‘SIPER’ (repository of research and innovation policy evaluations)
- **UPEMLV:** ‘Corporate Invention Board (CIB)’ (CIB)(invention activities of the 2000 world largest R&D performers) and dataset on ‘Nanotechnology Dynamics (NANO)’ (Nano) (publications and patents dealing with the production of nanotechnology)
- **USI:** ‘ETER’ (full dataset on European Universities with key indicators)

The main essence for the production and further usage of the reports on each dataset has been a harmonised and clear structure for the reports. The structure has been suggested by AIT as coordinator of WP6 and discussed among all partners during the kick-off meeting in Paris, 3-4 February 2014. The resulting structure has been used by all partners for the production of the reports.

The resulting 10 reports on content and technical structure of the datasets under consideration have been finally submitted – after two internal review rounds– to the Facility Coordination Board (FCB) of RISIS by AIT in November 2014. The basic structure and all reports are presented in Deliverable D6.1. They are available under [http://risis.eu/publications/](http://risis.eu/publications/).
As an important side activity of the Task 1 reports, all partners engaged in this task have produced a poster for the 19th international conference on Science and Technology Indicators (STI) in Leiden, Netherlands, 3-5 September 2014 (with the exception of University of Manchester, since their database SIPER has not been sufficiently developed at that time). The respective posters that have been produced by the partners and presented in a well-received poster session are to be downloaded at the RISIS website (http://risis.eu/data/).

3.2. Task 2: Coordination of the work on the categorisations used in RISIS facilities (activity 6.2)
Partners engaged: AIT (coordination) UPEMLV, VUA, UNIMAN, CNR, NIFU, UL, POLIMI, IFQ, USI
The objective of Task 2 was to coordinate the work on the different categorizations used in the 10 datasets. By categorizations, we refer mainly to sectorial classifications, such as economic or technological categories defined by a statistical authority or by own efforts. Under the frame of Task 2, one main activity was to stimulate a collective discussion among all WP6 partners on how the categorizations of the datasets under study may be harmonized in a reasonable way. Task 2 has instantly started after finalization of the reports on each dataset from Task 1 in autumn 2014 and has been finalized in spring 2015. The detailed work and outcomes are given in Deliverable D6.2.

3.3. Task 3: Insure robustness and potential for connectivity of technical aspects (part of activity 6.6)
Partners engaged: VUA (coordination), AIT, UPEMLV, USI
This task builds on the previous tasks as it uses the descriptions of the datasets produced in Task 1. The outcome of task 3 will be the establishment of a platform to facilitate the connectivity of the datasets in terms of semantic web technologies. The platform has been established under http://sms.risis.eu/. In the course of Task 3, following steps have been conducted:

I. The reports on the datasets have been scrutinized from the perspective of connectivity in terms of semantic web technologies – which is much more flexible than a database perspective. In order to answer the question of connectivity, for all the datasets in RISIS an inventory has been made of the objects (entities) and the attributes of the objects. This defines the connectivity options.

II. Not only the type of objects in the databases need to be identified, but also the domain. E.g., if datasets have research organizations as entities, than it is important to know to what extent the sets of covered organizations overlap. The less they do, the less de facto connectivity is possible.

III. The vocabularies used in the different datasets have been described and compared, in order to make clear what objects and attributes in the different databases are similar or not.

IV. Using semantic web technologies, the main requirement is that data is converted to a Resource Description Framework (RDF) format. However, in order to connect databases in practice, a user needs to use a series of relevant API calls (queries to the different datasets).

V. The level of access to the data is described as this determines the de facto connectivity – in contrast to the theoretical connectivity. The level of access leads to a classification of the APIs on different levels of access – depending on the users’ access rights.

3.4. Task 4: Coordination of principles and conditions of access (part of activity 6.4)
Partners engaged: UL (coordination), AIT, UPEMLV
The objective of this specific task was to develop principles and conditions of access as well as obligations for users, so as to develop a harmonised set of options, which will cover all databases. The main outcome was a central accreditation protocol for any researcher to access the RISIS infrastructure. Task 4 has been finalized in the end of the reporting period, with the establishment of the RISIS registration website. Detailed information is given in Deliverables D6.1a and D6.3.
3.5. Task 4 cross-cutting activity 6.6
Partners engaged: AIT (coordination), UPEMLV, USI, VUA
This activity bundles several smaller activities that are necessary for the preparation of the opening of the RISIS datasets. This includes:

- Development of a checklist for opening
- Building the registrations portal for user
- Formulating a selection procedure for projects
- Site visits, possibly with trial runs and supported by checklist, with subsequent reporting and assessment

Detailed information on these activities is given in Deliverable D6.1. The activity will be continued in the second reporting period.

4 – Outputs

Table 1: Overview of WP6 outputs and Deliverables first reporting period

<table>
<thead>
<tr>
<th>Task/Deliverable</th>
<th>Output</th>
<th>Month</th>
<th>Lead partner</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task 1</td>
<td>10 harmonized reports on the content and technical structure of the datasets under consideration in WP6</td>
<td>November 14</td>
<td>AIT</td>
</tr>
<tr>
<td>Task 1</td>
<td>9 posters and poster presentations on each dataset, present on RISIS website and during the STI conference in Leiden</td>
<td>September 14</td>
<td>AIT</td>
</tr>
<tr>
<td>Task 2</td>
<td>Screening of existent categorizations in the datasets; implementation of harmonization strategy</td>
<td>May 15</td>
<td>AIT</td>
</tr>
<tr>
<td>Task 3</td>
<td>Screening of all technical conditions and production of a technical checklist and a platform to facilitate connectivity (<a href="http://sms.risis.eu/">http://sms.risis.eu/</a>)</td>
<td>May 15</td>
<td>VUA</td>
</tr>
<tr>
<td>Task 4</td>
<td>Screening of all legal conditions and report</td>
<td>February 15</td>
<td>UL</td>
</tr>
<tr>
<td>Task 5</td>
<td>Clarification of access to online datasets (SIPER and ETER, has been clarified in Task 1 reports)</td>
<td>November 14</td>
<td>AIT</td>
</tr>
<tr>
<td>Deliverable D6.1</td>
<td>Report on technical conditions &amp; harmonisation of datasets</td>
<td>May 15</td>
<td>AIT</td>
</tr>
<tr>
<td>Deliverable D6.1a</td>
<td>Report on activities to conduct opening</td>
<td>May 15</td>
<td>AIT</td>
</tr>
<tr>
<td>Deliverable D6.2</td>
<td>Report on the harmonisation of categories</td>
<td>May 15</td>
<td>AIT</td>
</tr>
<tr>
<td>Deliverable D6.3</td>
<td>Report on conditions of access</td>
<td>May 15</td>
<td>UL</td>
</tr>
</tbody>
</table>
5 – Challenges

WP6 is one of the core work packages with respect to the establishment of a joint data infrastructure on science and technology studies which is the overall objective of the RISIS project. It includes a variety of partner organizations holding distinct datasets. In this context, the main challenge will be to coordinate different partner rationalities and different legal conditions of each dataset under consideration in order to finally produce an accessible and rather harmonized novel infrastructure. The main challenges are to be summarized as follows:

- Coordination of different rationalities among WP6 partners
- Opening of datasets to researchers taking into account respective legal conditions
- Establishment of a joint and agreed upon accreditation system for the usage of the datasets
- Ensuring potentials for connectivity between the datasets, given different data cleaning strategies, classification systems etc. (close interaction with WP8 in establishing an organisation registry)

6 – Program for next year and milestones

According to the overall work plan of WP6, the main elements of the programme for the next year will be the further coordination of the opening process and the conduction of Task 6 as given in the DoW. After the finalization of Task 1 – Task 5 opening of datasets can be started. However, there is still a lot of coordination work to be done during the course of opening, which will mainly be a WP6 task. These activities are the following:

- Organization of site visits to each dataset to be opened. This will be mainly done between July 2015 and July 2016. The site visits are intended to ensure a smooth opening of a dataset in joint discussions between the visiting team (FCB, VUA and AIT) and the dataset holders. It will be jointly considered whether the basic conditions for opening are fulfilled, and potentially raise issues that still need to be tackled before opening.
- Overseeing and coordinating the further development of the RISIS registration approach, based on feedback of incoming researchers and first experiences with visits. This also tackles for instance the issue on how to deal with combined datasets, or datasets that are coming into RISIS from the outside.

Table 2: Milestones of WP6 work for the next reporting period

<table>
<thead>
<tr>
<th>Task</th>
<th>Output</th>
<th>Lead partner</th>
</tr>
</thead>
</table>
| **Further coordination of opening** | - Contact point for all WP6 dataset holders in terms of opening  
- Organization of site visits  
- Coordination and synthesizing conclusions from experiences of first researcher visits | AIT          |
| **Task 6**                      | - Each dataset holder will produce a second report  
- The second report contains all updates after harmonization activities and full documentation | AIT          |
RISIS incorporates two platforms that are complementary in their objectives: CORTEXT aims at helping researchers to enrich, treat and visualise the larges textual corpuses they have built, often from heterogeneous sources; while SMS focuses on the ability for researchers to build new corpuses originating from public data, and mostly from the web. The latter is often confronted to mixing this public data with private ones (which can even be confidential), thus the need for a complex structure enabling such constructs.

The two platforms also differ in their state of affairs. CORTEXT has been developed thanks to previous French public engagements and was already in a Beta version when RISIS started, the objective for wide opening being to stabilise processes, insure quality, increase computer performance (both in term of speed and in term of capacity – the number of parallel users working on the platform) and enlarge the targeted population of researchers (by being more user friendly and in particular requiring less computer knowledge). The objective was thus to open as soon as possible the Beta version widely as a service (which has been done well in advance on the schedule) and to open the “RISIS version” for the summer 2015 (It will open in October 2015). Deliverable 71 accounts for all the choices made and the state of development at the end of 2014, and we have just here an account of the work done since (first term 2015).

SMS on the contrary is a new venture, which requires important developments, including the definition of its architecture. We thus have an extensive report presenting how the platform is, the on-going developments and the plans for the next phase. The ambition is to open it by mid 2016 for in-depth testing on topics of interest for members of the network, and through on-site visits. The wider opening is expected in 2017. The reader will find in annex to this document a case study about what can be expected from the SMS platform in the future.

1 – CORTEXT progress report

The work in this period was about preparing the opening of Risis-eu CorTexT, including the provision of a full description. More specifically, it addressed two issues:

- the design of an ergonomics user-oriented interface to access the scripts to be run on datasets;
- the prototyping of the “Risis-eu CorTexT” Facility to be open at M12 in a beta accessible version;

Phases of work during 2015 (Jan./June)

- Active participation to the RISIS week in Roma and in the June Governing Board (see proposal for a RISIS overall architecture)
- Development of the back-office of the interface (server, auth., architecture) thanks to regular meetings through Distant Communication Systems and occasional visits
- Integration of Scripts for Data Analysis in the architecture

Phase 6: the design of an ergonomics user-oriented interface to access the scripts to be run on datasets;

Design and integration of a Graphic Interface by subcontracting with a Designer

Detailed Report of IT Activities

a) Setting the Resources to ensure the feasibility of IT Development for the RISIS.Cortext Platform

- Development of an IT environment for coders (Cortext Box) in order to receive codes and realize tests with a “plug-and-play” accessibility
- Preparation and installation of 3 servers for data location and calculation resources
- Migration of some CorTexT Manager scripts and interface components to the installed Servers
- Securisation of installation devices while freezing the version of modules in a LTS (Long Term Support)
- Protocol of the Symfony framework.
Design and Implementation of a Software Architecture for the central system of jobs management in order to enable the use of the CorTexT Manager existing Libraries in the RISIS.Cortex infrastructure of calculation

b) Designing and implementing the environment to install, control and align Scripts
- Migration of the version of ZMQ to manage the dialogue interface between the modules of the workers at the Server Level and the automatisation of ZMQ installation
- Installation and configuration of Docker Technology in order to encapsulate scripts and to gain interoperability and more security on scripts
- Designing within Docker a controlled procedure of installation and of automatic tests of interdependency between scripts
- Alignment of scripts and logs manager within the Silex Framework and creation of a central Risis.Cortext repository

c) Developing a capacity to check, control and follow-up scripts and jobs in order to monitor the development of the software
- Uniformisation of logs reporting system and designing the logs visualisation interface
- Developing the resource to follow up the jobs at the use level
- Developing and Interface of Jobs Monitoring
- Creation of Users Fields
- Implementing solution to improve accessibility to scripts and files in the “Start Script” menu of Risis.Cortext
- Implementing the possibility to access the status of running jobs
- Improvement of the collaborative feature of the User interface (adding capacities for the previzualisation of output files)

d) Producing a Beta-Version of RISIS.Cortex
- Improving the process of testing the Risis.Cortext architecture
- Bugs correction
- Consolidation and normalization of the architecture
- First tests of interoperability with SMS
- Installation of the RISIS.Cortext platform

e) Graphic Design of Risis.Cortext Facility
- Benchmark of existing Infrastructure Graphics
- Design of a first proposal for the Graphic Design of the RISIS.Cortext Platform (Design by a subcontractors)

f) Subcontracting (17 days by Chloé Duloquin, cost 5 929 euros)
- Benchmark of various Digital Humanities Platforms
- Proposition of a Visual Identity
- Edition of technical specification of the Graphic Chart of the RISIS-Cortext Interface
- Creation of a template of the pages under fireworks

CORTEX planned forthcoming activities
Phase 7: the final prototyping of the “Risis-eu CorTexT” Facility
- Hackathon in Paris, 3 days for All Team of the SMS and CorTexT Platforms
- Edition of Technical Specification
- Phase 8: Opening access to CorTexT platform for October 2015
- Preparation of the Training Session in October 2015
2 – SMS progress report

2.1. SMS platform: a non-technical description of the idea and of progress

Quite some work was done on bringing SMS in a shape enabling use by researchers in the field - in order to realize the opening date in M26. The following steps were undertaken. After the kick-off, the RISIS project went into a direction of stand-alone datasets, and the idea of platforms was seen as merely supporting tools for analyzing these stand-alone datasets. Gradually, the RISIS project moved to an infrastructure focusing on data harmonization and integration – as that enables researchers to ask new and more complex questions. SMS aims at opening in February/March 2016 and for the moment there are no delays foreseen.

The aim of the platform is to provide data integration support and data analysis support, for science and innovation studies researchers, using semantic web technologies. The platform has a high level of flexibility, but exploiting that requires rather high-level computational skills from the user. Therefore, quite some effort is devoted to optimize the usability for social science researchers. Despite such efforts, one should expect that support by data scientists remains necessary, and this is the rationale behind offering the possibility to visit the infrastructure (and the related data scientists) as a service to researchers.

The SMS platform is now being reshaped into an architecture (Fig 1) based on the best practices from a successful example: the OpenPhacts system (awarded with the 1st prize in the 2015 Linked Data Applications competition of the Semantics.CC conference in Vienna). It consists of the following components, all of which will be built as services that can be used in a variety of end-user interfaces:

(i) Tools for integrating datasets, those within RISIS but also a variety of other linked open data (e.g. Wikipedia, Geodata), and proprietary data (e.g. Web of Science).
(ii) Metadata system, enabling users to search through the available data, but also supporting data integration. The RISIS website (WP4) is used as the delivery platform for the metadata system, making the website an integrated part of the infrastructure.
(iii) Entity recognition, needed for the identification of the objects of research in the data: organizations, people, grants, projects, cities, etc.
    Entity recognition also has a major role in disambiguation and in data enrichment.
(iv) Identity resolution (disambiguation). This is an essential feature to improve the quality of datasets, and especially for the integration of datasets at the level of the entities. Although the work on this has started, it is an important RISIS task (with WP25) for the second half of the project.
(v) Category services (or vocabulary linking services), in order to map at a basic level different ways characteristics of entities are measured, such as field of science, type of organizations, grant types, or geo-location; we will use the existing category systems, as well as the existing mappings.
(vi) Provenance data to enable researcher assessing the origin of data and what has been done with the data. This will support quality control, one of the challenges in WP25.
(vii) A set of APIs for advanced users to extract data from the platform. A set of queries for researchers with relatively low level computer skills to get often required data from the system, in a format needed for analysis or visualization.
(viii) An export facility to transform data to a format that can be imported by analytical tools for data analysis and visualization.
(ix) Several facilities for access control, in order to optimize accessibility and at the same time protecting non-open data.
(x) A service for domain adaptation which will adapt other SMS services (in particular Named Entity Recognition) to a specific domain of interest. For instance to only detect organization of type Public Research Organization when performing entity lookup.
(xi) A mechanism to manage similar or identical entities in our platform. Different identifiers might be provided for the same entities by different RISIS datasets. Therefore, it is a
crucial feature in SMS platform to keep record of identifiers for doing the data integration tasks.

Figure 1: the basic architecture

2.2. Work done in the first period

The major tasks for VUA in WP7 are documenting the SMS platform for users (task 2), streamlining the workflow for users (task 3) and supporting the inclusion of heterogeneous data (task 4). In the selected architecture, these three tasks are highly connected. In order to realize these tasks, several activities have been undertaken in the first 18 months of the project by the SMS team:

(a) Basic decisions about the approach:
- Reusing and adapting technologies developed at the VUA – especially the OpenPhacts architecture (www.openphacts.org). In order to do this, the current team involves OpenPhacts developers;
- Initially, the decision to have the platform in a version for online use and a version for local use, as the latter makes the inclusion of non-open data easier by individual data-owners who often hesitate to store those data in a central facility. However, within the consortium the awareness has grown that central facilities may be more powerful, and that data protection can be organized also at a central level.

(b) Implementing various components of the platform. Most of the core parts of the architecture are now available, and we foresee the test version by the end of 2015, which enables opening for visitors early 2016. It will not the final version, as for several components more
development work is needed. E.g., entity resolution (disambiguation) remains a crucial but also difficult issue.

(c) Converting relevant datasets, including RISIS data, into the general Resource Description Framework (RDF) format - the Linked Open Data standard.

(d) Searching available (Linked Open) Data that are relevant to researchers in the field and therefore worthwhile to integrate. Examples are Wikipedia data, geo-information, OpenStreetMaps, OrgRef, OpenAire, OpenCorporate, Fundref, Cordis, BASE.

(e) Integrating heterogeneous data strongly depends on entity recognition and resolution; and this also supports data enrichment: e.g., coupling ETER with DBpedia leads to much richer data about the HE institutions – which also may support more qualitative research designs.

(f) Developing a demonstrator that can show the potential of the platform to researchers in the field (see the annex).

(g) Managing the code repository of RISIS project. We have created an organization Github (https://github.com/risis-eu) which shares the code for RISIS open platforms. The code for RISIS datasets portal, SMS platform, APIs, API applications and cortex geoclustering algorithm is already hosted in there. This infrastructure helps for collaboration between developers on RISIS.

2.3. A non-technical description of the platform

The basic idea is of the infrastructure to support the whole data integration and analysis workflow. We now describe those functions, of which a part has been implemented already, and others are under development or being implemented currently. This is a description of a full-scale version of the SMS infrastructure, which also contains elements that we will not be able to (fully) develop during RISIS project. Prioritization will be done among others in consultation and coordination with users.

(i) Firstly, heterogeneous datasets have to be translated into RDF format, in order to link them. This requires often some preprocessing. For example, data may be in PDF format that has to be
transformed into a format that enables further processing. The platform will contain preprocessing tools for the most common data formats. Figure 3 gives a few examples.
* PDF is converted into TXT, and then the named entities (e.g. organizations, municipalities, people) are identified (see below step ii).
* Also other preprocessing can be done, like identifying words and phrases for textual analysis. Then the data are converted into RDF.
* Web of Science records are parsed, and then converted into RDF
* Database and spreadsheets are directly converted into RDF, such as RISIS datasets ETER or Leiden Ranking.

![Preprocessing Diagram](image)

Figure 3: Preprocessing

(ii) When in an accessible format, the named entity recognition service will identify the relevant entities in the data (people, organizations, projects, etc.) in order to link those between the datasets, using knowledge about those entities in open data. The RDF format makes a large connected graph of the available data, using the different entities as nodes (figure 4 for an example). This is crucial for data-integration, as increasingly Linked Open Data are available that can be mobilized for many tasks:
- Enriching the data
- Entity recognition: what is in the data?
- Disambiguation, by using background information about entities.

(iii) The aim is to integrate heterogeneous data. As Figure 1 shows, various types of data are distinguished:
Linked open data ‘out there’;
The RISIS open datasets, such as ETER, or JOREP
The RISIS non-open sets
Proprietary data such as Web of Science
Confidential datasets, containing e.g., privacy sensitive data

(iv) The platform is based on a set of data access methods, a data cache, and a set of integration tools that couple the services to the data. The design will allow users with private and confidential data to integrate the platform for themselves to deploy the functionality with their own data. These data will get the same format, but an access control service will regulate who can access what data. E.g., if a user has legal access to WoS data, this may enable to use those data through the SMS system, and if a user has not, this is blocked. The further refinement of the access control service will be coordinated with the Cortext work on protected virtual workspaces.

Figure 4: Linked data example

(v) Within and between datasets, entities may have different names, and different entities may have the same name. An identity resolution (disambiguation) service will be integrated in the platform to support the task of disambiguating as good as possible. Within the RISIS project, the development of these tools is an R&D task in WP25. Work is progressing but there is still quite some work to do. Disambiguation remains a core challenge, and this will remain a labor-intensive task, which we aim to support as good as possible. The philosophy is not to develop new tools for disambiguation, but to optimize the disambiguating process by combining existing tools and techniques. The work-package has started by scanning various existing techniques. A few of these are currently tested; but the testing is not yet finished:
* PSL, a disambiguation system based on probabilistic soft logic.
* The system developed by (Gurney et al 2012) and that seems to work well in smaller document sets, in combination with the related system developed by Sandström that works well on large sets, but then needs further refinement. Combining the two may optimize their different strengths: First applying the Sandström system on large corpora, and then for each individual author the Gurney system for further refinement.
* The Agdistis project, as it combines named entity recognition with disambiguation.

(vi) Disambiguation and named entity recognition are of course greatly supported by registers of the main entities relevant for our domain, such as research universities, other higher education
institutions, public research organizations, research intensive companies, or researchers (Orchid). Partly, registers are open available, such as the ETER dataset. Others are private (such as the ORBIS company database) or under development (such as a register for PROs). As increasingly open resources for creating those registers become available, the platform aims developing a service to generate those registers out of open data. We are collecting open data to do this (figure 5). This is part of WP8, where of course also conceptual issues have to be resolved.

Figure 5: Linked Open Data Cloud (top). Some open datasets ingested to our platform (bottom)

For integrating purposes, but also for supporting the users, the platform extracts metadata, and supports data-owners to enter the relevant metadata (Figure 6). The metadata system has been created in a way that it not only functions as search facility through the metadata as provided by the RISIS dataset owners, it adds to this function by relating the metadata to other knowledge about the entities that are covered by the datasets. Technically, this goes first through entity
recognition in the datasets, and second through coupling these entities to information that is “out there” in linked open data – such as DBpedia.

Figure 6: The metadata entry site (top)
Classification schemas as part of the metadata (bottom)

The metadata system also supports the coupling and integration of data, by using uniform entity descriptions and standard attribute definitions made available on the web. This is done by (1) reusing attribute definitions that fit our data modeling rationale, (2) generating new and specific attribute definitions that did not exist and (3) making the added definitions open, reusable and accessible on the web. In addition the metadata was categorized to accommodate different aspects of data integration within RISIS datasets.

The system goes beyond the RISIS datasets, as we will allow others to enter metadata into the system, and that enlarges the scope of data that can be integrated and used through the RISIS platform. More particularly, in the context of the SMS platform, we will do this for relevant open (and less open) data sets, and for attribute definitions.
A metadata search facility supports the user to identify those data needed for his/her study (figure 7). The RISIS dataset portal also allows for registration and access requests. While working on the website it gradually evolved into an element of the infrastructure itself, as the advanced metadata system has been integrated in the website, and the same will be the case for the access control.¹

(vii) Not only entity recognition is a major function of the platform, the same holds for the identification of the attributes of the entities. Entities may have similar properties in different datasets, but those properties may be classified and measured in different formats, such as science and technology fields, application domains, organization types, location, etc. For users that aim to deploy data from different datasets, it is important that they can use the same ‘category schemes’ for the different datasets. Data integration therefore also asks for category harmonization, and the category services provide this: existing category systems and their mappings. This is work done in relation with WP8 and WP9.

(viii) More specifically the platform will implement several basic geo-location services, as being developed within the context of WP9. Basic geo-services contain mainly the mapping of a variety of geo-localization system upon each other. We implemented e.g., a system that goes from an organization name or an address, to a standardized address and coordinates, and from that to NUTS-categories and from that to the list of municipalities in that NUTS-area (figure 9). The tool

¹ The website of the project has several functions:
- Providing general and specific information about the project and its activities.
- Providing potential users with detailed information about RISIS datasets available, about the platforms and other facilities available in the RISIS infrastructure, as well as relevant data ‘out there’.
- Registration of RISIS infrastructure users, as we require researchers that want to use data and the facilities to register.
- Requesting access to the RISIS data when online available, and apply for a RISIS-supported visit to one of the facilities. Connecting to RISIS protected data use environments.
- Monitor use of the RISIS infrastructure.
enables to do this visually, but of course underlying is the possibility to do it for datasets that then can be used for analysis. As research needs often needs classifications on a more detailed level, or classifications used by policy makers, we investigate whether it is possible to approximate the OECD Functional Urban Areas, using string matching techniques and using publicly accessible shape files.

![Transportation Time and Cost](image)

**Transportation Time and Cost**

Figure 8: advanced geo-concepts

(ix) More *innovative geo-services* will include other distance metrics, such as:
- Distance in terms of intensity, costs, and duration of travel (figure 8)
- Distance in terms of similarity in socio-economic and cultural characteristics.

This work will start in WP9 when the basic services are implemented.
The platform is a complex support tool. Experience with OpenPhacts has shown that it is too difficult for most users to access the data store directly, even for bioinformatics researchers. Next to that, direct access would also create possible overload problems for the system, given the size of some of the databases involved. OpenPhacts experience showed that the optimal solution is to develop a series of ‘standard’ queries as these satisfy much of the data demand. These queries were derived from intensive sessions with researchers from the field.
Also in SMS, we therefore will support the users with a considerable set of services built on top of predefined queries that address the typical types of research questions in the domain. A typical example of a query is in figure 10.

Data will be exported in a format enabling analysis and visualization. Some own visualization tools are available. But to a large extent, the data will be analyzed using existing analytical tools, such as statistical packages (SPSS, Stata, SAS, R), and network analysis and visualization tools (like Gephi). For some analytical platforms, interfaces will be developed in collaboration with those analytical platforms (Cortex).

Find all grant names and panels for applicants from Netherlands who are affiliated with VU University Amsterdam.

![Query example](image)

**Figure 10: Query example**

(xii) Interaction with potential users took place (e.g., during the RISIS week) to find out about the common data needs. It became clear that this remains too abstract for many STI researchers. Therefore we decided to implement some ‘demonstrators’ to show the possibilities of the platform to users. This may lead to a clearer formulation of data needs. Several demonstrators – real research projects - are under exploration and development:

- A demonstrator on grant selection research, which we plan to finish by the end of 2015.
- A demonstrator on research and innovation collaboration of large companies foreseen for 2016.

Demonstrators were selected as they reflect together the complex situation a researcher may be confronted with. In case of the first demonstrator: ETER, Leiden Ranking (both RISIS), Cordis, Crunchbase, DBpedia (all open), WoS (license based), Grant Applications data, review reports and CVs (all confidential and therefore closed data). See the ANNEX for an example.

(xii) For more advanced users, APIs\(^2\) will be provided to export the data in various formants

(xiii) Quality control: The opening of datasets asks for bringing those on a high level of quality and transparency – a big task for the owners of the individual datasets, but even more for platforms

---

\(^2\) API = Application Programming Interface
supporting the integration. That asks for even higher levels of harmonization and quality control. In order to support quality control, a checklist for the opening of datasets was developed (with AIT), describing a series of steps to be undertaken before making the datasets available for users. A main instrument for checking the data quality was provided by the developed metadata system (see above): by completing the metadata, the data owner runs through a set of steps that help improving data quality, transparency, and usability. The checklist was based on a state of the art overview of data harmonization and quality considerations and techniques (deliverable D6.1). In WP25 further steps will be done towards improving quality control, using again the linked nature of the data, as well as using provenance approaches (with La Sapienza).

2.4. SMS, a wish list for the next period

We plan to continue work on a variety of issues, all aiming at the further implementation and improvement of the SMS platform. The list below should be seen as a wish list that cannot be fully realized. What we will select from the list and how far we will come needs further discussion with partners in the project and with the users (researchers in the field), and depends on the resources we can mobilize. We do not consider this as problematic, as the development of research infrastructures is a long-term activity in which we cannot expect to have everything ready in a relatively short period defined by the RISIS project. Leading the work is that the platform can be used in a meaningful way by the researchers.

- RISIS website:
  Further development of the website as not only a portal for the RISIS data and services, but as an integrated part of the RISIS infrastructure that supports scholarly work with integrated data and analytical tools, and at the same time safeguarding the non-open and sensitive nature of part of the data that can only be accessed with specific permission. So access control and data protection will be important tasks in the further development of the RISIS website. As the website is integrated in the infrastructure, we decided not to outsource the production and maintenance of the website, but doing this in-house too. That seems a more efficient way of using the resources.

- Finalizing the first release of the SMS platform
  - Finalizing the implementation of the platform.
  - Finalizing the description for the users.
  - Opening the platform in M26
  - Extending the RISIS datasets with external dataset. We will apart from including already open data – also approach data-owners of other datasets to add their data to the platform, e.g., data on careers and mobility of researchers.
  - Coupling the platform with analytical tools (CorTexT network analytical tools; SMS/Saint bibliometric analysis tools).
  - Adding the category services
  - Extending the set of APIs, based on (potential) user consultation.
  - Access regulation
  - Adding the results of WP8 and WP9:
    - Organization registers
    - Basic geo-services
    - Innovative geo-services

- Further development of the platform will be done after the previous actions have been finalized:
  - Adding entity recognition (with WP25): We will further develop and extend the activities on disambiguation in the second half of the project. The goal remains to test combinations of disambiguation tools to determine their effectiveness in relation to the

---

disambiguation task at hand. In this way we plan to develop a disambiguation toolbox that can be applied in a workflow of large-scale data integration and harmonization.

- Adding quality control tools – provenance based (with WP25): This task will start with a workshop to discuss and select the existing quality management and provenance systems. Based on that we will produce a state of the art overview of provenance technologies to support harmonization within and integration between data sets. Again the philosophy is: what is available out there, what can be reused, and what should be developed. When that is done, we will experiment with provenance approaches for RISIS.

- As bibliometric data are crucial in most of the research in our field, we started to explore how this can be organized and implemented within the SMS platform. The conversion of WoS records into RDF can successfully be done. We envisage to have two different interfaces with publication and citation data: through an existing tool for analysing smaller sets of bibliometric records, and through another existing tool for doing analysis on the large ('complete') bibliometric corpora. Especially the latter is of crucial importance, given that indicators more and more need to be (field) normalized, which requires access to whole corpora.

- As one needs bibliometric indicators are needed for analysis and not the raw data, we explored how to integrate tools for deriving advanced bibliometric indicators in the platform. For this, collaboration with prof Ulf Sandström (KTH Stockholm & Orebro University) has been established. Access to these kinds of proprietary data remains of course restricted to those with the correct rights.

- Showing how the platform can be used in one or more demonstrators, in order to support users.

- Training and support material (With WP5).

- Adding more data to the platform.
Several datasets are already included, such as a few confidential datasets of the GendERC project, which will be used for the SMS demonstrator, WoS data needed for that project, several RISIS datasets (ETER, Leiden Ranking), and a variety of open data. See figures 4 above and figure 3 of the ANNEX. Ideally, also the other RISIS datasets could be included in the system.
PART TWO
ACTIONS ACCOMPANYING DATASET OPENING

The central report has already presented the key results we derived from the ENID conference (WP4) (a full report about the Leiden ENID conference is available). We thus only present here complementary information on training activities (WP5) and about the two core harmonisation processes that will guide RISIS integration of datasets: organisation based harmonisation (WP8) and geography-based harmonisation (WP9).

The report for WP8 does not correspond to what we expected; for two central reasons. We have divided the work in two subsections: firms and public sector research. For firms we have found it more logical that the effort is combined with the efforts in WP20 (efforts made are thus presented in WP20 report). For public research organisations the principle of building one or more registers has progressively emerged. In order to get a clear view of the constraints of a register we have asked people working on ETER to propose us the functionalities of what a register should do by fully explaining this through the deployment of ETER, this will already address one central issue of public sector research: universities (see report below). It still leaves unaddressed issues of alliances and mixes (with PROs and with hospitals). This will be the objective of the work during the second part of 2015 so that options are proposed for discussion in the RISIS week in Amsterdam. But the major issue we face deals with PROs, for which the review made has shown that resources available were fragmented, partial and with high level of inconsistencies, and that each dataset has developed ad-hoc solutions. So a meeting has been organised (August 28, 2015) to define a strategy and an operational approach that will nurture the discussion in the RISIS week in Amsterdam (January 2016).

1 – TRAINING ACTIVITIES

1 – Objectives of WP as of Description of Work

RISIS is a research infrastructure for increasing the researcher capacities, and for diffusing more widely knowledge about and issues on indicators, but it also have to have effects on the wider community, and on stakeholders.

The objective of the training can be summarized as follow:

• Raising the research skills needed for using large data. This objective can be pursued through courses dedicated to supply knowledge about:
  – Introduction to the use of the infrastructures (platforms and datasets), including the introduction into theoretical background
  – Indicators and data handling and indicators design
  – Statistic and econometric analysis, scientometric tools (sophisticated treatments methods)

• Supply a concrete evidence of the impact of RISIS on the scholars’ community. Training is one of the most important evidence of the importance of the Project and of the impact it has had.

Training activities developed within RISIS have to take into account the need to be integrated and harmonized with the training activities already developed within our research organizations and within other research networks.

RISIS training has specific characteristics that make it different from other existing courses, and with a specific value with respect to other courses supplied at European level.
We consider three main characteristics here, namely:

- **Target addressed**
  - Senior scientist, early career researchers and PhD students at the last phase of their PhD training (PhDs)
  - People from the policy making level (e.g. funding agencies)
  - Research intermediaries (e.g. research association like Science Europe)

- **Courses are generally associated with the use of one or more infrastructure -the latter solution is the best choice; training activities on methods must be contextualized and related to the datasets and platforms;**

- **Courses demonstrate the added value compared to existing courses, which means that there should be a focus on areas less covered in our field;**

- **Training is free for European participants**
  - No fees to be paid
  - Venue of the trainees (travel and accommodation) covered only in case of researchers, early researchers and PhDs coming from European countries
  - No costs are covered for people not involved in research activities (i.e. people from intermediaries or policy level).

RISIS has put in place three type of courses: Short course A (SCA), Short Courses B (SCB) and Summer Schools (SS).

**SCA are two days courses aimed at illustrating the content and the use of the datasets and the platforms. They are:**

- Focused on the presentation of the infrastructure (characteristics, potentialities, limitations, etc.)
- Accompanied by a full documentation of the infrastructure, slides and other relevant materials available on the RISIS website. These materials have to be prepared by each participant within the activities developed for opening the infrastructure.

SCA may adopt different types of organization, according to the characteristics of the dataset; for instance it is possible to group different datasets in one SCA, to have a half day introductory session, in which a general introduction on the subject that can be addressed through the dataset is provided, or to have a practical computer session proposed to the non specialized audience, including PhD students with different background.

SCA can also be organized as one-day courses dedicated to a specific community, when there is a case of many people interested and located in one site. This should NOT be the rule: courses are also aimed at having people interacting from different communities, sharing ideas and views for research, providing useful interactions between them.

**SCB are two days of training on a given subject as:**

a) The use of specific dataset(s) and indicators
b) Statistics, econometric and scientometric approaches using the dataset(s)
c) Constructing and treatment of heterogeneous data
d) New developments of indicators and data analysis (e.g. textual data treatment and web-based indicators).

SCB may foresee different levels of skill in the targeted audience; they must always be associated to the use of one or two datasets, and in no case they can provide basic teaching on quantitative techniques.

Some suggestions come from FCB, namely:

a) It would be good that courses dealing with semantic analysis in practice connected to CORTEXT are opened to European participants (from the second part of 2014 onwards)
b) UPEM can organise a course on the use of patents in connection wit the 4 datasets directly interested (CIB, Nano, EUPRO and VICO)
c) Further ideas deal with quality control of datasets, with the use of event history to study organisations.

Further suggestions are: positioning indicators, network analysis, web-based indicators, performance measurements.
Summer schools are one-week courses (starting on Monday lunch time and ending at Friday lunch time); SS are generally devoted to early career researchers or PhD student in an advanced phase of their PhD training, although senior researchers are not excluded. SS are a collective endeavour in RISIS: we foreseen a minimum 3 RISIS members with different disciplinary foci as main organisers in an integrated way. SS can cover:

- Thematic issues (internationalization, collaboration, performance assessment, mapping, etc.)
- Methodological issues (articulating and agglomerating the RISIS databases)
- Policy issues (use of data and indicators, for performance assessment, European integration, etc.)

2 – Work programme Description

A process for submission and assessment of the training proposals coming from the RISIS partners has been implemented. Proposals are submitted using dedicated formats available on the website. Up to 15 participants per course can be envisaged. We have different types of format according to the 3 types of courses

SCA: simple format with dates, places, infrastructure(s) involved;
SCB: more elaborated proposal with details about the content of the training and infrastructure(s) involved, an articulated programme, targeted audience (specific skills requested to the participants), facilities supplied to the attendees
SS: here the format requires a structured proposal, which must involve several RISIS participants (at least three), and must highlight the aim of the training and the relevance of the issue proposed for RISIS, a detailed programme and the organization of the School. A budget articulation is also expected.

Proposals are assessed on the base of the following criteria:

SCA and SCB:

a. Clarity of the proposal,
b. Presence of a number of trainers with adequate skills,
c. Focus on one specific training content to be exploited using the dataset(s) involved,
d. Precise identification of the targeted audience. Please consider that in principle scholars attending the courses are supposed to have basic knowledge about statistical tools. In case of courses based on advanced statistics, the audience shall have enough competences in order to benefit of the course.
e. Facilities provided to the participants
f. Relevance for the RISIS aims
g. Clear added value of the course with respect to the existing ones (see General features, section 2)
h. Presence of updated documentation to be supplied to the students (the documentation for SCA is supposed to be updated according to the changes of the datasets and platforms).

SS:

a. Clarity of the proposal,
b. At least three RISIS organizations involved,
c. Presence of a number of trainers with adequate skills,
d. Focus on one specific training content to be exploited using the dataset(s) involved,
e. Precise identification of the targeted audience. Please consider that in principle scholars attending the courses are supposed to have basic knowledge about statistical tools. In case of courses based on advanced statistics, the audience shall have enough competences in order to benefit of the course.
f. Facilities provided to the participants
g. Relevance for the RISIS aims
h. Clear added value of the course with respect to the existing ones (see General features 2)

i. Reliable budget articulation

j. Presence of updated documentation to be supplied to the students

The idea is not to have a traditional “selection process”, rather a constructive process of the Committee in charge of training management (see section 7) with the proposers, so that their proposals fit better with the RISIS approaches. According to the RISIS project, only 7 partners (UPEMLV, VUA, CNR, UL, POLIMI, IFQ, USI) can organize the courses; all partners can participate in specific lectures or sessions, and partners not directly involved can organize courses in collaboration with other RISIS partners.

Costs
RISIS project foresees the following costs for training:
Preparation and delivery: 1 PM for Short courses – 2 PM for Summer schools
Logistical costs for rooms, documentation etc: 2.000€ for Short courses – 3.000€ for Summer schools
Costs to cover the venue of the trainees (travel and subsistence): 7.500€ for the Short courses, 12.000€ for the Summer schools
In case of one-day course, only travel costs are covered.
RISIS allocates the money directly to the organization in charge of the activity.
The work-programme foresees at least 30 training activities during the 4 years of RISIS duration. The expectation is to have as a minimum:
• 1 Short course-A per year associated to the data sets/platforms (or combining data sets/platforms);
• 4 Summer Schools to be harmonized with the existing offer at the university level.

3 – Activities engaged and participants

A call for proposal is permanently open for SCA, SCB and SS on the RISIS website. Proposals are assessed four-times per year (15 March, 15 June, 15 September, 15 December). The first deadline was June 15th, 2014. The activities engaged in the considered period are the following:

• Coordination and management the process of submission and analysis of the RISIS training activities as well as the reporting activities of the RISIS partners, which develops courses.
• IRCRES CNR has been in charge for managing the whole training in close cooperation with the FCB.
• A Committee supported the whole activity. The composition is:
  o Kate Barker (MiOIR),
  o Giovanni Cerulli (IRCRES-CNR),
  o Sybille Hinze (IFQ),
  o Emanuela Reale (IRCRES-CNR),
  o Benedetto Lepori (USI) on behalf of the FCB
• IRCRES CNR through the Committee provided a preliminary assessment of the proposals, based on the criteria listed in the overall planning. FCB provided the final approval.
• The Committee had one face-to-face meeting on June 2014 in order to take decisions about promotion of the training activities, accompanying actions for people submitting ideas, and assessment of the proposals submitted. Assessment of the courses, and reporting activities has also been discussed. A summary of the decisions has been transmitted to the FCB on June 2014. After this first meeting the other interactions were carried out using Skype or e.mail in order to speed up the process. It was decided that physical meetings can be organized during the RISIS week, which takes place every year in January.
The management activity developed included:

- Preparing the formats for the training activities (proposals, applications and reporting)
- Circulation of the training offers
- Managing the selection process of the courses
- Collecting the short Final Report on each course
- Elaborate an online questionnaire for assessing the courses
- Repository of participants and materials
- Summary for the annual RISIS week

Website
A space on the RISIS website is dedicated to the WP Training, with all the information available. Furthermore information has been also circulated by mail. More specifically the website provides:

- Call permanently open with the indication of the deadlines for submitting (text of the call downloadable)
- Formats for RISIS partners for presenting proposals of training courses (downloadable Formats of the three types of courses: SCA, SCB, SS)
- List of courses open with information on aims, contents, deadline for application, etc. (downloadable documents for people that want to apply)
- List of courses done with information on aims, contents, materials, reporting and assessment
- Questionnaire to be filled by all the people participating to the Training courses
- Format for reporting activities at the end of the courses.

4 – Outputs

The outputs produced by the RISIS Training can be summarized using some data on courses developed, people attending the courses and the overall satisfaction of the participants, which come from the results of the questionnaires submitted after each course. Table 1 presents the RISIS courses developed in the period 1.01.2014/30.06.2014, and Table 2 presents the courses that have been positively assessed, and are planned for 2015-2016.

Table. 1 - RISIS courses (period 01/01/2014 – 30/06/2015)

<table>
<thead>
<tr>
<th>Type</th>
<th>Title</th>
<th>Venue</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>SCA</td>
<td>Joint and Open REsearch Programmes: features and potentialities of the JOREP database</td>
<td>Rome</td>
<td>November 27-28, 2014</td>
</tr>
<tr>
<td>SCB</td>
<td>Applications of multi-level models to research policy and higher education studies</td>
<td>Lugano</td>
<td>January 8-9, 2015</td>
</tr>
<tr>
<td>SCA</td>
<td>What we know about Venture Capital in Europe: exploring VICO</td>
<td>Milan</td>
<td>February 4-5, 2015</td>
</tr>
</tbody>
</table>
Table. 2 - RISIS courses (already planned for 2015-2016)

<table>
<thead>
<tr>
<th>Type</th>
<th>Title</th>
<th>Venue</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>SCA</td>
<td>Early career researchers: features and potentials of the doctoral candidates and doctorate holders panel study ProFile</td>
<td>Berlin</td>
<td>September 23-24, 2015</td>
</tr>
<tr>
<td>SCA</td>
<td>Characteristics of the ‘Corporate Invention Board’ (CIB) and ‘Nano’ patent databases and exploration of their potential use in S&amp;T studies</td>
<td>Noisy-le-Grand (Paris)</td>
<td>October 8-9, 2015</td>
</tr>
<tr>
<td>SCB</td>
<td>Social Network Analysis. Introduction to methods and applications to EUPRO database</td>
<td>Lugano</td>
<td>February 16-17-18, 2016</td>
</tr>
</tbody>
</table>

Table 3 presents the participants to the RISIS courses by country and course attended. Interestingly enough the courses attracted also researchers working in Government departments and in private associations.

Table. 3 - Participants to RISIS courses per country and course

<table>
<thead>
<tr>
<th>Country</th>
<th>JOREP course (Rome)</th>
<th>Multilevel course (Lugano)</th>
<th>VICO course (Milan)</th>
<th>ETER course (Rome)</th>
<th>ETER course (Oslo)</th>
<th>ETER course (Leiden)</th>
<th>ETER course (Wien)</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Austria</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>Belgium</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Colombia</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Estonia</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Finland</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>France</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>Hungary</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Iran</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Israel</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Italy</td>
<td>8</td>
<td>6</td>
<td>14</td>
<td>10</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>39</td>
</tr>
<tr>
<td>Netherlands</td>
<td>2</td>
<td>0</td>
<td>14</td>
<td>10</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>39</td>
</tr>
<tr>
<td>Norway</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>11</td>
</tr>
<tr>
<td>not available</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Portugal</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Spain</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>TOTAL</td>
<td>15</td>
<td>15</td>
<td>17</td>
<td>10</td>
<td>11</td>
<td>11</td>
<td>7</td>
<td>86</td>
</tr>
</tbody>
</table>

Figure 1 presents the overall satisfaction of the participants to the RISIS courses. All the courses reached very high scores of overall satisfaction, which is reported as a promising result for the forthcoming years of the RISIS project.
Participants were asked to report the best features of the courses and to point out the aspects to be improved. They mostly considered positively the practical sessions of the training concerning the concrete use of data, the expertise of the teaching staff and the organizational aspects of the courses. As aspects to be improved they suggested mainly to extend the duration of the courses in order to provide more room to practical sessions.

Fig. 2 presents the percentage of participants to the RISIS courses that do not belong to one of the RISIS partners; it is important to see how attractive the courses have been for the community beyond the partners of the RISIS Consortium.

Further materials of the Training activities (leaflets with the aim and objectives of the courses and the Final Report produced by each course) are available on the website.
5 – Challenges

SCA courses proved to be easy to organize, and well accepted by participants. A different solution is under experiment, with training provided directly in different locations. This solution will be assessed in the next year. Nonetheless the number of proposals received is still too low if compared with the work-programme. The number of proposals for this type of courses should increase as soon as the activities linked to the preparation of the infrastructures for opening will be completed. In this respect the need to have at least one SCA course for each infrastructure before the opening has been outlined during the RISIS week;

SCB courses did not receive enough attention from RISIS participants. This means that a strategy toward stimulating the initiatives and proposals from the RISIS community must be elaborated;

RISIS courses proved to be highly appreciated and attractive for people outside the participants’ organizations. A strategy toward a large involvement of people in the STI field should be prepared by the Committee and discussed with the FCB;

Summer Schools are objects to be more precisely defined by the Committee. The Committee should elaborate ideas about possible proposals for Summer School in order to press the RISIS participants to action.

6 – Program for next year and milestones

• To address the challenges identified
• To promote the development of SCA-SCB courses linked to the opening of the research infrastructures
• To promote Summer-Winter School related to the demonstrators of the combined use of data for the investigation of key research topics

2 – GEOGRAPHY BASED HARMONISATION

1 – Objectives

The WP9 has 6 main objectives:
1. Comparing different approaches to knowledge clustering at the European level:
   a. Understanding the approaches for geographical aggregation at the 2 RISIS datasets most concerned by this issue, Nano and EUPRO.
   b. Learning the OECD methodology developed in order to delineate Functional Urban Areas (FUA).
2. Propose a ‘layered’ strategy (adapted to the appropriate level of geographical aggregation), based either on the OECD method, the UPEM method, a third method, or on a combination of the different approaches.
3. Develop and implement the selected approach to the two primary RISIS datasets EUPRO and Nano. Examine other possible candidates for this purpose.
4. Using the selected method for different purposes, as testing number of dominant assumptions in the literature.
5. Developing the RISIS platforms in relation to the geographic dimension.
6. Considering whether the geography identified within RISIS datasets can be generalized.
2 – Work programme

WP9 aims at harmonization and integration of the datasets centred on geographical dimensions. The harmonization activity results from the adoption of possible methods and options: the new approach of clustering developed by the OECD; the spatial clustering method developed by UPEM; a third possible method; combination of the different methods.

The activities were grouped into four main categories:
1. Depicting and selecting preferable approaches to knowledge clustering;
2. Operationalizing and testing the selected approaches in RISIS datasets;
3. Proposing S&T map of Europe;
4. Reporting activities.

“Depicting and selecting preferable approaches to knowledge clustering” (see below the work plan suggested) is dedicated to preparatory activities, concentrated on choosing a method or methods of spatial clustering which will be applied to the RISIS datasets, and will be later integrated at the project’s platforms (Deliverable D9.1). It is a prerequisite input for the “Operationalizing and testing of the selected approaches in RISIS datasets” category (see below the work plan suggested). This category is devoted to developing and integrating the chosen methods in RISIS datasets. It includes its implementation, initially at the Nano and the EUPRO datasets, and later at VICO and FGMF. In addition, this category relates to developing the RISIS platforms in the geographic dimension (Deliverables D9.2 and D9.3). The application of the second category would be a key component in enlarging the gained knowledge into a third category of activity dedicated to “Proposing S&T map of Europe” (see below the work plan suggested) (Deliverables D9.4 and D9.5).

3 – Activities engaged (and participants)

We engaged in November 2014 the first activity – that is depicting and selecting preferable approach(es) to knowledge clustering. It is made of 4 sub-activities that will close with the workshop planned in September 2015.

Activity 1.1:
The objective of this activity is to explore approaches for geographical aggregation. As the existing administrative classification (e.g. NUTS3 regions) was found to be insufficient for mapping STI activities, an in-depth examination for identifying alternative agglomeration methods was initiated. The research revealed two relevant approaches that suit the project’s objectives: The OECD’s method and UPEM’s method.

The OECD has developed an approach of ‘functional urban areas’ (FUA). The OECD currently applies this method to most European countries, the US and Japan. The FUA method was discussed extensively during the RISIS week in Rome. The plan is to use the FUA’s as a background map for urban areas and to embed it within the RISIS platforms and datasets. As such, it could be used as a relative characterizing of clusters tailored to one issue. Currently discussions are taking place with the OECD on receiving access to the FUA’s GIS layers (shape

4 Urban areas is defined within OECD countries using population density to identify urban cores, while travel-to-work flows are being used in order to identify the hinterlands whose labor market is highly integrated with the cores. Delineating urban cores and their hinterlands enables to define different metropolitan areas (FUA) that are comparable in 29 OECD countries (out of 34 members of the organization). Its vast applicability has numerous advantages to the objectives of WP9. For example, this method makes it possible to compare functional urban areas of similar size across countries. For further explanation of the FUA method see: OECD (2012), ‘Redefining "Urban": A new way to measure metropolitan areas’ OECD Publishing.

5 For example, UPEM plan to use it to its field of interest, Nano.
files). These geographical layers will enable us to geocode the FUA's and embed them within the different RISIS datasets. Deepening the collaboration between RISIS and the OECD could be beneficial for both sides.

As aforementioned, the second method of geo-clustering relates to UPEM. UPEM has developed an approach for analysing the distribution of S&T activities (through patents and publications) and estimating its aggregation impacts by identifying the geographical space where dense STI activity takes place. The method uses different density thresholds, in order to show their effects on the geographical distribution of clusters.

They project all addresses in the geographic space and build a first partition using a density algorithm, DBScan (Ester, Kriegel & Sander, 1996), which is parameterized using a minimal distance between two points (two couples of coordinates) and a minimal amount of addresses for building a cluster (sum of addresses in the perimiter). The borders of the convex hulls are built using these points. These envelopes build the smallest geographical divisions (initial clusters).

Using the documents (publications and/or patents) as units of analysis, we apply the CHAMELEON method (Karypis, Han & Kumar, 1999). It enables to identify collaboration regimes between individuals (authors and inventors) located in given geographical spaces. We then compare nearby initial clusters (i.e. that are less than 100 km distance) two by two, using the internal and external collaboration values of the initial clusters. Two measures are calculated: relative inter-connectivity and relative closeness that enable defining thresholds for merging initials clusters (see detailed presentation of the method).

Activity 1.2:

The objective of this activity is to test the applicability of UPEM method in other RISIS datasets and to find other methods that could be beneficial to RISIS goals and objectives. Following the discussions in the RISIS week, it was agreed that the EUPRO dataset will be used for testing the UPEM approach. The AIT provided to the UPEM the set of addresses for the FP3. FP3 has about 19716 distinct addresses across about 7 000 standardized organization names. The dataset contains the standardized organization name with the country and the address. The geographical information is divided in Postal Codes, City and Region. The UPEM has used the method developed to geocode and cluster these information to analyze their geographical distribution. This is to be discussed in the meeting planned in Vienna (10-11 September 2015).

The second part of this activity centers on exploring other approaches for spatial clustering. It is important to take into consideration the possibility that the OECD's FUA method will be either proven to be unsuitable for our needs (e.g. due to insufficient coverage for emerging economies such as China), or that the negotiation with the OECD aimed at obtaining the geographical layers will fail. Therefore, additional approaches will be considered, in order to identify, if possible: (a) other methods of urban clustering (b) other methods suitable for the RISIS S&T-based clustering. The VUA is preparing a full review for the September workshop planned for the WP9 (see below). The output of the VUA’s activity will be to provide different tools for supporting geo-coding methods for the project's datasets and platforms. Furthermore VU has explored the possibility of developing a systematic method to generate ‘shapefiles’ from list of institutionally defined geographical entities: this would enable us to progress without waiting for the outcome of the OECD negotiation, and also to enter in a process of building ‘quasi FUA’ in countries not covered by OECD.

---

6 Similar correspondences were initiated with Israel's CBS, as this country is not covered by the OECD's FUA. The target was to explore the reason for the absence and to overcome the obstacles in order to add Israel. There is no progress up to this point, due to technical issues.

7 For example considering methods that uses measures that are broader than simply geographical distance (e.g. traffic congestion, language factor, flight routes, etc.).

8 These tools (or services) are targeted to define basic geo-services to be used for geographical data harmonisation in RISIS datasets. Such services could be as conversion between different geo-coding systems, geographical boundary-related services and geo-tagging services.
Activity 1.3:

The objective of this Activity is to investigate within the limits of the 'choosing' phase other possibilities of geo-clustering in RISIS datasets. Consequently, beyond Nano & EUPRO datasets, the study in this Activity includes testing to some degree knowledge clustering at VICO dataset and at the new fast growing mid-sized firm dataset being under construction (FGMF). Other datasets, beyond WP9 participants could be also considered. Geographical clustering methods could be applied to the ETER and MORE databases. To this end, in the next few weeks exchanges with the USI and the NIFU would be made in order to better identify in geographical terms the nature of their potential needs.

Activity 1.4:

The objective of this activity is to organize a workshop that finalizes the strategy for WP9. A two-day workshop is planned in September 10th-11th for this purpose. The workshop is targeted at: 1) selecting the strategy and roadmap for the geographical clustering approach (or approaches) and 2) setting up a working schedule for carrying out the aforementioned activities that at this stage remains quite tentative. Technical arrangements are currently being conducted by the AIT, the host of the workshop.

4 – Outputs

The first deliverable of WP9 was planned to be due 15 month after the kick-off meeting ("Report on the comparison of the approaches by the 2 datasets and proposed multi-layered strategy"). However, due to administrative and research constraints, activity in WP9 started only at the end of 2014. An updated overall plan was prepared in accordance, including new timetables of deliverables. The first deliverable is planned due on December 2015. It will summarize the work of the first category, including the workshop decisions (see above).

5 – Challenges

The main challenge for the next months is the attainment of the GIS shape files, representing the OECD FUA's. As talks with the organization proceed, the need for collaboration is becoming more evident. The OECD website that presents the FUA's, provides poor evidence for knowledge and innovation dynamics. It is expected from RISIS to build and develop a set of indicators that outline innovation and STI dynamics, thus enriching the OECD's platforms. Specifically, we need to geocode the FUA and implement it to different datasets (ETER, PRO, IFRIS-PATSTAT, VICO and FGMF), while transforming RISIS data into relevant indicators that are both representative and provide good visualizations of the relative positions of the different FUA in knowledge and innovation dynamics. These tasks are challenging, as they require to convert and consolidate the existing geographical taxonomy present in the various databases into a unified and robust geographical dimension that could be applicable to all RISIS platforms.

6 – Programme for next months

Next months will be devoted to complete “Depicting and selecting preferable approach(es) to knowledge clustering”. All the activities in this category are expected to be finalized in November 2015. By the end of 2015, the second category will be started. This category defined as a 'developing and integrating' phase, deals with 'Operationalizing and testing the selected approach(es) in RISIS

\(^9\) See ‘Description of Work’ (DOW).
datasets’. It will be running for one year, until the end of 2016. Operationalizing and testing concerns two aspects: geocoding addresses and applying the clustering approach selected. As stated, it might add a third one integrating the FUA as a reference point. In this category we include the following activities:

2.1 Operationalizing the approach selected in category 1 with the necessary documentation (UPEM, VUA, SNI).

2.2 Applying the approach in the core datasets concerned –Nano, EUPRO, VICO and FGMF with SNI/UPEM support (UPEM, AIT, POLIMI, SNI).

2.3 Developing a suit of tools that can be integrated to both SMS and Cortext. That is, exposing a set of basic and innovative services through Linked Data API embedded in SMS platform. The different tools can be used both stand-alone and combined with UPEM geographical clustering approaches. (VUA, UPEM).

2.4 Developing tests in order to compare clusters arrived at, and develop an approach for cluster-based integration of datasets, along two dimensions. 1. Testing key assumptions of the literature on geographical agglomerations (e.g. the anchor tenant hypothesis, privileged relations between core nodes, etc.). 2. Testing key policy issues (e.g. on the circulation of knowledge) (UPEM, AIT, POLIMI).

3 – WP8 REPORT ON THE ETER REGISTRY

by Benedetto Lepori, Università della Svizzera italiana, Michael Ploder and Daniel Wagner-Schuster, JOANNEUM RESEARCH, Hebe Gunnes, NIFU (Version 31.07.2015)

1 – Basic characteristics

1.1. Overview of the facility

The ETER-EUMIDA facility is a set of databases providing a register of European Higher Education Institutions and containing basic statistical information on them, including descriptors, geographical information, students and graduates, personnel, finances, and research activities. These databases, created by merging data from national statistical authorities, are the only available comprehensive information on European higher education, and thus are of fundamental value for analytical purposes.

The ETER-EUMIDA facility is composed by three components:

- The ETER main database including data on the years 2011 and 2012 (2013 and 2014 to be collected). It is available on-line at http://eter.joanneum.at/imdas-eter, but off-line copies (in csv, xls and SPSS format) are also available.
- The EUMIDA database including data on the year 2008. It is available as off-line file (SPSS format).
- A file including additional data for the years 2008, 2011 and 2012, including the numbers of scientific publications, impact, participation to European programs. It is also available off-line as SPSS data file.

10 Subjected to the outcomes resulted from the negotiations with the OECD.

11 If the OECD data will be available, it will be possible to convert FUA data as a new dataset in RISIS platform and expose this data through VUA Linked Data API.
The three files share the same identifier system (see below 0) and therefore can be easily matched and combined for purposes of analysis.

Coverage is very extensive as both databases include not only doctorate-awarding HEIs, but also second-tier HEIs in binary systems, as well as a large number of specialized schools at the tertiary level. The estimated coverage of tertiary education at the bachelor, master and PhD is around 100%. Coverage currently includes all EU-28 countries, Iceland, Norway, Liechtenstein, Switzerland, and FYORM (some countries are missing for some years). The two datasets are built on the same basic approach, but also display some relevant differences in coverage, data sources, methodology and especially, engineering.

This report presents the characteristics of the ETER dataset and of its infrastructure. We will also highlight differences with EUMIDA, since a major issue for RISIS is to provide joint access to both datasets and present the additional data provided together with ETER and EUMIDA for purposes of scholarly research.

This report refers to the status of ETER in July 2015, after completing the first two waves of data collection have been completed. It draws heavily on the ETER handbook (Lepori, Bonaccorsi, Daraio, et al 2015a), as well as on the accompanying report on data collection (Lepori, Bonaccorsi, Daraio, et al 2015b).

This report reflects the status of the facility at the end of July 2015. It will be subsequently updated as soon as additional releases will be available.

1.2. Context and aims

EUMIDA and ETER have to be seen in the context of the movement towards the provision of micro-data at the level of individual Higher Education Institutions, which started with the PRIME-AQUAMETH project (Bonaccorsi and Daraio 2007). PRIME-AQUAMETH was an experimental project, which proved the feasibility of collecting HEI-level data from different countries and publicly available sources. Further, the project proved that, despite existing comparability problems (Bonaccorsi, Daraio, Lepori and Slipersaeter 2007), these data could be used to provide interesting analytical insights on European higher education (Lepori, Probst and Baschung 2010, Daraio, Bonaccorsi, Geuna, Lepori and et. al. 2011).

Following these results, the European Commission (DG EAC and EUROSTAT) funded in 2010-2011 a large-scale feasibility study for a European Register of Tertiary Education. The EUropean MIcro DAta (EUMIDA) project developed a consistent methodology for the delineation of a census of HEIs in Europe and for the data collection from official sources. Further, the project managed to provide the first register of European Higher Education and to collect a large number of data (Lepori and Bonaccorsi 2013, Bonaccorsi 2014).

The European Tertiary Education Register (ETER) was launched in 2013 in order to consolidate and establish the EUMIDA methodology and to prepare for a regular data collection on European Higher Education. It is a service contract from the Directorate General of Education, Audiovisual and Culture (DG-EAC) and realized by a consortium of four partners (Università della Svizzera italiana, JOANNEUM RESEARCH, NIFU and Università Roma La Sapienza) in cooperation with a number of national experts and national statistical authorities.

ETER builds on the results and experience of the EUMIDA (EUropean MIcroDAta collection) study, and has the following goals:

- To further develop the indicators tested in EUMIDA towards a set of more complete indicators while characterizing HEIs in the ERA by following their main activity dimensions.
- To extend the coverage of the EUMIDA dataset and consolidate the European Higher Education perimeter, i.e. the list of HEIs officially included in the dataset.
- To collect, validate, and publish the data for the defined perimeter for the years 2011 and 2012.
- To document the methodology developed for the project in a methodological handbook, while providing suggestions for the consolidation of the European Tertiary Education Register and for a regular data collection on European HEIs.
As of July 2015, the contract has completed the first two waves of data collection, referring to the year 2011 (respectively the academic year 2011/2012) and 2012 (academic year 2012/2013), finalized the data cleaning, and the dataset was made available on-line.

In July 2015, the European Commission awarded a follow-up contract for two additional years to the same consortium (plus the University of Pisa). The new contract foresees two additional waves of data collection for the year 2013 (to be published in June 2016) and 2014 (to be published in June 2017); some additional variables are foreseen, as well as a stronger focus on policy analysis and dissemination.

1.3. Legal name of the operating organization

The current ETER database is operated by JOANNEUM RESEARCH, Institute of Information and Communication Technology, Graz on behalf of the ETER consortium composed by Università della Svizzera italiana, JOANNEUM RESEARCH, NIFU and Università Roma la Sapienza. While the dataset has been collected through a service contract of the European Commission, and therefore is owned by the EU, the database system itself is proprietary to JOANNEUM RESEARCH.

1.4. Database location and type of access

The ETER database is located at JOANNEUM RESEARCH and is accessible on-line through a web application: http://eter.joanneum.at/imdas-eter/.

Users can perform searches on the datasets and then download either the whole dataset or parts of it in an Excel format. Additionally, it is possible to download metadata and demographic information on the included HEIs, as well as methodological information on the project (methodological handbook). Public access is available for most of the data. Some data is available for research purposes only. For this data, restricted access under a non-disclosure agreement is possible.

The EUMIDA dataset and the ETER-EUMIDA additional data are available as stand-alone files, which can be provided to interested users under the signature of a non-disclosure agreement. They are available to registered users from the RISIS website.

2 – Data content of ETER

2.1. Definition and description of observations

A key task of EUMIDA and ETER was to provide a clear definition of the observation units and of the perimeter to be considered, i.e. which units should be included in the register and in the data collection.

The unit of observation in ETER are Higher Education Institutions (HEIs). These are defined as entities

- which are recognizable as distinct organizations,
- which are nationally recognized as HEIs, and
- whose major activity is providing education at the tertiary level (ISCED 2011 level 5, 6, 7 and/or 8). R&D activities might be present, but are not a necessary condition for inclusion.

In practice, the definition of HEIs included universities (doctorate-awarding), non-university higher education (like Fachhochschulen) and specialized schools delivering education at the tertiary level. We provide below more detailed information on coverage in respect to the whole of national tertiary education.
The focus on HEIs implies that all data in ETER are aggregated at the level of the whole HEI and no subunit data are provided (even in case of multi-site organizations). The only relevant exceptions are foreign campuses, which follow standard practices in educational statistics and are treated as stand-alone organizations. ETER includes only descriptive information on the existence and localization of other education sites, but no disaggregated data at this level. The definition of the organization is taken as adopted by national statistical authorities. In most cases, it corresponds to the legal entity, but some departures are possible for associated units and research centres. Different national practices in handling resources, staff, and students, meaning the actual definition may slightly differ by country. A threshold has been implemented in the perimeter. This implies that institutions must have at least 200 students or 30 employees to be included in the ETER dataset. EUMIDA did not have such a threshold, and all research active HEIs, regardless of size, were included in the EUMIDA perimeter; as a matter of fact, practices in this respect differed between countries, with a few also including very small HEIs, while others were more restrictive.

2.2. Number of units and demography

ETER includes 2675 unique units referring to the year 2011, while EUMIDA includes 2471 units referring to the year 2009. However, the EUMIDA-ETER additional file includes only the 1378 HEIs in the so-called restricted set, for which enough data are available. ETER introduced stable IDs across time for HEIs, as well as demographic notations to handle processes like births, deaths and mergers. In particular, HEIs common to EUMIDA and ETER carry the same ID and thus can be easily matched: more precisely, 1974 of 2471 HEIs in EUMIDA have been carried over in ETER. Most new cases concern countries not covered by EUMIDA. About 350 HEIs included in EUMIDA have been excluded in ETER because of not reaching the size threshold. Importantly, no other variable is associated uniquely to IDs, but all HEI characteristics are collected and recorded by individual years, including variables such as HEI name, country, and location. Technically, unique records in the ETER database are characterized by the pair HEI-ID and YEAR. Additionally, consistency across time is maintained as IDs are carried over from year to year when HEIs do not witness demographic changes. Rules for handling demographic events correspond to the approach of the Business Units Register. For instance, in the case of mergers, the parent organizations ID’s are reserved, but become non active, whereas the new organization receives a new ID. A separate record of demographic events is maintained in order to be able to track dependencies (see the ETER handbook for full reference). The table below provides an overview of demographic events between EUMIDA and ETER. Note that changes in the name of an organization do not necessarily constitute a demographic event and imply a change of ID, which is also the case when an institution changes its legal status.

<table>
<thead>
<tr>
<th>Event code</th>
<th>Description</th>
<th>N EUMIDA</th>
<th>N ETER</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>no demographic event</td>
<td>1974</td>
<td>1974</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>entry</td>
<td>0</td>
<td>557</td>
<td>Related to the integration of new countries in the dataset</td>
</tr>
<tr>
<td>1</td>
<td>entry</td>
<td>0</td>
<td>80</td>
<td>Entries in countries already covered by EUMIDA</td>
</tr>
<tr>
<td>2</td>
<td>exit</td>
<td>382</td>
<td>0</td>
<td>Most cases because HEIs are below the size threshold of ETER</td>
</tr>
<tr>
<td>3</td>
<td>birth</td>
<td>0</td>
<td>24</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>death</td>
<td>36</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>merger</td>
<td>32</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>split</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>take-over</td>
<td>42</td>
<td>21</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>spin-out (spin-off)</td>
<td>0</td>
<td>2</td>
<td>Two cases of foreign colleges created by existing HEIs in the perimeter</td>
</tr>
</tbody>
</table>

Source: Lepori et al. (2014b).
2.3. Data acquisition and processing (e.g. data cleaning)

As a general rule, ETER data are provided from official sources, like the National Statistical Authorities and/or Research and Higher Education Ministries, meaning data should be considered as officially validated. Primary sources are in most cases administrative data collected at the national level for management and statistical purposes (like the preparation of the EUROSTAT educational statistics).

There are a few exceptions to this rule: in many cases, organizational descriptors and geographical information were provided by the consortium itself based on Internet sources. In a few countries, statistical data were also elaborated by the consortium based on data published online by the NSA. A full documentation of sources is provided in the metadata.

The process was organized in the following main steps.

a) Definition of the perimeter. For the countries participating in EUMIDA, the National Statistical Authorities received a pre-filled sheet including all EUMIDA HEIs in their respective countries and were requested to update the sheet by indicating changes which occurred in the period 2009-2011. A systematic notation of demographic events was provided. The perimeter also included the official name of the HEI (in the national language and English) in 2011 as reference.

b) Data collection. Based on the perimeter, the consortium provided master MS-Excel files for the data collection. These sheets include all variables, as well as remark fields. Moreover, metadata were collected in separate sheets for each individual variable. NSAs are also requested to provide detailed methodological explanations on some variables for which comparability is expected, specifically financial and staff data.

c) Checking of the data. Data sheets received were subject to a set of extensive checks for inconsistencies and problematic values by the consortium (see table 2 below). This was important in order to address issues related to underlying problems in the data, misinterpretation of guidelines, as well as simple mistakes in the handling of the data. A list of checks which have been performed can be found in the ETER handbook.

2.4. Cleaning of the EUMIDA datasets

The original dataset for the year 2008 produced by the EUMIDA project displayed a number of differences in respect to ETER. The most important ones are:

- The list of variables is partially different, not only concerning their names, but also less variables were included especially for what concerns descriptors and geographical information.
- Some classifications differed, as EUMIDA used the old ISCED-1997 classification, as well as FOE-2007 for the fields of education.
- No systematic coding of missing data was available, they were included as blank cells in the database.
- No flags and less remarks were included in the dataset.
- The documentation of the dataset and metadata is less systematic and less usable, as it was provided in the form of a large word file.

The cleaning of the dataset had the goal of making EUMIDA as similar as possible to ETER in order to allow users to combine in a smooth way the two datasets for the purposes of longitudinal analysis. Following main changes were made:

- Descriptors and geographical information was systematically compared between ETER and EUMIDA and inconsistent cases were cleaned. To the extent of possible, the additional variables introduced in ETER have been integrated in EUMIDA as well (for example postcode and geographical coordinates).
- All EUMIDA variables have been renamed identically to ETER. Variable codes and labels have also been made uniform.
- Blank cells have been coded to ETER standard codes.
- Classifications have been updated. For example ISCED-F-2011 is adopted (the two missing fields not included in FOE-1997 consistently recoded as "xc").
- Flags and remarks columns have been introduced in order to be able to track any specific issues.
- Accuracy and consistency checks have been run.

The revised EUMIDA file will be available in SPSS format, accompanied by the original metadata file and by a file specifying changes which have been made.

Table 2. Description of data validation procedures in ETER

<table>
<thead>
<tr>
<th>Type of checks</th>
<th>Description</th>
<th>Procedure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy checks</td>
<td>Accuracy checks verify that data entered have the right format foreseen by the handbook (for example years as a 4-digit numeric code) and that no logically impossible values are found (foundation year&gt;2011).</td>
<td>Accuracy checks are performed in the data collection sheet and on delivered data. Simple mistakes are corrected directly, whereas unclear cases are reported back to NSAs/NEs for clarification and correction.</td>
</tr>
<tr>
<td>Completeness checks</td>
<td>Particular attention in ETER has been devoted to the correct coding of blanks and missing values. As a general rule, ETER does not include any blank cells and the null value (“0”) is reserved for the case where it is known that the corresponding value is null. All other cases are coded with special codes like missing (“m”), not applicable (“a”), included in totals (“x”) or in another column (“xc”) or row (“xr”).</td>
<td>Blank cells are highlighted automatically. Clear cases are recoded directly and ambiguous cases (for example between missing and not applicable) are reported back to national experts and NSAs.</td>
</tr>
</tbody>
</table>
| Consistency checks      | a) These checks control for logical consistency between different variables, for example when the highest degree delivered is at ISCED 7 level, all values for students and graduates at ISCED8 level should be not applicable. Rules in this respect are stipulated in the handbook.  
  b) Further, these checks also control whether the sums of breakdowns by subcategories equals the total and numerical relationships between values (example R&D expenditures lower than total expenditures). | Clear cases are recoded directly and ambiguous cases (for example between missing and not applicable) are reported back to national experts and NSAs. Remaining inconsistencies are flagged. |
| Deviant cases           | Standard ratios are calculated (for example students to graduates) and compared to the national averages. See the handbook for the list of ratios. | Deviant values are identified and checked. In case there are specific reasons, an explanation is added to the metadata for that specific HEI. |
| Check of missing data   | An analysis of missing data is performed (also including issues of breakdowns by subcategories). | When it is expected that data should be available, possibly with some limitations, this is requested to NE/NSAs. |
| Control of metadata completeness | Metadata are systematically controlled for completeness, taking into account also issues emerging from the checks on the data. | When metadata are missing or incomplete, further information is requested. Quality of metadata is critical for the exploitation of the database. |
| Expert checks           | Expert checks based on knowledge of national systems, as well on information available on the Web and EUMIDA data, are performed in order to ensure that provided data are realistic. | Potentially problematic cases are notified back to national experts and NSAs. When these are related to methodological issues, the corresponding remarks are integrated into the metadata. |
2.5. Information all variables/indicators

Both ETER and EUMIDA include the following groups of variables:

- Institutional descriptors, e.g. the name of the institution and the foundation year.
- Geographical descriptors like the NUTS region, the city of the main seat and its postcode.
- Data on numbers of students and graduates divided by ISCED-2011 level, by gender, fields of education, citizenship and mobility.
- Data on HEI expenditures and revenues.
- Data on the number of staff, divided between academic and non-academic, available both as headcount (HC) and full-time-equivalents (FTE), as well as on the number of professors (HC).
- Data on research activities (PhD students, R&D expenditures).

Differences between EUMIDA and ETER are as follows. ETER includes a few additional variables, especially concerning geographical information and staff. Moreover, in EUMIDA, many variables and breakdowns are only available for the so-called research-active HEIs. These institutions comprise about half of the EUMIDA sample, but most of the students and research activities. Finally, ETER adopts the more recent release of classifications of educational levels (ISCED-2011) and educational fields (ISCED-F).

To the extent of possible, the current version of EUMIDA has been updated and extended to match ETER definitions (see details below). Variables in ETER and EUMIDA also have identical labels in order to make a merge of both datasets easier.

For full reference on the revision and cleaning of the EUMIDA dataset the reader is requested to read the corresponding document.

List of variables

The table below provides a full list of ETER variables, with additional information on their availability in EUMIDA. For complete information and definitions the reader should refer to the ETER handbook. In general, most definitions concerning students, graduates, staff and finances are derived from the UOE data collection handbook (UOE 2006), respectively with the OECD 2002 Frascati manual for R&D expenditures and thus comply with official statistics at EUROSTAT and OECD.

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Variables</th>
<th>Differences in EUMIDA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identifiers</td>
<td>ETER ID</td>
<td>Same variables in EUMIDA.</td>
</tr>
<tr>
<td></td>
<td>Country code</td>
<td></td>
</tr>
<tr>
<td></td>
<td>National identifier (optional)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Institution name (in own language)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>English institution name (if available)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Year</td>
<td></td>
</tr>
<tr>
<td>Basic institutional descriptors</td>
<td>Country Code</td>
<td>The same variables are also included in EUMIDA.</td>
</tr>
<tr>
<td></td>
<td>Legal status</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Institution category, national definition (in own language)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Institution category, national definition (in English, if available)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Institution category standardized</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Foreign campus</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Foundation year</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Legal status year</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ancestor year</td>
<td></td>
</tr>
<tr>
<td></td>
<td>University hospital</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Institutional website</td>
<td></td>
</tr>
<tr>
<td>Geographic information</td>
<td>Region of establishment, NUTS2 code</td>
<td>The same variables are also included in EUMIDA.</td>
</tr>
<tr>
<td></td>
<td>Region of establishment, NUTS3 code</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Name of the city</td>
<td></td>
</tr>
<tr>
<td>Postcode</td>
<td>Multi-site institution</td>
<td></td>
</tr>
<tr>
<td>----------</td>
<td>------------------------</td>
<td></td>
</tr>
<tr>
<td><strong>Educational activities</strong></td>
<td>Highest degree delivered Number of enrolled students at ISCED levels 5, 6, 7, by fields of education, gender, citizenship and mobility Number of graduates at ISCED levels 5, 6, 7, by fields of education, gender, citizenship and mobility Distance education institution</td>
<td></td>
</tr>
<tr>
<td><strong>Research activities</strong></td>
<td>Research active institution Number of enrolled students at ISCED levels 8, by fields of education, gender, citizenship and mobility Number of graduates at ISCED levels 8 (doctorates), by fields of education, gender, citizenship and mobility R&amp;D expenditure Personnel expenditure Non-personnel expenditure Capital expenditure Accounting of capital expenditure</td>
<td></td>
</tr>
<tr>
<td><strong>Expenditures</strong></td>
<td>Core budget Third party funding Private funding Tuition fees Student fees funding</td>
<td></td>
</tr>
<tr>
<td><strong>Staff</strong></td>
<td>Number of academic staff in FTE Number of academic staff in headcounts Number of academic staff, by fields of education gender and citizenship in headcounts Number of administrative staff in FTE Number of administrative staff in headcounts Number of full professors (HC) Inclusion of PhD students Number of total staff in FTE Number of total staff in HC</td>
<td></td>
</tr>
<tr>
<td><strong>Indicators</strong></td>
<td>Share of women among students ISCED6 and ISCED7. Share of women among graduates ISCED6 and ISCED7. Share of women among PhD students and graduates (ISCED8). Share of women among academic staff and full professors. Share of foreigners among students ISCED6 and ISCED7. Share of foreigners among graduates ISCED6 and ISCED7. Share of foreigners among PhD students and graduates (ISCED8). Share of mobile among students ISCED6 and ISCED7. Share of mobile among graduates ISCED6 and ISCED7. Share of mobile among PhD students and graduates (ISCED8). Subject concentration of undergraduate students (ISCED5-7). Subject concentration of PhD graduates (ISCED8). Subject concentration of academic staff. PhD intensity Full professors as share of academic staff (headcounts). Academic staff as share of total staff (headcounts). Core budget as a share of total budget. Third-party funds as a share of total budget. Students’ revenues as a share of total budget.</td>
<td></td>
</tr>
</tbody>
</table>

Number of students: only aggregated values for ISCED5-7 are available (no breakdown between levels 5,6 and 7). The 1997 FOE classification is used, hence ISCED-F04 included in ISCED-F03 and ISCED-F06 included in ISCED-F05.

The 1997 FOE classification is used, hence ISCED-F04 included in ISCED-F03 and ISCED-F06 included in ISCED-F05.

Staff data only in FTEs. No data by fields and on numbers of professors. No data on gender.

Not currently available in EUMIDA
Additional variables
The additional dataset for ETER-EUMIDA includes variables not covered by the core dataset, but relevant for purposes of analysis of higher education. The current set of variables includes:

- The number of participations of EU Framework Programmes derived from the EUPRO database.
- The number of publications of HEIs derived from the Leiden rankings.
- A correspondence table with the OECD/EUROSTAT functional urban areas, which allows matching ETER with data on regional development (http://www.oecd.org/regional/redefiningurbananewwaytomeasuremetropolitanareas.htm).

This set of variables might be expanded in the future. For full reference on the additional dataset the reader should consult the corresponding document with detailed description of variables and sources.

2.6. Temporal, geographical and sectoral coverage

Temporal coverage
ETER-EUMIDA data are in principle collected for every year. EUMIDA provides data for the year 2008, whereas ETER for the year 2011 and 2012. Data for the years 2013 and 2014 will be available in summer 2016, respectively 2017.

Depending on the nature of the data and the practices of data collection, individual data refer to slightly different periods as detailed in Table 4. Departures from these reference periods for individual countries are recorded in the metadata.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Reference period/date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Descriptors and geographical information</td>
<td>Last day of calendar year (31st of December).</td>
</tr>
<tr>
<td>Expenditures</td>
<td>Calendar year (1st January – 31st of December).</td>
</tr>
<tr>
<td>Revenues</td>
<td>Calendar year (1st January – 31st of December).</td>
</tr>
<tr>
<td>Personnel FTE</td>
<td>Calendar year based on person-years.</td>
</tr>
<tr>
<td>Personnel headcount</td>
<td>End of first month of beginning of academic year.</td>
</tr>
<tr>
<td>Students</td>
<td>End of first month of beginning of academic year.</td>
</tr>
<tr>
<td>Degrees (including PhD degrees)</td>
<td>Academic year or calendar year (to be specified).</td>
</tr>
</tbody>
</table>

Geographical coverage
ETER covers all 28 European Union member states, EEA-EFTA countries (Iceland, Liechtenstein, Norway and Switzerland), as well as candidate countries (the Former Yugoslav Republic of Macedonia, Montenegro, Serbia and Turkey), for a total of 36 countries.

Data for the following countries are currently missing:
- EUMIDA 2008: France, Iceland, Liechtenstein, the Former Yugoslav Republic of Macedonia, Montenegro, Serbia and Turkey.
- ETER 2011: Belgium (French part), Romania, Slovenia, Montenegro, Serbia and Turkey.
- ETER 2011: Belgium (French part), Romania, Slovenia, the Former Yugoslav Republic of Macedonia, Montenegro, Serbia, Turkey, Iceland and Hungary.

Sectorial coverage
ETER aims to provide a fairly complete coverage of the European Higher Education sector (tertiary education), as defined by graduation of at least the ISCED-2011 level 5 (diploma at the tertiary level). The criterion that tertiary education should be a major activity excludes professional organizations (delivering only some curricula), as well as public research organizations (even if they employ PhD students as researchers).
Almost all ETER organizations belong to the higher education sector. There are a few cases of research organizations included in ETER either because they award a large number of PhD degrees or because they are associated with a university. In general, the ETER coverage matches fairly well the official list of HEIs considered as part of the higher education system at the national level, with the exclusion of very small HEIs. As previously mentioned, a threshold has been implemented; meaning institutions must have at least 200 students and 30 staff (FTE) to be included in the ETER dataset. Coverage at the bachelor, master and PhD level in ETER should be considered as quite complete. The same applies for research activities in the higher education sector. In respect to all tertiary education (graduating from ISCED level 5 onwards), ETER covers educational providers delivering short diplomas of less than 3 years only to a limited extent (ISCED level 5), like professional schools in the vocational sector or preparatory classes before university studies. In terms of the number of students, ETER includes 85% of the tertiary education students in the countries that delivered data, respectively 65% in the countries that provided information about the perimeter, i.e. the number of higher education institutions.

2.7. Classifications used in the database

Type of HEI
This variable specifies a European-level standardized classification of Higher Education Institutions. It is relevant in order to provide comparative analysis of higher education systems and analyze subgroups. It is available in ETER only.

The following categories are used:
- UNI (university). These HEIs display a largely academic orientation (without excluding some focus on applied research), they have the right to award doctorate degrees, and can bear the full name of “University” (including variants like technological university, etc.). In general, awarding doctorates should be the main criterion to classify HEIs in this category, even if a few doctoral-awarding HEIs might be included in the two following categories.
- UAS (university of applied sciences). These institutions are officially recognized as a part of higher education, though not as universities (see definition above). Commonly these institutions have a focus on professional education. In most cases they do not have the right to award a doctorate (exceptions are possible). Examples are Fachhochschulen (Austria, Germany), Hogescholen (Netherlands), colleges (Norway), and Polytechnics (Finland). This institutional category applies only to countries that have a binary HE system, where these institutions are given a specific legal status. Examples include Norway, Switzerland and the Netherlands.
- Other. All institutions that do not fit the description of university/university of applied science will be categorized as “other.” This may apply to institutions such as art academies and military schools. Also technological and professional schools in countries without a binary system (like the UK or France) should be classified in this way.

Legal status
Consistent with the UOE manual, a classification of HEIs by legal status is provided both in ETER and EUMIDA.

The distinction between public and private is made according to whether a public agency or a private entity has the ultimate control over the institution. Ultimate control is decided with reference to who has the power to determine the general policies and activities of the institution and to appoint the officers managing the school. Ultimate control will usually also extend to the decision to open or close the institution. As many institutions are under the operational control of a governing body, the constitution of that body will also have a bearing on the classification.
Private institutions are further divided between government dependent – which either receives more that 50% of their core funding from government agencies or whose teaching staff is paid by a government agency – and independent private. Thus, this classification includes three categories: public, private, private government-dependent.

**Geographical classification**
ETER provides extensive geographical information on HEIs. For each HEI, the following geographical data is included:
- Country code.
- NUTS 2 and NUTS 3 code of the main seat (only NUTS included in EUMIDA).
- The name of the city and the postcode of the legal seat (not included in EUMIDA).
- Geographical coordinates (derived from the postcode) are foreseen in one of the next releases (not included in EUMIDA).

Additional information is provided for multi-site institutions, i.e. HEIs having establishments in different cities. This includes additional NUTS codes and descriptive information on the locations.

**Levels of education**
The International Standard Classification of Educational Degrees (ISCED)\(^2\) provides information on the level of curricula and is therefore used to break down numbers of students and graduates. ETER adopts the more detailed ISCED-2011 classification, which includes separate classifications for bachelor (ISCED6), master (ISCED7) and PhD (ISCED8). The classification scheme used in ETER singles out the so-called long degrees (ISCED7 long), i.e. 4-5 year curricula without an intermediate classification. EUMIDA used a simpler distinction between diploma, bachelor, master and PhD, which can, in principle, be matched to ISCED-2011.

The table below provides detailed information on the classification used in ETER.

**TABLE 5. ISCED-2011 CLASSIFICATION**

<table>
<thead>
<tr>
<th>ISCED-2011 level</th>
<th>Definition</th>
<th>Criteria</th>
</tr>
</thead>
</table>
| ISCED 5 short-cycle tertiary education                | Programs at ISCED level 5, or short-cycle tertiary education, are often designed to provide participants with professional knowledge, skills and competencies. Typically, they are practically based, occupationally specific and prepare students to enter the labor market. However, these programs may also provide a pathway to other tertiary education programs. Academic tertiary education programs below the level of a Bachelor’s program or equivalent are also classified as ISCED level 5. | Duration: 2-3 years  
Entry requirements: ISCED 3 or 4                                                                       |
| ISCED 6 Bachelor’s or equivalent levels               | Programs at ISCED level 6, bachelor’s or equivalent level, are often designed to provide participants with intermediate academic and/or professional knowledge, skills and competencies, leading to a first degree or equivalent qualification. Programs at this level are typically theoretically-based but may include practical components and are informed by state of the art research and/or best professional practice. They are traditionally offered by universities and equivalent tertiary educational institutions. | Duration: 2-3 years  
Entry requirements: ISCED 3 or 4  
Usually: first degree at tertiary level                                                                  |
| ISCED 7 Master of equivalent level                    | Programs at ISCED level 7, master’s or equivalent level, are often designed to provide participants with advanced academic and/or professional knowledge, skills and competencies, leading to a second degree or equivalent qualification. Programs at this level may have a substantial research component but do not yet lead to the award of a doctoral qualification. Typically, programs at this level are theoretically-based but may include practical components and are informed by state of the art research and/or | Duration: 2-3 years  
Entry requirements: ISCED 6  
Usually: second degree at tertiary level  
Direct access to ISCED 8 level                                                                           |

**Fields of education**

The Fields of Education and Training classification allows breaking down numbers of students and graduates by field of study. It is envisaged to introduce this classification also for academic staff. ETER adopts the most recent version of the classification, i.e. the Fields of Education and Training 2013 classification. EUMIDA used the previous Fields of Education classification 1997: the main difference between the two classification schemes is that in ISCED-F 2013 two new fields have been distinguished, namely Business, Administration and Law (included in social sciences in the previous schemes) and ICT (included in natural sciences in FOE-1997).

For students and graduates, the breakdown by field of education is provided separately by level of education.

**Table 6. Fields of Education and Training Classification**

<table>
<thead>
<tr>
<th>Code</th>
<th>Name</th>
<th>Subfields</th>
<th>ISCED 1997 FOE</th>
</tr>
</thead>
<tbody>
<tr>
<td>00</td>
<td>General programs and qualifications</td>
<td>001 Basic programs and qualifications 002 Literacy and numeracy 003 Personal skills</td>
<td>01 Basic programs 08 Literacy and numeracy 09 Personal development</td>
</tr>
<tr>
<td>01</td>
<td>Education</td>
<td>011 Education</td>
<td>14 Teacher training and education science</td>
</tr>
<tr>
<td>02</td>
<td>Humanities and Arts</td>
<td>021 Arts 022 Humanities 023 Languages</td>
<td>21 Arts 22 Humanities</td>
</tr>
<tr>
<td>03</td>
<td>Social sciences</td>
<td>031 Social and behavioral science 032 Journalism and information</td>
<td>31 Social and behavioral science 32 Journalism and information</td>
</tr>
<tr>
<td>04</td>
<td>Business and law</td>
<td>041 Business and administration 042 Law</td>
<td>34 Business and administration 38 Law</td>
</tr>
<tr>
<td>05</td>
<td>Natural Science, mathematics and statistics</td>
<td>051 Biological and related sciences 052 Environment 053 Physical sciences 054 Mathematics and statistics</td>
<td>42 Life sciences Part of 62 (natural parks and wildlife) 44 Physical sciences 46 Mathematics and statistics</td>
</tr>
<tr>
<td>06</td>
<td>Information and communication technologies</td>
<td>061 Information &amp; Communication Technologies</td>
<td>48 Computing</td>
</tr>
<tr>
<td>07</td>
<td>Engineering, manufacturing and construction</td>
<td>071 Engineering and engineering trades 072 Manufacturing and processing 073 Architecture and construction</td>
<td>52 Engineering and engineering trades (plus most of 85 environmental protection) 54 Manufacturing and processing 58 Architecture and building</td>
</tr>
<tr>
<td>08</td>
<td>Agriculture, forestry, fisheries and veterinary</td>
<td>081 Agriculture 082 Forestry 083 Fisheries</td>
<td>62 Agriculture, forestry and fishery (minus natural parks and wildlife) 64 Veterinary</td>
</tr>
</tbody>
</table>

---

**Citizenship and mobility**
ETER includes classification of students, graduates and staff by:

- Citizenship, distinguishing between nationals and foreigners.
- Mobility (students and graduates only), distinguishing between residents (nationals and foreigners who earned their qualifying education in the country) and mobile students.

Both classifications are compliant with standard EUROSTAT definitions from the UOE manual (UOE 2006).
These breakdowns are provided separately by ISCED level (but not by field of education).

3 – Quality and accuracy of data
ETER data have been subject to an extensive data validation and quality control procedure, which has been coordinated by the University of Rome La Sapienza together with the other consortium partners. Since ETER has no control over the primary sources, much of the data quality process is concerned with documenting methodological departures and comparability problems in order to make the users aware of potential problems.

EUMIDA data have been subject to some data quality controls, but in a somewhat less systematic way. Documentation of methodological and comparability issues is provided in an accompanying document, but currently not systematically linked to the data themselves.

3.1. Overall data quality approach
The ETER approach to data quality is based on the combination of two integrated processes:
- preliminary level quality and validation checks performed within the data collection phase on a country basis in order to allow for an easy return on the respondents and the correction of data before online integration (see also chapter 2), and
- a final quality and validation phase which has the role of performing more complex controls that can provide hints to use data in the appropriate way and improve the quality of the collection in the second wave. In this phase the quality control is performed on data at both a “global” and “local” level.

Different methods are applied:
- A systematic analysis of internal data quality, more specifically referring to four dimensions: format accuracy, completeness, consistency, and timeliness (see the ETER handbook for details).
- Advanced statistical methods for outlier detection by estimation of the distribution of observations according to a model distribution (mostly a lognormal distribution) and by identification of extreme values which do not fit into this distribution (see the ETER handbook and Ruocco and Daraio 2013, for full details).
3.2. Accuracy and consistency

**Accuracy** evaluates the compliance of the value to the requested format, as defined in the data chapter of the ETER handbook, respectively in the definitions of each variable. This includes characteristics like being non-negative for all financial values, and student and graduate data being integer variables, among others. This also includes the correct coding of missing and null values. **Consistency** verifies possible violations of semantic rules defined over the involved data, and specifically between different variables.

Given the nature of the ETER dataset, there is a high number of mutual dependencies between variables, which can be exploited for purposes of data quality analysis. In broad terms, they can be regrouped in the following categories (see a complete list in the ETER handbook):
- Logical dependencies between categorical variables and values. For example, when the highest degree delivered is ISCED 5, all numbers of students and graduates at levels 6-8 have to be coded as “not applicable.” Similarly, if an HEI is non-research active, R&D expenditure should be “not applicable.” Most of these rules are already stipulated in the definition of these variables.
- Sums of breakdowns of variables equal to the total, for example the sum of male, female, and gender unclassified students should be equal to the total.
- Relationships between valued variables. For example, R&D expenditure should be lower than total expenditures. The ancestor year should precede the foundation year of the actual HEI (which, according to the definition in ETER, should precede or be the same as the legal status year).

These dimensions have been systematically checked during the data collection process. A final check on the complete dataset has been performed before publication.

Overall, the current version of the ETER dataset reaches a very good level of accuracy and consistency. Very few remaining cases are due to national specificities or simply to rounding errors. These have been flagged in the dataset.

3.3. Completeness

Completeness of data strongly varies by variable and by country. Overall, we can summarize the situation as follows:
- Descriptors are generally available for all countries, with the exception of a few cases where information on foundation years was not available.
- Financial data (revenues, expenditures, R&D expenditures) are available for only about one-third of the countries in 2011.
- Staff data are generally available in most countries. The main exceptions are countries which provided for the time being only the descriptors and no statistical information. However, the breakdown of academic staff between national and foreign citizenship is available for a much smaller number of countries.
- Students and graduates data are available for most countries, including breakdowns by gender, nationality, and fields of education. The breakdown by mobile students is less widely available. Since this is a breakdown requested by EUROSTAT, but not yet implemented in all countries, the number of countries providing this breakdown is believed to increase.
- The situation is similar for PhD students. Some major countries, such as Spain and the UK, did not deliver data on PhD-students at all.

The level of completeness varies largely by country. In the analyzed dataset there is a group of countries with a very high level of completeness (over 85%) including CH, CY, DE, DK, IE, IT, LI, LU, MT, NO, SE, a second group with an acceptable level (50%-85%) including AT, BG, CZ, EE,
ES, FI, GR, IS, LT, LV, MK, NL, PL, PT, and a third group with minor data availability (below 50%) including BE (for French part only descriptors are available), FR, HR, UK.

3.4. Ratios and outlier identification

A central component of the data quality process in ETER was a careful analysis of ratios between variables. This involved two main types of checks:

a) whether ratios are in a reasonable range which could be expected from the process considered (for example whether ratio between students and graduates is near the usual length of the curriculum), and

b) whether ratios for individual HEIs and for whole countries are within a reasonable range in the overall distribution. The latter analysis was performed by comparing the distribution of ratios with hypothetic distributions (notably, lognormal) in order to ascertain whether observations are “unlikely” to be generated from the empirically (robustly) estimated distribution. This analysis identified a number of deviant cases, which were then checked manually. Most cases could be explained by specificities of the observed HEI and have been documented within the dataset. Other cases revealed mistakes which were corrected in agreement with NSAs.

3.5. Metadata and comparability problems

In order to highlight problems of comparability between ETER figures across countries, specific metadata have been collected together with quantitative variables. Although the degree of completeness of metadata is lower than the average level of the dataset and information is sometimes incomplete, metadata are an essential resource in order to understand problems highlighted by the quality control of the data, as in many cases revealed data problems are already explained by the providing NSA. In this respect, data quality and metadata analysis are complementary.

Some emerging issues revealed by the metadata are the following:

- Total expenditure is not perfectly comparable for countries which do not include capital expenditures or with a definition of capital expenditures different from the others.
- The breakdown of revenues by categories, although not always recalled in metadata, may hide different classification choices which can have a minor impact on the comparability of figures.
- Minor specificities about inclusion and classification of staff across countries and within countries among HEI categories (typically universities vs. colleges) may impact full comparability.
- Staff data are available in some countries only in headcounts, in others in Full-Time-Equivalents (FTE).
- Classifications of students by new ISCED levels of education are not straightforward in every country. Nevertheless the problem was solved for the majority of cases, with a few exceptions where no disaggregation was possible.
- Similar consideration applies for classification by field of education: in several countries the ISCED-97 classification was used.
- A breakdown of students and graduates by mobility status is not fully comparable among countries, since different criteria to identify mobile students are adopted (residence, place of prior education).
- Information on R&D expenditure per HEI is available only in a subset of countries.
- Metadata can be accessed directly on the ETER website and are sorted by variable and country. Concerning EUMIDA, metadata have been summarized in a written document that can be downloaded together with the data.
3.6. Flags

In order to alert users concerning data and comparability issues, data flags are introduced directly in the dataset (in separate columns alongside the data columns). The dataset includes also short remarks explaining the flag and referring to metadata for full explanations. For example a flag like “d” (definition differs) might be accompanied by the remark that mobile students are counted based on residence (rather than on place of prior education).

The table below provides a full list of flags for the ETER dataset.

**Table 7. List of Flags**

<table>
<thead>
<tr>
<th>Code</th>
<th>Description</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>b</td>
<td>Break in time series</td>
<td>When changes in definitions or data collection procedures imply that the data are not comparable across years. This flag will be relevant in the framework of multi-annual data collection.</td>
</tr>
<tr>
<td>d</td>
<td>Definition differs</td>
<td>Differences in definitions adopted for data collection imply that the value of the marked cells differs significantly from those complying with the ETER methodology.</td>
</tr>
<tr>
<td>i</td>
<td>See metadata</td>
<td>There are specific conditions which imply that the value of a cell should be interpreted in a different way or not directly compared with others.</td>
</tr>
<tr>
<td>ic</td>
<td>Inconsistent</td>
<td>Either when the sum of break down differs from the total or another semantic rule is violated.</td>
</tr>
<tr>
<td>rd</td>
<td>Rounding differences</td>
<td>When a sum of data does not fully correspond to the total because of rounding differences.</td>
</tr>
<tr>
<td>c</td>
<td>Confidential</td>
<td>When data are available, but restricted to public access (this flag is relevant only for user with unrestricted access).</td>
</tr>
<tr>
<td>ms</td>
<td>Missing subcategory</td>
<td>This flag is applied to totals in order to warn users that the total does not include one relevant subcategory (for example total expenditures not including capital expenditures).</td>
</tr>
</tbody>
</table>

3.7. Special codes

Special codes replace blank cells, which are not allowed in the dataset, providing more information on why a numerical value cannot be provided. These special codes largely follow standard conventions by Eurostat.

**Table 8. Special Codes**

<table>
<thead>
<tr>
<th>Code</th>
<th>Description</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>m</td>
<td>Missing</td>
<td>The data is not available.</td>
</tr>
<tr>
<td>a</td>
<td>Not applicable</td>
<td>This variable is not applicable for the specific case. For example, number of PhD students in non-doctorate awarding HEIs is coded in this way.</td>
</tr>
<tr>
<td>x</td>
<td>Included in totals</td>
<td>The data are not available, but it is included in the total. This applies for example when data on revenues from student’s fees are not available, but nevertheless these amounts are included in the total revenues of the HEI.</td>
</tr>
<tr>
<td>xc</td>
<td>Included in another column</td>
<td>The data are not available, but are included in the value for another column. This applies for examples when educational fields cannot be split.</td>
</tr>
<tr>
<td>xr</td>
<td>Included in another row</td>
<td>The data are not available but are included in another row. This happens in the rare case that two HEIs are part of a holding and have a common budget.</td>
</tr>
<tr>
<td>c</td>
<td>Confidential</td>
<td>When data are available, but restricted to public access, this code is displayed in place of the data.</td>
</tr>
<tr>
<td>s</td>
<td>Below threshold</td>
<td>This code is displayed in the public dataset when the count of a cell is so low that data protection issues might arise as individuals can be identified (for example number of students below 3).</td>
</tr>
</tbody>
</table>
4 – Legal issues encountered and access conditions

4.1. Owner of the raw data

National Statistical Agencies and/or Ministries of Research and Higher Education are the owners of most of the raw data in ETER, and in particular, almost all statistical data. The consortium itself has collected descriptors from public sources. Since ETER is under a service contract of the European Commission, the owner of the dataset is the European Commission. The situation concerning EUMIDA is similar, as the dataset is legally owned by the European Commission. There is however no systematic documentation of data sources and thus it is impossible to track ownership of the raw data (even if most probably originates from the NSAs).

4.2. Access to the data and conditions of use

In many countries, raw data reported in ETER are publicly available at the national level. The NSAs in these countries have confirmed this to the ETER consortium. For other countries with restricted access, the NSAs have signed an agreement for data disclosure, and delivered this to the ETER consortium. These agreements allow public access for the largest part of the data, while a few data (mostly financial) are under restricted access for research purposes (under the condition of non-disclosure of individual data points). Access to restricted data is possible under signature of a non-disclosure agreement. Additional datasets and the EUMIDA dataset (as soon as it is available) will be available through the RISIS projects. To this aim, potential users need to register to the RISIS system, accept the RISIS code of conduct and to sign a non-disclosure agreement (which covers both the ETER confidential data and additional data in RISIS).

5 – Technical structure of the ETER database

5.1. Information on the database system

The ETER project provides an infrastructure for data collection, which allows for standardization and systematization of the process. This infrastructure includes:

- Templates for data collection including documentation (e.g. flags and special values as commonly used for EUROSTAT-statistics), which guide national data sources (statistical offices, national authorities, other sources) and country experts addressing and supporting national data sources.
- A master database including an upload interface and documentation of database activities (with time and active person).
- A web application, which allows for individual data exports (downloads of all variables for all countries is also possible, such as downloads of staff data for a specific country).

The specific requirements of the ETER project suggest a centralized web server based collection tool, which will be specified for the needs of the two data collection rounds in ETER. The advantage of such a solution is a closer linkage between data collection, feedback and revision.
(identification of problems and coordinated support) and finally, integration to a raw data set ready for advanced quality, consistency checks and analysis.

The data management tool has been derived from existing tools (imdas/archivis) engaged by JOANNEUM RESEARCH. The existing model of a data management system and centralized data collection was adapted according to the specific needs of the ETER project. Imdas was programmed in Gupta and C# and accesses a relational database, namely an MS SQL Server database. The database is managed via direct and central access by JOANNEUM RESEARCH, which guarantees data security, consistency and quality.

5.2. Technical variable definition

The following table summarizes the set of variables used in ETER and the format the delivered data should have. As the ETER data collection provides special codes, for example missing data, in reality all data are types of “text.” Detailed definitions are provided in the ETER handbook.

**TABLE 9: LIST OF VARIABLES IN DATASET**

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Variables</th>
<th>Format</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identifiers</td>
<td>ETER ID</td>
<td>Text</td>
</tr>
<tr>
<td></td>
<td>National identifier (optional)</td>
<td>Text</td>
</tr>
<tr>
<td></td>
<td>Institution name (in own language)</td>
<td>Text</td>
</tr>
<tr>
<td></td>
<td>English institution name (if available)</td>
<td>Text</td>
</tr>
<tr>
<td></td>
<td>Year</td>
<td>Integer</td>
</tr>
<tr>
<td>Basic institutional descriptors</td>
<td>Country Code</td>
<td>ISO code Nominal</td>
</tr>
<tr>
<td></td>
<td>Legal status</td>
<td>Text</td>
</tr>
<tr>
<td></td>
<td>Institution category, national definition (in own language)</td>
<td>Text</td>
</tr>
<tr>
<td></td>
<td>Institution category, national definition (in English, if available)</td>
<td>Nominal</td>
</tr>
<tr>
<td></td>
<td>Institution category standardized</td>
<td>Binary</td>
</tr>
<tr>
<td></td>
<td>Foreign campus</td>
<td>Integer</td>
</tr>
<tr>
<td></td>
<td>Foundation year</td>
<td>Integer</td>
</tr>
<tr>
<td></td>
<td>Legal status year</td>
<td>Integer</td>
</tr>
<tr>
<td></td>
<td>Ancestor year</td>
<td>Integer</td>
</tr>
<tr>
<td></td>
<td>University hospital</td>
<td>Nominal</td>
</tr>
<tr>
<td></td>
<td>Institutional website</td>
<td>Website</td>
</tr>
<tr>
<td>Geographic information</td>
<td>Region of establishment, NUTS2 code</td>
<td>Text</td>
</tr>
<tr>
<td></td>
<td>Region of establishment, NUTS3 code</td>
<td>Text</td>
</tr>
<tr>
<td></td>
<td>Name of the city</td>
<td>Text</td>
</tr>
<tr>
<td></td>
<td>Postcode</td>
<td>Text</td>
</tr>
<tr>
<td></td>
<td>Multi-site institution</td>
<td>Binary</td>
</tr>
<tr>
<td>Educational activities</td>
<td>Highest degree delivered</td>
<td>Nominal</td>
</tr>
<tr>
<td></td>
<td>Number of enrolled students at ISCED levels 5, 6, 7, by fields of education, gender, citizenship and mobility</td>
<td>Integer</td>
</tr>
<tr>
<td></td>
<td>Number of graduates at ISCED levels 5, 6, 7, by fields of education, gender, citizenship and mobility</td>
<td>Integer</td>
</tr>
<tr>
<td></td>
<td>Distance education institution</td>
<td>Binary</td>
</tr>
<tr>
<td>Research</td>
<td>Research active institution</td>
<td>Binary</td>
</tr>
<tr>
<td></td>
<td>Number of enrolled students at ISCED levels 8, by fields of education, gender, citizenship and mobility</td>
<td>Integer</td>
</tr>
<tr>
<td></td>
<td>Number of graduates at ISCED levels 8 (doctorates), by fields of education, gender, citizenship and mobility</td>
<td>Numeric</td>
</tr>
<tr>
<td></td>
<td>R&amp;D expenditures</td>
<td>Numeric</td>
</tr>
<tr>
<td>Expenditures</td>
<td>Personnel expenditure</td>
<td>Numeric</td>
</tr>
</tbody>
</table>
The ETER IDs in combination with the respective reference year is used as a unique identifier, since each ETER ID can only be found once a year.

Besides data on the institutional level, the database includes metadata information in order to consider institutional and country specific characteristics in the data. Metadata include:

- metadata at the institutional level,
- metadata at the country level.

Country level metadata are structured into metadata for descriptor and quantitative variables, which contain information about:

- the content and deviation from ETER definitions,
- the reference period, and
- data sources.

“In-depth information” contains more detailed additional information about the quantitative variables covered in ETER. Finally, country level data also includes the perimeter descriptions and thus provides information about the higher education landscape in the observed countries.

### 5.3. Description of the Entity Relationship Model

The following figure shows the structure of the database and their first data level, the reference year. After the reference year, data are structured by country. For each country, variables are collected at the institutional level, which also includes a set of institutional metadata in order to provide the possibility of detailed data descriptions for institutions. Additionally, the database provides the opportunity to flag data, e.g. in the case of incomparable data, after quality control, and where flags are available for all quantitative variables. In the case of corresponding sub-categories of variables (e.g. male and female students) the flag marking a statistical footnote will apply for all corresponding sub-categories.
FIGURE 1: STRUCTURE OF THE ETER DATABASE

Country level data

- Year
- Country level data

Institutional level data

- Identifier
- Geographical information
- Basic institutional descriptors

Variable metadata

- Descriptors
- Quantitative variables

Descriptive variables

- Basic institutional descriptors
- Geographical information

Descriptive variables

- Expenditure
- Education – Students
- Revenues
- Education – Graduates
- Research

In-depth information

- Currency
- Exchange rate
- Conversion rate to PPP
- Expenditure
- Staff
- R&D expenditure

Institutional level metadata

- Revenues
- Expenditure
- Staff
- R&D expenditure

Education - Students
- Highest degree delivered
- Students enrolled at ISCED 5, 6, 7 by gender, citizenship, mobility and field of education
- Distance education institution

Research
- Research active institution
- Students enrolled at ISCED 8 by gender, citizenship, mobility and field of education
- Graduates at ISCED 8 by gender, citizenship, mobility and field of education
- R&D expenditure

Research
- Staff
- Personnel
- Non-personnel
- Accounting system of capital expenditure
- Academic staff by gender and citizenship
- Full professors
- Inclusion of PhD students
- Administrative staff

Education - Graduates
- Graduates at ISCED 5, 6, 7 by gender, citizenship, mobility and field of education
5.4. Interfaces for access and to other infrastructures

Completed and validated data collection sheets are imported into the database *imdias pro*, and all collected data are hosted in a secured server environment by JOANNEUM RESEARCH, which is also responsible for data management (cleaning, preparing for upload, uploading, updating) and backup tasks in the process. The existing model of a data management system and decentralized data collection has been adapted according to the specific needs of the ETER project. This will be done by JOANNEUM RESEARCH coordinated with the core partners.

The database also provides the foundation for ETER web application developed from JOANNEUM RESEARCH. The web application enables the user to retrieve data from the whole ETER data set in order to conduct research on micro data of the European higher education sector. The following figure gives an overview about the process starting from collected data until their exploitation in other facilities.

**Figure 2: From collected data to statistical analysis and validation**

The ETER web application includes a short description of the ETER project and the performing consortium members. Starting from the homepage of the ETER web application, three paths are prepared for the user to define an individual query (depending on the information required):

- **path1**: the user wants to get an overview of the included higher education institutions.
- **path2**: the user wants to export demographic information or metadata on the country level, or
- **path3**: the user wants to export data from the ETER micro data set.

**Getting an overview about higher education institutions in ETER**

The selection field “Higher Education Institutions” offers the possibility to have a closer look at the included higher education institutions in the ETER project. The data are prepared by year and country, and by using the plus sign, a list of all included HEIs in a specific country’s perimeter will appear. Using the magnifier symbol leads the user to a detailed view of the variable (year, country or institution), which includes “ETER ID”, “English Institution Name”, “Year”, and “Country Code”.

---

RISIS – First Reporting Period – Strategic Report
Export country level metadata and information about demographic events
By choosing the selection field “Demographic Events & Metadata”, users can gather information about demographic events or country level metadata. The user is forwarded to a platform where the following information can be downloaded:
- an Excel-file where all demographic events for all countries and all institutions since 2009 were collected,
- one Excel-file for every country, where country level metadata for all variables and perimeter descriptions are collected in four unified sheets, and
- an Excel file with four sheets (descriptors, quantitative metadata, in-depth information, perimeter description), where all corresponding metadata are collected and organized by year, country and variable. Using a filter enables the user to choose the desired metadata by year, country and/or metadata variable.

To simplify the handling of institutional data and metadata, the application enables one to switch back directly from the metadata area to the last search by using “Last Search Result”.

Export data from the ETER database
In order to follow path3 the user needs to access “Download ETER Data”, where two different types of access will be provided:
- An open public access, where small numbers and all data, for which public access was restricted by a national statistical authority, are coded.
- A restricted access, where accredited users receive access to the entire dataset for research purposes under the condition that individual data are not disclosed. To this aim, it is envisaged that ETER makes use of the accreditation system which will be established by the EU-FP7 Infrastructure Initiative on Research Infrastructure for the Assessment of Science, Technology and Innovation Policy (RISIS).

In order to follow path3, the user needs to access “ETER Micro Data”, which leads to a mask where the data can be selected by year and/or country. While choosing a year is mandatory to get a search result, the user will get information for all countries at once if she/he does not select a country.

The user can choose to view:
- the entire data set, or
- data sets for one or several specific countries in one or more years.

After choosing the required information, the ETER web application displays the search result, which can be exported by using the selection field “Export Results / create reports.” The following figure shows the result of a search for the year 2011, which leads to all data in the data set for this year.

The result shows all institutions in the data set, which can now be exported to Excel. The export function enables an export of:
- all variables at once, using “Export Full Spreadsheet,”
- a specific group of variables (e.g. staff), or
- a predetermined group or related variables (e.g. all revenues and expenditures).

An export of a specific group includes the basic variables
- “ETER ID”,
- “National Identifier”,
- “Institution Name”,
- “English Institution Name”, and
- “Year” plus all variables assigned to a group.

While the export function “Export Full Spreadsheet” enables the export of all variables included in ETER at once, all data for a specific group of variables (see groups and included variables in chapter Error! Source du renvoi introuvable.) can be exported by selecting their name in the export function. Additionally, four more export possibilities are provided, which cover related variables and present a useful extension for research purposes. The following table will give an overview of these export possibilities and the included variables.
### Table 10: Additional Export Possibilities

<table>
<thead>
<tr>
<th>Export possibility</th>
<th>Variables</th>
</tr>
</thead>
</table>
| **Export All Expenditures and Revenues** | Personnel expenditure  
Non-personnel expenditure  
Capital expenditure  
Expenditure unclassified  
Total expenditure  
Accounting of capital expenditures  
Core budget  
Third party funding  
Private funding  
Tuition fees  
Student fees funding  
Revenue unclassified  
Total revenues  
Research active institution  
R&D Expenditure |
| **Export Student Graduates and Research**| Highest degree delivered  
Number of enrolled students at ISCED levels 5, 6, 7, by fields of education, gender, citizenship and mobility  
Distance education institution  
Number of graduates at ISCED levels 5, 6, 7, by fields of education, gender, citizenship and mobility  
Research active institution  
Number of enrolled students at ISCED level 8 by fields of education, gender, citizenship and mobility  
Number of graduates at ISCED level 8 by fields of education, gender, citizenship and mobility  
R&D Expenditure |
| **Export All Students**                  | Highest degree delivered  
Number of enrolled students at ISCED levels 5, 6, 7, 8 by fields of education, gender, citizenship and mobility |
| **Export All Graduates**                 | Highest degree delivered  
Number of graduates at ISCED levels 5, 6, 7, 8 by fields of education, gender, citizenship and mobility |

The user can access directly from the search mask to the country level metadata.

**Additional functions**

The selection field “Last search result” enables the user to call their last query and presents a quick link to the results if necessary. Subscribed users can change their assigned password by using selection field “Settings” and the function “Change Password.” There are no restrictions for the chosen password.

### 6 – References

PART THREE
NEW DATASET DEVELOPMENT

We have organised the work along 5 problems we consider crucial for a better understanding of research and innovation dynamics and policies: (a) innovation dynamics of firms; (b) types and dynamics of public research organisations; (c) European integration of research activities; (d) PhD and researchers’ careers; (e) repository of policy evaluations. The following sections present the activity reports produced by the coordinators of the corresponding work packages.

1 – INNOVATION DYNAMICS OF FIRMS (WP20)

1 – Objectives

The WP20 follows 3 main objectives:
1. To update and enlarge two existing datasets:
   a. CIB: updating is about the delineation of large firms and the yearly coverage of patents and enlargement is about adding publications to patents for large firms;
   b. VICO: updating information on existing small high tech firms (from 2006 to 2012) and adding the new small high tech firms supported by Venture Capital since 2006.
2. To develop a new dataset dealing with innovation in fast growing mid-sized firms (FGMF).
3. To apply the harmonisation approaches developed in WP8 and WP9.

Thus, in WP20 we aim to extend two existing infrastructures (CIB on large firms and VICO on small high tech firms). Furthermore, we aim at creating a new dataset on FGMFs. We will devote a substantial effort for the definition of the perimeter of the FGMF dataset, the selection of large enough sample to cover geographical and sectoral dimensions, and the strategy adopted for defining the boundaries of groups (as sets of economically bound legal entities – for instance on average one large firm gathers 150 distinct legal entities). Finally, we will apply the different harmonisation approaches developed in WP8 and 9 (on names and geography) to insure a wide inter-connectivity.

2 – Work programme

WP20 aims at developing a robust monitoring device covering the economic and technological development of three pillars of the industrial and technological system: large corporations, FGMF and small high tech firms. The information collected for these three key segments of the productive fabrics will include financial, legal, technological and scientific data extracted respectively from ORBIS, PATSTAT and publication databases. The three new databases resulting from this WP20 will benefit from harmonisation and geographical advance produced in WP 8 & 9.

We grouped activities in three main categories:
1. Preliminary activities;
2. Data collection and harmonization activities;
3. Opening and reporting activities.

“Preliminary activities” are dedicated to substantial preparatory activities concerning the definition of the FGMFs, the structure of the FGMF dataset and the possibility to extend the VICO dataset. It is an essential input for specifying the “Data collection and harmonization activities”, which are devoted to the data collection of legal, financial and innovation information from multiple sources and to the harmonization
according to WP8 and WP9 (see the DOW). Finally, “Opening and reporting activities” comprise all dataset-
specific activities to prepare for opening up (according to Milestones MS1, MS2, MS3 and MS4, see the
DOW) and the preparation of the final version for reports (Deliverables D20.1, D20.2, D20.3, D20.4, D20.5
and D20.6) and the policy brief (D20.7).

3 – Activities engaged (and participants)

“Preliminary activities” started in June 2014 and ended in December 2014. In this category we
performed the following activities (participant in round brackets):

1.1 Definition of the population of European FGMFs: criteria and preliminary analysis (POLIMI).
1.2 Collection of qualitative information on FGMFs (POLIMI).
1.3 Extension of VICO to Israeli firms: exploratory analysis aimed at collecting quantitative information
on VC-backed firms to be integrated in VICO (SNI).
1.4 Exploratory enlargement of CIB from technological (patents) to scientific (publications) information
(UPEMLV and SPRU).

“Data collection and harmonization activities” started in 2015. In this category we are performing the
following activities (participant in round brackets):

2.1 Data collection on legal and financial information for FGMFs (POLIMI).
2.2/2.3 Data collection on legal and financial information for VICO firms and Israeli VC-backed firms
(POLIMI)
2.4 Data collection on legal and financial information for CIB firms (POLIMI).
2.5 Data collection on patents for VICO, CIB and FGMFs (UPEMLV).

4 – Substantive analysis of what has been done and where do the WP stands

Activity 1.1
We defined the criteria for the identification of the population of fast growing mid-sized firms (FGMFs).
Partners agreed that the FGMF Database will include firms that at the beginning of the observation period
(see below) can be classified as “Mid/Intermediary-sized” firms, i.e. firms that belong to the following
categories:

- EUROSTAT medium-sized firms: firms with a number of employees between 50 and 249, and either a
turnover of not exceeding €50 million or a balance sheet total of not exceeding €43 million.
- Entreprise de taille intermédiaire (ETI): firms with a number of employees between 250 and 4999, and
either a turnover of not exceeding €1.5 billion or a balance sheet total of not exceeding €2 billion. A firm
that has less than 250 employees but a turnover of more than €50 million and a balance sheet total of
more than €43 million is also considered an ETI.

The FGMF database will therefore include firms that at the beginning of the observation period have
employees between 50 and 4999, and either turnover lower than €1.5 billion or a balance sheet total of not
exceeding €2 billion. We will measure fast growth in 3 different observation periods: 2008-2011, 2009-2012
and 2010-2013. The observation periods start in 2008, 2009 and 2010, respectively.

As to the data collection process, we proceeded as follows:

- Identification of the potential population of Mid/Intermediary-sized firms by applying the Turnover OR
Balance Sheet criteria:
  - Selection of firms with turnover lower than €1.5 billion OR balance sheet total lower than €2
  - Data source: 2008, 2009, 2010 DVD versions of ORBIS in order to not to lose firms gone
bankrupt/have been acquired.
• Selection of the final population of European Mid/Intermediary-sized firms applying the employees threshold (50 and 4999 employees). For firms for which data on employment are missing, we imputed the employees on the basis of industry, age and accounting information.

Activity 1.2
The original objective of Activity 1.2 was to collect qualitative information from the CEOs of a limited number of fast growing mid-sized firms to define the structure of the FGMF dataset. Partners agreed to involve CEO and key executives of fast grow firms once the FGMF Database will be ready, in order to discuss the main findings/results from the use of the FGMF dataset. However, we defined the list of variables to be included in the FGMF dataset (see the next section).

Activity 1.3
The objective of this activity was to formulate a methodology for consolidating Israeli VC backed firm-level data with the VICO database and to bridge the existing differences between the two types of data sources. Activity 1.3 within WP-20 is a preliminary step for the data collection and data construction stages. The following presents the main tasks that were undertaken within the framework of this activity:
• In-depth study of the VICO database (e.g. variables, level of resolution, data construction).
• Definition of the criteria for selecting Israeli VC backed firms.
• Identifying the possible sources for data collection.
• Formulating a methodology for consolidating Israeli data on VC-backed companies with the VICO database.

Activity 1.4
The objective was to test the feasibility of an enlargement of the CIB from technological (patents) to scientific (publications) information. The tasks performed for this exploratory activity aimed first at testing the appropriate way of querying the Thomson Reuters Web of Science and second at assessing the quality and completeness of extracted information.

Activity 2.1
Using the criteria defined in the Activity 1.1, we selected a list of approximately 67,000 firms that showed high-growth performances (in terms of employment or turnover) in the 3 different observation periods, i.e. 2008-2011, 2009-2012 and 2010-2013. Starting from this list, we are currently collecting accounting information and ownership data for these firms.

Activity 2.2. and Activity 2.3
Originally SNI was in charge to collect data on VC-backed firms in Israel (Activity 2.2) and POLIMI to collect data on VC-backed firms in Europe (Activity 2.3). However, WP20 partners decided to perform these two activities together, by POLIMI, as to avoid the problem related to the integration from multiple sources and the cleaning of data. The objective of these activities was first to define a list of firms to be included in VICO update. To this aim we resorted to three different public and commercial datasets on VC investments in both Europe and Israel: Zephyr, Thompson One and Crunchbase. We thus define a sole list of VC-backed firms by merging these three different sources. We are now collecting financial data for the list of firms included in VICO update and for Israeli VC-backed firms.

Activity 2.4
The objective of this activity was to define a list of firms to be included in CIB update. To this aim we started from two different public datasets, firstly, on industrial R&D investment (yearly IPTS Industrial R&D Investment Scoreboard, from 2008 to 2014) and, secondly, on top applicants at WIPO (yearly WIPO list on applicants with more than 9 PCT applications, from 2008 to 2014). This drove to a large extension of possible candidates (from 2000 to over 6000).
In a second step, these firms should be matched (when possible) with the Orbis database in order to identify, as a first step, the ultimate owner (with more than 50.01% ownership) for all these candidates and, as a second step, the list of subsidiaries (with the 50.01% of ownership criteria) that can be considered as included in the groups’ perimeter.

3 – Outputs

**Activity 1.1**
List of Mid/Intermediary-sized firms from ORBIS.
Preliminary statistics on fast growing firms by applying different indicators of growth.

**Activity 1.2**
List of variables to be included in the FGMF Database. Variables cover the following topics:
- General company information: age, industry, address.
- Accounting information.
- Market presence: international presence, number of sectors (primary and secondary) in which the firm operates.
- Innovation: patents and publications.
- Ownership data.

**Activity 1.3**
A report describing the mechanism and the methodology for collecting and consolidating Israeli VC backed data with the VICO database was presented in the RISIS group meeting held in Milan on October, 22 2014.

**Activity 1.4.**
Scientific publications have been extracted for firms belonging to two science intensive industrial sectors (first chemistry; second pharmaceutical and biotechnology) over a 20 year period of time (1991-2010). A similar pilot exercise has been carried out for a subsample of firms included in VICO dataset. This test can be considered as a proof of concept according to the criteria of technical feasibility: the WoS tag “organisation” was efficient for retrieving publications for industrial firms. Moreover this test has validated the interest of looking for publications authored not only by researchers belonging to the GUOs (global ultimate owner) but as well by researchers belonging to subsidiaries of these GUOs. Nevertheless, the volume of publications extracted (in the range of 300 000 only for the two above-mentioned sectors) and the huge amount of false negatives produced when using a simple query raise the question of the opportunity/necessity of developing an automatic filtering; which has not been implemented in this phase of test, based on a manual check. This automatic filtering could be developed adapting the process used for the large-scale patents’ extraction: discarding stop words in firms’ names, using context dependant restrictions (geographical, disciplinary…). But these operations require accessing an off-line version of a scientific publications database, either Thomson Reuters Web of Science or Elsevier Scopus. The opportunity of such an access and the corresponding negotiations with scientific databases producers is being investigated within the framework of the RISIS project.

**Activity 2.1.**
Accounting and ownership data for European FGMFs

**Activity 2.2 and 2.3.**
List of VC-backed firms in both Europe and Israel
Data on VC investments and principal indicators (economic and financial) of annual reports

RISIS – First Reporting Period – Strategic Report
Activity 2.4.
After eliminating the non ambiguous duplicates and discarding the non industrial actors, the two initial lists extracted from the Industrial R&D Scoreboard and from the top applicants at WIPO, provided a first set of candidates which are considered, over a seven year period, as top industrial technology actors considering the amount of R&D investment or the proprietary technology produced. The first round of harmonization carried out (combining manual check and the use of Google refine) produced a list of 5871 candidates, which are currently (summer 2015) matched with the Orbis financial database. These two steps have turned out to be far more problematic than expected and drove UPEM to allocate complementary resources well above what was planned by the project.

4 – Challenges

The first challenge is related to the definition of the perimeter of the FGMF. Specifically, in the next months we intend to develop a strategy for defining the boundaries of groups (as sets of economically bound legal entities – for instance on average one large firm gathers 150 distinct legal entities), using the ownership information that we are currently collecting with Activity 2.1. Another issue is related to the coordination of the data collection for patents, publications, and accounting information for the VICO & CIB databases. Partners agreed on a centralized data collection process for patents (CIB), accounting information (POLIMI) and publications (Us).

5 – Programme for next months (with milestones anticipated)

Next months will be devoted to complete “Data collection and harmonization activities”. First we conclude the following activities:
- Data collection on legal and financial information for FGMFs
- Data collection on legal and financial information for VICO firms and Israeli VC-backed firms
- Data collection on legal and financial information for CIB firms
- Data collection on patents for VICO, CIB and FGMFs.
More over, next months of 2015 will be devoted to the other two activities included in “Data collection and harmonization activities”:
- Data collection on scientific publications for VICO, CIB and FGMFs (which requires a different approach than the one initially envisaged)
- Harmonization of geographical information and names for VICO, CIB and FGMFs according to WP8 and WP9
The next period should also see the achievement of two critical milestones (according to the DOW):
- MS1 Opening of enlarged CIB;
- MS2 Opening of enlarged VICO.
The actors to be analyzed in the context of this project are Public Research Organizations (PROs) which are a mixed type of organizations which aim is generating fundamental knowledge, providing R&D support to business and advice to public authorities (OECD, 2014).

The aim of this WP is to contribute to the advancement of the study of PROs as R&D organizations and contextualize the evolution of the different organizational models, the current framework of their activity and also the identification of different typologies of centers. Besides, the evaluation of their performance that is becoming an important topic in a context of budget cuts and economic competitiveness will be also reviewed.

Public sector research organizations have received less attention than universities and this is why we need a systematic study about the role played by these institutions, the evolution that they have experienced during the last decades, and their performance in a changing context. Then the main objective of WP 23 is to close this gap in the literature analyzing systematically PROs through the construction of a database in the EU that allows for monitoring, analyzing and studying the dynamic of PROs in the research system. Doing this requires necessarily:

1. Developing a **conceptual framework** for the study of PROs linked to key generic policy questions.
2. Establishing a comprehensive **typology** of PROs.
3. Identifying, developing and defining relevant **indicators** and data to gather, taking into account the different types of PROs previously identified.

These three steps constitute **Task 1 and Task 2** of the initial Work Plan for the WP 23. Both of them are included in **Activity 1**: Conceptual development which output is a report on the theoretical framework and a set of indicators including the methodology for data collection strategy.

The following box summarizes the planned work and what has been made so far until the end of June 2015.

### RISIS– WP 23 Activity 1: Conceptual development

<table>
<thead>
<tr>
<th>TASK 1</th>
<th>AIM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Definition of a conceptual framework including a set of criteria enabling to identify PROs and to classify them according to different relevant characteristics. Identification of an initial list of PROs.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SUB-TASKS</th>
<th>Stage</th>
<th>Main results</th>
</tr>
</thead>
</table>
| Literature review of generic policy issues related to PROs | Completed | - Positioning PROs in historical perspective and under competing policy paradigms.  
- Changes and evolution in PROs governance and structure: economic competitiveness; New public management; Hybridization processes; Funding cuts |
| Review of previous studies proposing typologies of PROs | Completed | 1. Literature about teams or research labs  
2. Laboratories as organizations |
| Development of PRO institutional framework for analysis, boundaries and definitions | Completed | - Critical dimensions for understanding the structure and change through ideal types: mission(s), governance, autonomy and authority structures, external links.  
- Theoretical attributes for the classification of different public research organizations  
**Definition**: Legally independent institutions, either founded under private or public law i.e. regardless their legal form, established with the main goal of providing basic or applied research, technology and innovation services to other agents. PROs will be somehow linked to those other actors (firms, universities, public authorities or
society) through several channels as training, publication production, technology transfer, cooperation in research or advisory services. All with not-for profit purposes and partially oriented to overcome market failures associated with technological and industrial development. These organizations will be influenced to a greater or lesser extent by the public sector either through their financial structure, their ownership or the type of mission and vision established in their statutes.

### Development of typology of PRO

<table>
<thead>
<tr>
<th>Development of typology of PRO</th>
<th>Completed</th>
</tr>
</thead>
</table>

Four ideal types of centers that mainly in terms of the level of governmental influence (funding and ownership) and the type of research activities they do (mission).

1. Mission oriented centers (MOC)
2. Public research centers and councils (PRC)
3. Research technology organizations (RTO)
4. Independent research institutes (IRI) and Hybrids

### Preliminary selection of PROs for the pilot exercise

<table>
<thead>
<tr>
<th>Preliminary selection of PROs for the pilot exercise</th>
<th>Completed</th>
</tr>
</thead>
</table>

For the purpose of testing the feasibility of a database a series of cases have been identified and selected, mainly based on three different criteria: location, type of center and size in terms of number of publications and patent applications.

National and regional research centers (not European or international) independent from universities will be included in the sample, keeping in mind the diversity of the types.

Criteria of including bigger cases (in terms of outputs identifies) have been also taken into account.

Based on the composition of the team, and following the three aforementioned criteria, a sample of research centers in Austria, France, Germany, Italy, Netherlands, Norway, Poland and Spain have been selected; this analysis allows to cover economies in Northern, Western, Southern and Eastern Europe. The sample from Spain is bigger but much more varied in compositions of types, to deal with the practical problems of database construction in deep.

As conclusion we expect to generate a database of the most important research centers, regarding the size, combining these two different sources (patents and publications) in order to avoid biases towards some type of institutions, i.e. either closer to basic research or industrial development. Each one of the selected centers has been classified taxonomically in MOCs, PRCs, RTOs and IRIs.

**Up to now the database of cases selected in the 8 countries includes 216 PROs**

### Task 2

<table>
<thead>
<tr>
<th>AIM</th>
<th>SUB-TASKS</th>
<th>Stage</th>
<th>Main Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Definition of the relevant PRO indicator and data to gather, taking into account typologies</td>
<td>Systematic review of previously</td>
<td>Completed</td>
<td>Data about funding sources or availability of human resources as proxies of inputs, and data about</td>
</tr>
</tbody>
</table>

**DELIVERABLE 1.**

All the previous work has been done. Information just needs to be organized in the deliverable.
| proposed and used PRO indicators | publications or patents as outputs of the research activity are the most used in the literature background on the evaluation of R&D performance in PROs. However, for the purpose of this report, also structures and process variables have been considered. |
| Systematic review of guidelines of indicators for universities as relevant research actors in order to see similarities and differences with PROs | Completed |
| **DELIVERABLE 2.** | Review of ETER project and EUMIDA project in order to establish new metrics and not to take into account some of them, as those ones related to undergraduate students. |
| Elaboration of methodological guidelines: concepts, measurements, indicators and data collection strategies. | In progress |

The WP23 team will meet on September 30th, 2015, and discuss the draft documents and approve them before dissemination.

### 3 – FOSTERING EUROPEAN INTEGRATION (WP22)

#### 1 – Objectives of WP as of Description of Work

European Framework Programs (EUFPs) and Joint programmes (JP) are key components of the European research funding system. They provide evidence of the government rationales funding R&D, where R&D is the main goal and mission; they allow understanding the role that funding agencies, and other organizations executing the 'programming function', are playing at the research policy level.

EUFPs and joint programming show the national and European strategies toward collaboration, coordination and integration, the level of concentration of funding on a limited group of actors and a limited number of programs, the geography of collaborations in research activities and in research policy design. Programmes and projects can also allow studying the strengths and the dynamics of the different research fields in national communities, as well as the growing collaborative commitment to European R&D goals.
EUPRO and JOREP can thus play an essential role supplying evidence that can be used for the analysis of the ERA dynamics and Europeanization processes. At the moment EUPRO focuses on EU projects that have been funded and their beneficiaries; JOREP focuses on the amounts spent for programmes and their organizational characteristics. Thus they represent on the one hand, the policy rationales of joint public project funding at European and national level, and on the other hand the research projects effectively funded at European level and the collaborations that have been generated among scholars. Maintaining the facilities alive, improving their coverage and develop an interoperability between the two datasets are essential activities in order to investigate Europeanization.

2 – Work programme Description

WP 22 foresees four main types of activities:

**Activity 1**
*Selection of the programs to be included in the datasets:* this activity to be developed in the first six months of the WP22 (from June 2014 to December 2014) wants to screen and identify the most significant multi-lateral and joint programmes, as a sound basis for selecting those that have to be considered in WP22 further work. In the end, it will only be a limited number of larger joint programmes that make a difference for joint research in Europe, and we need to concentrate on them when updating of JOREP and enlarging activities of both EUPRO and JOREP are concerned. Within this activity the issue of feasibility is an important one for testing the concrete possibilities to develop the planned activities, checking the best procedures for updating and for the enlargement (Activity 2 and 3).

**Activity 2: Updating:**
This activity wants to maintain a sound coverage as to the year of funding (2013). A substantial amount of work to be done on EU-PRO, although the data set coverage is at 2012. The operators of the CORDIS website do not document changes of already existent project data (new participants, additional or reduced budgets, extension of project duration, etc.). Therefore, an update of EUPRO in 2013 means in fact a total download of the complete CORDIS data to consider all changes conducted since the last update. Updating is of course a very consistent amount of work for JOREP where the existing data cover information until 2009. Updating in the case of JOREP is an activity that overlaps the important phase of re-engineering of the dataset to be developed under the WP6. In both cases (EU-PRO and JOREP) updating 2013 will be developed using web-based resources as far as possible. The activity shall takes 6 months (from January 2015 to June 2015).

**Activity 3: Enlargement:**
There is a mutual interest of EU-PRO and JOREP to enlarge the coverage of the infrastructures including the Joint Programs in the former and the EUFPs in the latter. This objective would allow to having a common base for the two datasets. In the case of JOREP we can also foreseen the possibility to enlarge as far as the EU funded programs are concerned, the country coverage of the dataset, which is actually limited to eleven European countries.

- Other specific activities related to JOREP are:
  - Test the possibility to collect data on Programmes generated by Funding Agencies in countries not yet included by JOREP
  - Disaggregate the existing data on large programs (e.g. EUREKA)
  - Integrate the PROs acting as funding agencies in different countries

- Other specific activities related to EU-PRO are:
  - Test the possibilities to collect data at the project level on Joint Programmes covered by JOREP
  - Integrate data of selected Joint Programmes (e.g. EUREKA, COST)

The effective possibilities to enlarge the datasets involved in WP22 depend on the results of the Activity 1. It means that Activity 3 could be revised accordingly. In any case Activity 3 is supposed to take one year, starting from June 2015.
Activity 4: Exploiting complementarities:
EUPRO and JOREP datasets are suitable to improve the possibilities for investigating Europeanization from different perspectives. In order to achieve this goal the AIT and CNR teams will develop strategies based on a step-by-step approach, in order to allow the possibility of complement the results of the two datasets. An entry point can be the ID on organizations participating in EUFPs, which can be integrated in the datasets. Other activities shall be related to the harmonization of the classifications used in the datasets. This activity is supposed to be developed from 2016 on.

The activities mentioned have been developed both separately by AIT and CNR (Activity 2), and with a joint collaboration, especially during the Activities 1 and 3, in order to facilitate the achievement of the results.

We developed a coordination of the efforts for gathering the information, and a shared organization for collecting data mainly using web-based resources.

Collaboration with USI was also developed, since it is useful for the matching of EUPRO with ETER/EUMIDA and for the enlargement of the JOREP dataset.

Moreover, several items were discussed (ID of programs, unit of analysis, classifications, coverage, etc.) in order to decide the best strategy for the exploitation of complementarities. Analysis of the complementarities between the two datasets is an issue also to be supplied to the on-site visitors and to be integrated in the training.

Exploring possibilities of integrated uses of the databases with other facilities is another possible development to be foreseen in a more mature phase of RISIS; in this respect, the most interesting datasets are: ETER/EUMIDA and EUPRO as to the possibility to matching project collaboration with the data of the organizations; bibliometric resources, CIB, Nano and VICO in order to find out the research outputs of programmes and projects.

3 – Activities engaged and participants

The feasibility study was entirely developed by IRCRES-CNR (former CERIS-CNR) and foresaw:

- Selection of programmes with a high data availability (basically the European ones) and draft analysis in order to check their significance for research purposes on Europeanization and Era dynamics
- Creation of a strategy to retrieve data on programmes with low data availability. A pilot can be envisaged based on the most relevant programmes already included in the JOREP dataset, in order to check data availability and reliability;
- Test possibilities for enlargement of the dataset in countries not already included (one large country and one small country can be select for this purpose).
- Since the availability of data on joint R&D programmes is scattered in multiple sources and not generally high (with the only exception of the European programmes), the study is determining the borders for the future expansion of the infrastructure, specifically in terms of selection of new units of observation and geographical coverage

Three sub-activity within the feasibility study were developed:
1. Selection of programmes with a high data availability in EU28 and Associated countries (basically the European ones);
2. Creation of a strategy to retrieve data on programmes with low data availability (basically the bilateral ones);
3. Test possibilities for enlargement of the dataset in countries not already included.

4 – Outputs

The sub-activity 1 is completed. Around 69 new European–based Joint programmes were selected for the inclusion in the dataset, beyond the 44 already included and to be updated, with 211 Research Funding Organizations involved in the funding and management of the programmes. (Tab.1) The data collection of the descriptors is completed; data on funding is completed for the 61% of the programmes.
According to the new relational structure of JOREP, the inclusion of these new programmes is revealing useful to undertake utility and feasibility controls for:

(i) The current list of JOREP variables in order to limit data redundancy to the necessary for the integration of tables;
(ii) The solutions adopted to trace the demography of the programmes through the new fields inserted in order to register transformations of the joint R&D programmes;
(iii) The disaggregation or unpacking data on large programmes to obtain more specific information.

Tab. 1 – European-based Joint programmes collected in the new version of JOREP (baseline 2013-2014)

<table>
<thead>
<tr>
<th>Programme Type</th>
<th>Baseline 2009</th>
<th>Baseline 2013-2014</th>
</tr>
</thead>
<tbody>
<tr>
<td>ERA-Net/ERA-Net + actions</td>
<td>33</td>
<td>50</td>
</tr>
<tr>
<td>Joint Programming (JP)</td>
<td>-</td>
<td>9</td>
</tr>
<tr>
<td>Joint Technology Initiatives (JTIs)</td>
<td>2</td>
<td>3*</td>
</tr>
<tr>
<td>Article 185 of the EC Treaty</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>European Social Fund (ESF)</td>
<td>2</td>
<td>-</td>
</tr>
<tr>
<td>COST + EUREKA + ESA</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td><strong>44</strong></td>
<td><strong>69</strong></td>
</tr>
</tbody>
</table>

*The Joint Technology Initiative ECSEL merged the ARTEMIS embedded systems JTI and the ENIAC nanoelectronics JTI in 2013, so the three programmes are all included in JOREP due to the launch of calls of ARTEMIS and ENIAC on 2013 and ECSEL on 2014.

Tab. 2 - Participation to the European-based Joint programmes included in the new version of JOREP (baseline 2013-2014)

<table>
<thead>
<tr>
<th>Country Name</th>
<th>Number of programmes signed by the Funding Agency</th>
<th>Number of programmes financed by the Funding Agency***</th>
</tr>
</thead>
<tbody>
<tr>
<td>Austria</td>
<td>44</td>
<td>31</td>
</tr>
<tr>
<td>Belgium</td>
<td>53</td>
<td>42</td>
</tr>
<tr>
<td>Bulgaria</td>
<td>10</td>
<td>4</td>
</tr>
<tr>
<td>Croatia</td>
<td>10</td>
<td>5</td>
</tr>
<tr>
<td>Cyprus</td>
<td>17</td>
<td>13</td>
</tr>
<tr>
<td>Czech Republic</td>
<td>18</td>
<td>9</td>
</tr>
<tr>
<td>Denmark</td>
<td>40</td>
<td>29</td>
</tr>
<tr>
<td>Estonia</td>
<td>28</td>
<td>14</td>
</tr>
<tr>
<td>Finland</td>
<td>44</td>
<td>28</td>
</tr>
<tr>
<td>France</td>
<td>66</td>
<td>47</td>
</tr>
<tr>
<td>Germany</td>
<td>61</td>
<td>41</td>
</tr>
<tr>
<td>Greece</td>
<td>30</td>
<td>17</td>
</tr>
<tr>
<td>Hungary</td>
<td>22</td>
<td>14</td>
</tr>
<tr>
<td>Ireland</td>
<td>32</td>
<td>22</td>
</tr>
<tr>
<td>Israel</td>
<td>27</td>
<td>22</td>
</tr>
<tr>
<td>Italy</td>
<td>50</td>
<td>34</td>
</tr>
<tr>
<td>Latvia</td>
<td>20</td>
<td>18</td>
</tr>
<tr>
<td>Lithuania</td>
<td>14</td>
<td>10</td>
</tr>
<tr>
<td>Luxembourg</td>
<td>14</td>
<td>12</td>
</tr>
<tr>
<td>Malta</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>Netherlands</td>
<td>51</td>
<td>31</td>
</tr>
<tr>
<td>Norway</td>
<td>41</td>
<td>39</td>
</tr>
</tbody>
</table>
The sub-activity 2 identified the most significant bilateral R&D programmes, as a sound basis for selecting those that have to be considered in WP22 further work. In fact, only a limited number of large bilateral programmes (Nordforks, D-A-CH, LAAs) that make a difference for joint research in Europe are being included in the JOREP dataset. The selection of bilateral programmes is still in progress and shall be completed by Autumn 2015. At that time a new version of JO REP (JOREP 2) can be released and prepared for opening.

Concerning the sub-activity 3, some exploratory checks during the sub-activity 1 displayed a variety of situations regarding the availability of data on country participation to joint R&D programmes. These checks constitutes the first step to test possibilities for the enlargement of the dataset in countries not already included in the dataset, even if definitive decisions can be made only after the sub-activity 2 on bilateral programmes is completed.

The variable geometry of the joint programmes at European level is confirmed. Moreover the inclusion of programmes beyond the JOREP data collection (mainly JPI and JTI) provide more empirical evidences for analysis on the ERA dynamics

As to EUPRO, the activity developed under WP22 improve the technical robustness and usability of the database within RISIS in terms of updating the data set, implementing a new Entity Relationship Model (ERM) and enlarging the coverage of EUPRO by adding project information of different funding programmes besides EU FP. The activities developed were:

1) Updating EUPRO maintaining a complete coverage as to the year of funding (2013). Since the operators of CORDIS do not document changes of already existent project data, an update of EUPRO in 2013 meant in fact a total download of the complete CORDIS data to consider all changes conducted since the last update. This activity implies the standardization of organizations names and the assignment of organisation types and NUTS2-regions.

2) The implementation of a new ERM will build up the essential requisites for linking different data sets within RISIS. It implies to implement new table of unique organisations (registry in EUPRO), table ‘participations’ intended to link tables ‘projects’ and ‘organisations’, giving participation-specific characteristics, and add variables and classifications, e.g.: supra-national classification of organisations, geographical clustering, identification of public research organisations and large firms.

3) The enlargement of EUPRO focuses in a first step on the integration of project data of the programmes EUREKA (approx. 4,000 projects) and COST (approx. 1,000 projects).

5 – Challenges

- The capability to cover all the countries EU28 the most important bilateral existing at European level is the main result we want to achieve as to the data collection.
- The distinction of funding for beneficiaries proved to be not feasible, or feasible only for some very large programmes.
- Developing the Activity 4: Exploiting complementarities: EUPRO and JOREP datasets to improve the possibilities for investigating Europeanization from different perspectives (see the program for the next years, Annex 1).
6 – Programme for next year and milestones

The whole activity of the WP22 has been reworked according to the outputs achieved in the first period of activity. The new program is outlined in the Annex 1, which also points out the research objectives to study the Europeanization of the research activities, to demonstrate the added value of combining different research infrastructures, namely JOREP and EUPRO with others like ETER and Bibliometrics.

Annex – The dynamics of the Europeanisation of research activities: New insights from RISIS research infrastructures

Background

The development of the European Research Area (ERA) has consolidated over the past few years, in particular in terms of the instruments that have been put in place to foster transnational cooperation. However, there is a need to better understand what effects the instruments have on the actual research activities in Europe, their patterns and orientation. Ultimately, the ERA must demonstrate to deliver a value added in terms of the research conducted, if it is pursued as a main pillar of European policy.

According to Barre et al (2013), there are three important functions that ERA should fulfil, and according to which the evolution of ERA should be monitored, namely orientation, programming and performance of research. Two of these dimensions – programming and performance - can be usefully addressed by drawing on RISIS datasets. The third dimension would require looking into key policy strategies and documents (which is currently not foreseen in RISIS, although this would be possible using semantic or content analysis of policy strategies).

These functions can be performed in different modes of relationships, namely either by simple juxtaposing national policies and programmes, by coordinating national policies and programmes across Member States, or by integrating policies and programmes at European level (Barré et al 2013, Lepori, Reale, Laredo 2013).

Functions and modes of relationship provide a matrix-type framework that allows for a differentiated mapping and monitoring of the evolution of ERA in structural and organisational terms. Funding data, could then be used to depict the levels of resource mobilization and concentration in different years, while profiles of the institutional frameworks and policy instruments, allow an interpretation of modes of relationships.

However, in order to fully exploit the potential of analysing these patterns, they need to be tied to further indicators at organisational level. There are various types of such indicators that would allow meaningful interpretation. Publications and citations, for instance, can serve as performance indicators of research performing organisations. The projects they conduct, whatever the source of funding, provide insights into what research performing organisations actually engage in. As regards programming, either by funding agencies or also by PROs themselves, an analysis of their priorities and funding volumes in different domains could be informative in order to assess the extent to which the programming function in ERA has changed.

Key questions – scientific and policy-oriented

Against the background of this general approach, it would be possible to better understand the patterns of Europeanisation in research activities, for instance in terms of:

- The pace of change and the variable geometries emerging. These characteristics of Europeanization could be analysed in more depth, for instance in sectorally differentiated terms.

- The need and interest to include other countries beyond the Western European ones in the process of integration and coordination and ask for an in depth analysis of barriers and constraints that impede the countries to move beyond cooperation/juxtaposition;

- The need to find good evidence about the evolution of Europeanization in research activities in the actors’ strategies (national and supra-national policy actors, funding actors and performers).

Of particular relevance is yet another important policy issue linked to the ERA dynamics, namely the geography of research in Europe. Literature points out the presence of a core Europe, differentiated according to disciplines and fields, with a progressive polarization around some nodes, which tend to become more and more attractive and mutually exclusive; thus, participation in European transnational initiatives asks for the re-organization of research structures and policies in order to exploit the opportunities coming from transnational initiatives.

From the point of view of the policy actors (especially at the EU level) the challenge of the European research geography is not yet fully perceived: despite the political rhetoric and fairness preoccupations,
funding is almost exclusively aimed at supporting research collaborations between excellent actors in few research fields. Assuming the lack of participation of groups and countries as a result of failure of research capabilities (institutions, framework conditions, etc.), the analysis of the dynamics of Europeanization should also investigate how far funding instruments provide opportunities to actors with less research capabilities.

The objective is to demonstrate how to address the organizational dimension of Europeanization in the knowledge production using the RISIS infrastructure.

The work should focus on three main research questions:

- What is the importance of trans-national collaborations of research performers in different research fields and why they maintain the cross-border relationships along the time (what do research performers do)?
- What kind of research projects do research performers use for their trans-national research activities whatever the funding come from (what do research performers engage to)?
- What programmes RFOs (Research Funding Organisations) and UPRO (Umbrella Public Research Organisations) put into action, what are the organizational features of those programmes and their relationships with other funding actors at European level (what research funding organizations and large umbrella PROs do)?

This suggests a range of possible hypotheses, which are organised here in three groups, with the latter being decomposed in two sub-groups:

- We witness an overall polarisation in a limited ‘set of nodes’ that: (i) have intensive linkages with one another; and (ii) furthermore aggregate other places around them. The nodes tend to become more and more attractive and mutual exclusive, following patterns of preferential attachment, producing Matthew effects, and lock in mechanisms;
- The geography of nodes varies between fields either because of the organizational characteristics of the fields (e.g. the presence of key specialized actors, both funders and performers at national level) or because of specific coordination instruments (e.g. specialized research infrastructures) or because of the effect of cognitive and spatial proximity;
- The degree (intensity and coverage) of Europeanisation depends on the combination of two factors:
  o The relative importance of project-based funding available in the field, and the variety of instruments (e.g. presence or not of federal-type project-based funding; existence number and size of trans-border funding programs; degree of extension (in term of participating ‘spaces’ and their level of engagement) of trans-border programs
  o The structure of the organisational field (active presence of PROs, role of major infrastructures, existence or not of leading research universities)

Data infrastructure: The RISIS contribution

RISIS has two datasets, JOREP (run by CNR) and EUPRO (run by AIT) dealing with ERA dynamics. The former covered transnational research programmes funded by the EU and by Member states in eleven European countries. Actually the JOREP dataset includes (baseline 2014) all the EU transnational research programmes in the EU28 countries, plus CH and NO. JOREP allows analyses of the modes of the ERA dynamics (how functions are developed, what kind of relationships have been built up) using also network analysis are likely to show the dynamics of Europeanization in what funding agencies (including large PROs) do. In order to make JOREP fully effective for analysing the above-mentioned hypotheses, it needs to be extended to EU Framework Programmes and Horizon 2020 in order to complete the picture of the transnational research programmes.

EUPRO relates to the projects funded under the European Framework Programmes starting from the very beginning of the policy instruments. It allows depicting participations and projects of the programmes identifying the actors involved as performer of the ERA dynamics. Also in this case functions and relationships of the different performers can be investigated, providing evidences on the research performers. In order to improve the coverage of EUPRO, other transnational research projects coming from schemes not linked to EU Framework Programmes need to be integrated.

In addition, a new datasets is currently being developed by RISIS partners (CNR, USI, AIT) on Public Research Funding, the so-called PREF database, which – in the future - might complete the picture of the juxtaposed, coordinated and integrated research programming and performance. Through the standardised register of organisational entities, all three datasets can be linked to various performance indicators (publications, patents, size, no. of employees, etc.). RISIS can supply resources to work for improving the two datasets exploiting their complementarities, and – as a mid-term option – integrate PREF in the portfolio of RISIS datasets. Given the involvement of RISIS
partners in PREF development, matters of harmonisation with RISIS standards can be taken into account already during the development of PREF. In the longer term, semantic analysis (e.g. using project abstracts) is a possibility to keep in mind, in order to better investigate the substantive features of ERA development. It could be an option to draw on established high-level categories, but frame them a dual issue of 'sub-fields' and 'inter-fields'. In sum, RISIS is not directed to support the maintenance of the datasets, but it is specifically geared towards demonstrating the added value of using the infrastructures to address key research policy questions.

**Direction of work**

A clear starting point is that although an exploration of new issues generating new knowledge on Europeanization dynamics can be developed, it is neither possible to be exhaustive nor to cover all the existing dynamics at European level. A feasible objective is trying to shed light on not yet covered side of Europeanization, looking at the most important actors and funding instruments (programmes and projects) in key emerging research fields. A second point is to follow a two-step approach, where the first step is the data building, and the second is the development of key indicators to perform the analysis - although one should have in mind what kind of exploration data should allow. A third point is the need to be selective addressing specific fields/domains/sectors in order to explore the dynamics of Europeanization in a more concrete way. Thus the choice is not to focus on policy dimensions (integration, excellence, SGC or KET) rather to look at dynamics in action focusing on emerging fields like for instance nanotechnology, biotechnology, or environmental sustainability. Of course, a central issue is linked to the definition of the fields/domains/areas we want to organise the analysis around (think of nano which is lost in normal classifications under chemistry, physics and biotechnology/molecular biology) and of the geographic entities to analyse (in this regard, ongoing work in WP 9 should be built upon). These delimitations are critical to delivering meaningful results and should be given preference to any ‘policy-based’ categories. Overall, the aim of the research effort on databases should be oriented toward:

a) Maintaining the capability of the infrastructures to monitor the modes of relationships and functions performed, and providing a periodical update. This effort would be translated in different versions of the dataset that can be open to the public exploitation, and it is based on the CNR and AIT own activities;

b) Enlarge the datasets to improve the coverage of transnational research -programmes (the funding instruments) and projects (the performers implementation of the funding instruments);

c) Combining the analysis of the datasets with data useful to the analysis of the Europeanization of research activities coming from other statistical resources on R&DI (e.g. bibliometric databases), or matching with other RISIS datasets on research organizations (e.g ETER).

---

4 – INTEGRATING FRAMEWORK AND DATASET FOR ANALYSING RESEARCHER CAREERS (WP24)

1 – Objectives of WP as of Description of Work

As outlined in the overall work plan for WP24 - Integrating framework and dataset for analysing researcher careers, which was approved by the RISIS Facility Coordination Board in July 2014 the objective of WP24 is to develop a framework which facilitates the systematic transnational comparison of data by harmonising and integrating available datasets. The ambition of this research activity thus is to develop a conceptual framework and an operational database to bring together disparate information sources. We strive for the three major goals: a) enabling access to multiple local datasets that are currently only used by their local developers; (b) developing a theoretical framework that allows to identify the need
for certain data on different levels of aggregation and to formulate research hypotheses that can be tested based on the data at hand and (c) initiating further comparative as well as research and empirical inquiries on research careers at the EU level (including the role of mobility).

2 – Work programme Description

In a first step WP24 aims at developing and proposing a conceptual framework for the study of researchers’ careers. The ambition is to systematically reflect upon available theories and empirical developments leading to the proposal of a conceptual framework that will identify: i) main career types of PhD holders; ii) key milestones in career development of each type; iii) main factors affecting the career decision-making and development process, including personal, organizational, disciplinary, job market and systemic factors. The work will be structured around a systematic review and meta-analysis of the relevant literature and around a review of institutional frameworks shaping research careers in different countries. In the course of developing this framework we will select data sources that provide a sound empirical basis in order to analyse career-related decision making processes and career trajectories. The theoretical framework of researcher careers will serve as the structural basis of a platform that allows the incorporation of these existing data sources. To a degree possible, it will provide access to these data sources or – where enabling direct access remains impossible due to IPR and confidentiality issues - offer structured information about the data sources. As a starting point, datasets compiled by partners involved in this WP – IFQ, NIFU and CSIC – will be incorporated in the platform. In a second step anonymised micro-data will be supplemented by official data providing information on the relevant national research systems (e.g. data from statistical offices, data collections as provided by ERAWATCH\(^{14}\) and the MORE projects\(^{15}\) can offer some orientation in this regard) as well as information regarding the institutional and organisational level wherever possible without conflicting with data privacy. Documentation files will be developed providing background information and guidance on comparability, options and limitations for analysis and interpretation.

WP24 is organized in four tasks.

- Task 1 mainly focuses on the elaboration of a systematic framework for researchers’ careers (Activity 1) and, as outlined above, comprises a) a systematic review and meta-analysis of the relevant literature and b) a systematic review of the institutional context shaping research careers. This activity will result in synthesizing the theoretical model for research careers.

- Task 2 focuses on the definition of a platform which will be based on the framework to be developed in the preceding Task 1 and will allow the incorporation of different existing datasets. This task includes the preparation of our own datasets for the integration into the platform (Activities 2 and 3) and comprises necessary sub-tasks such as a) the definition of the conditions under which broad access to the data can be granted (these in some cases will have to be negotiated with the agencies funding the respective surveys as well as other participating institutions [universities and non-university institutions]), b) defining rules for ensuring data security and, c) the anonymization of micro-data. This task is related to WP6 and WP15 and WP19. Furthermore, it includes the harmonization of category descriptions and terms used in these data sets allowing for the integration of the data sources to an extent possible, desirable and reasonable (Activity 4). Furthermore, data sources providing information on relevant background data and data sets such as data on institutional information on careers in different national research systems (information and data on framework conditions) will be checked and documentation files providing background information and guidance on comparability, options and limitations for analysis and interpretation will be developed (Activity 5).

- In Task 3 the partners contributing to this work package – IFQ, NIFU and CSIC – will incorporate their own data sources (European and national data) into the platform as well as the information gathered on external data sets as outlined above (Activity 6). Following the implementation of Task 3 on-site visits to the facilities (facilities MORE and EARLY CAREERS) are envisaged allowing the exploitation of the facilities for analytical purposes.

- Task 4 is devoted to the presentation and dissemination of the new framework and ‘platform’ for incorporating various ‘local’ datasets. The dissemination activities should highlight analytical options but also point to desiderata regarding empirical data (Activity 7).
3 – Activities engaged and participants

Following the structure of the project as outlined above the following Activities in the course of the project are currently under preparation:

Task 1

• Activity 1
  Objective: Development of a systematic theoretical framework for researchers' careers
  Duration: June 2014-June 2015
  Partners involved: CSIC (lead), IFQ, NIFU

Task 2

• Activity 2
  Objective: Preparation of MORE data set for opening
  Duration: September 2014-January 2015
  Partners involved: NIFU

• Activity 3
  Objective: Preparation of iFQ data set for opening
  Duration: September 2014- June 2015
  Partners involved: IFQ

The following activities have not yet started but were foreseen in the planning of the work package:

• Activity 4
  Objective: Harmonization of data sets on researchers' careers
  Duration: September 2015 - December 2015
  Partners involved: IFQ (lead), NIFU, CSIC

• Activity 5
  Objective: Identification, analysis of data sources providing background and framework information on careers in different national research systems. Compilation of documentation files and guiding manuals
  Duration: October 2015 - September 2016
  Partners involved: IFQ (lead), NIFU, CSIC

Task 3

• Activity 6
  Objective: Definition of platform and incorporation of national and European data sets and background information
  Duration: June 2016 - June 2017
  Partners involved: IFQ (lead), NIFU, CSIC

Task 4

• Activity 7
  Objective: Presentation of the new framework and ‘platform’ for incorporating various ‘local’ datasets and outlining analytical options. Pursuing own research tasks based on the new platform
  Duration: January 2017 – December 2017
  Partners involved: IFQ (lead), NIFU, CSIC

4 – Outputs

Task 1, Activity 1

Objective: Development of a systematic theoretical framework for researchers’ careers
Duration: June 2014-June 2015
Partners involved: CSIC/INGENIO (lead), IFQ, NIFU

The activity started in July 2014 after the final approval from the FCB. To kick off the activity a meeting was organized and hosted in Valencia. Representatives from all participating partners attended the meeting which was held on July 23rd-24th2014.
The main output of the meeting was a first working document towards developing an integrated framework for analysing research careers which was made available to all partners and serves as a basis for further analysis and discussion. Following the workshop in Valencia the literature review and analysis towards developing a theoretical framework for research careers started and was conducted. The following working steps were carried out:

- Bibliographic searches aiming at covering all ranges of contributions on research careers in the social sciences and organisation of a Mendeley Reference Library on research careers implying a systematic organisation of the literature. The output from this exercise is a comprehensive bibliography covering relevant literature from psychology, management, economics, sociology and STS.

- Analysis of the literature in line with the draft scheme for an integrative framework proposed in the initial working document. A review and analysis of theoretical and empirical approaches to research careers was conducted – with a focus on typologies of careers, career stages and factors shaping career decision making and development processes. In addition, recent research projects and policy initiatives that provide descriptive approaches to research careers were also reviewed. The outcome of this work is an updated version of the working document towards developing an integrated framework for analysing research careers. This working document is submitted as deliverable with this report (as internal output 1 and deliverable D24.1 – report on integrated framework). It constitutes the basis for the scientific paper that the Activity will produce as final outcome. The research paper to be submitted to a scientific journal is currently under preparation and a draft for circulation will be discussed during a project meeting scheduled for September 3rd in Lugano (during STI2015). While we aimed at finishing also the research paper by End of June 2015 due to maternal leave of the lead researcher and the lacking chance to discuss the paper further with the whole team after the RISIS week in January 2015, we now attempt to submit the research paper by the end of September 2015. The draft of the paper will be circulated among the team members by the end of July. Further development and integration of comments from all groups within the team will best be done in the context of a face-to-face meeting. Thus, as mentioned above the team has scheduled a meeting in Lugano on the 3rd of September, coinciding with the 20th International Conference on Science and Technology Indicators. Comments and suggestions will be collected during the meeting with the purpose of finalizing the research paper by the end of September 2015. The final paper and thus the final version of D24.1 will be submitted to a scientific journal and will be submitted to the European Commission.

- Systematic review of institutional frameworks shaping research careers in different countries based on the information collected by the MORE2 project (internal output 2 delivered with this report).

- Concerning internal output 3) “guidelines for the empirical assessment of research careers and the operationalization of the proposed conceptual framework”, we propose elaborating the operationalization guidelines after the team has collectively signed off on deliverable D24.1.

Task 2, Activity 2
Objective: Preparation of MORE data set for opening
Duration: September 2014-January 2015
Partners involved: NIFU: Eric Iversen

The activity started in October 2014. Taking into consideration delays in securing the MORE2 data (see below), the preparations for the opening of the MORE Facility have largely gone according to plan. However, the facility was originally intended to be opened by June 2015. Several factors have slowed the progress (cf. deviations from the initial plan [in the facility report WP 15]).

Here we report according to the following categories in the standard procedures now codified by WP6 (“The process of opening datasets”).

The steps concerning the work on essential access criteria have been conducted both for MORE1 and MORE2. This includes coordination and clearance with data-holders (for MORE2) and with the data-owner (Commission) about access criteria. The data-owner (Commission) asks specifically that MORE1 and

RISIS – First Reporting Period – Strategic Report
MORE2 will not be integrated due to the fundamental differences. Based on WP6 standards, we are ready to implement the access criteria in a form to be signed by accredited users.

In terms of data quality, the cleaning, categorizing, and labelling of MORE1 has been done as has the weight-scheme. MORE2 has been cleaned, categorized and labelled, although some questions about the weights (post-stratification) still need to be resolved.

In terms of the user-handbook and meta-data, drafts of the metadata for both datasets have been generated. The documentation on MORE has been made along the lines of basic structure elaborated for WP6. This has resulted in reports for each dataset: some finishing touches are needed for MORE2 as we resolve remaining questions. The issue that needs to be finalized for the handbook involves finding the best means for weighting for the open dataset. A preliminary merge with ETER has been conducted for both datasets, in part to test whether this can improve accuracy.

The progress can be summarized in terms of the results that were originally scheduled.

i) Documentation of the datasets (SUF)
   a. Result: discussions with WP24 partners at the Valencia meeting on the conceptualization of career-stages
   b. Result: preliminary discussions with subcontractor (IDEA Consult) on appropriate categories and definitions

ii) Developing the model agreements for data utilization
   a. Result: preliminary discussions with partners and with subcontractor on a model agreement for (i) internal RISIS project members and (ii) external users

iii) The Scientific Use of the MORE micro-data
    Result: The Scientific Use File (SUF) has been prototyped and reported on in the poster for the meeting at the STI2014 and at the RISIS week 2015 in Rome. The reports for both MORE1 and MORE2 are now on the RISIS site and can be found in the annex of this report.

   iv) Handbook of the SUF
    Result: The Documentation information has been developed and can be found in the report.

   v) Report describing the data and outlining analytic potential of mobility data
      a. Result: ongoing communication with partners in WP24, Task 1.

We have made additional steps towards opening the facility in the following venues. (cf. WP4)
- WP24 meeting on the role of the MORE2 datasets in modelling researcher careers (WP24, Task 1)
- Presentation of MORE1 for RISIS session at STI conference (STI2014)
- Demonstration of MORE1 data at STI conference, including researchers and policy-makers
- Preparation of input on MORE2 for RISIS session at STI conference (STI2015)

Task 2, Activity 3
Objective: Preparation of IFQ data set for opening
Duration: October 2014- June 2015
Partners involved: IFQ: Jakob Tesch, Janine Lange

The activity started in October 2014, one month behind the original schedule. The duration foreseen is ten month. The activity continues the work which was started in Activity 1 of WP6 (“Report on the content and technical structure of the ProFile infrastructure”). The objective of the activity is the preparation of IFQ’s Early Careers dataset (“ProFile”) as Scientific Use File (SUF) for on-site access (WP19). The main outcome of this task is to enable researchers to use the dataset to run scientific analysis based on their individual research interest and provide them with an
adequate documentation of the data. The activity will be the basis for a training event (WP5) to be hosted in Berlin in September of 2015.

In what follows we report on the progress achieved thus far.

The focus of the activity so far was on merging, cleaning, categorizing, and labelling the data of different panel waves. As a result a master data set was created which contains all the relevant information needed to extract individual scientific use files from the individual panel waves according to the individual research interest.

A concept for user documentation has been developed. The document, which includes all questionnaires, field reports, and codebooks is presently being compiled. In parallel, conditions for on-site access were negotiated with the data protection officer and the IT service in house. A model agreement for data utilization between the individual user and IFQ is being finalized.

Effort also went into the parallel dissemination and information about the facility and the training options provided. In order to inform as many people as possible about the training (WP5) in September research was conducted online to generate a list of suitable contacts. The call for the training is being distributed among this list. Moreover, a poster will be presented at the STI 2015 in Lugano showing the results of research achieved using the ProFile data and thus presenting the analytical potential the facility offers.

The progress can be summarized in terms of the results that were originally scheduled.

i) solve privacy and access regulation issues and create concepts for SUF use
   a. Result: a model agreement for data utilization between the individual user and iFQ is being synchronized with the requirements of the internal data protection officer. It will comprise – among others – paragraphs dealing with the prohibition of data dissemination/copying and deanonymization, criteria for the publication of results and other duties of the individual user.
   b. Criteria for adequate guest researcher workplaces and respective regulations have been developed (e.g. on- site use, personalized access, restricted computer environment, program documentation of syntax) and are being synchronized with the requirements of the internal data protection officer as well as the IT-representative.
   c. Work on the agreement for data utilization as well as the criteria for guest researcher workplaces will be finished in September 2015

ii) Data related work
   a. Result: The panel waves were reorganized and merged into the master data set. A unique identifier across all waves now allows the identification of specific individuals.
   b. Criteria for “analyzeable questionnaires” have been developed and applied. A concept for the description of data accuracy has been developed and the document is currently being assembled.
   c. Cleaning and categorization were started and will finish in September 2015.

iii) documentation
   a. A concept for the documentation of all questionnaires has been developed and document is being finalized.
   b. Ad hoc-restructuring routines have been developed for the first probable use cases which are i) cross sectional analysis of doctoral candidates an ii) longitudinal analysis of doctoral candidacy

iv) dissemination
   a. The structure and substantive content of ProFile was presented as a poster at the STI2014 in Leiden.
   b. Analytic potential of ProFile was presented as a poster and oral presentation at the RISIS week in Rome, January 2015
   c. Poster highlighting research results will be presented in Lugano at the STI in September 2015
   d. Site visit with FCB is scheduled for September 17th in Berlin
   e. Type A Short Course for Training (WP5) has been approved and will take place on September 23rd/September 24th at Humboldt-Universität zu Berlin
   f. Calls for training have been disseminated via various mailing lists and forums
5 – Challenges

The time of writing the report coincides with the wrap up of two of the three activities currently pursued (Activity 1 and Activity 3). The AS for Activity 4 (harmonisation) has been submitted to FCB and is currently under review and waiting for approval. The submission of the AS for Activity 5 (identification of relevant data sources) has been postponed to October 2015 as it relies on the results of Activity 1, to be completed in September 2015. We are facing a slight delay for Activities 4 and 5 due to the problems encountered in the preparation and approval of the preceding activities and the delay caused through these problems. We are however confident that the delay does not compromise reaching the overall objective of the work package and, that none or only minor consequences regarding the planning and the execution of the activities for Task 3, scheduled to start in June 2016, and Task 4, scheduled for January 2017, are to be expected.

6 – Programme for next year and milestones

In the following year work will focus on the integration of the different data sets, harmonization of categories and start the development of the platform for the integration of the data sets (Activity 4-6) (MS9 Technical platform developed due in month 24). Moreover, in WP19 transnational access to the ProFile infrastructure will be organized as well as transnational access to MORE in WP15.

5 – SCIENCE AND INNOVATION POLICY EVALUATIONS REPOSITORY (SIPER, WP6, WP21)

Though SIPER is a new dataset, as there had been an experimental predecessor, it was thought that SIPER developments could both follow the joint opening process while having its own developments. The following table presents the work programme and what has been achieved end of June 2015.

<table>
<thead>
<tr>
<th>Workpackage:</th>
<th>WP6</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Objectives:</strong></td>
<td>Organise the preparation (technical, substantive and legal) for the opening of the 10 datasets.</td>
</tr>
<tr>
<td><strong>Work programme:</strong></td>
<td></td>
</tr>
<tr>
<td>Re Task 1:</td>
<td>Preparation of report on (intended) content and technical structure of IPER dataset</td>
</tr>
<tr>
<td>Re Task 2:</td>
<td>Assessment of I categorisations for potential harmonisation.</td>
</tr>
<tr>
<td>Re Task 4:</td>
<td>Work on principles and conditions of access: Investigations of legal conditions on use of public sourced evaluation reports.</td>
</tr>
<tr>
<td>Re Task 5:</td>
<td>Rebuilding of access to dataset contents (target November 2015). This involved:</td>
</tr>
<tr>
<td></td>
<td>o Close liaison with University IT Services to build host database and web site, including provisions for external (user) access and internal project team administration.</td>
</tr>
<tr>
<td></td>
<td>o Pilot assessment of database functionalities and adjustment as required.</td>
</tr>
<tr>
<td></td>
<td>o Implementation of design of Web page and integration with database functionalities (ongoing discussions with MBS Web Services and University IT Services).</td>
</tr>
<tr>
<td></td>
<td>o Reconfiguration of database information collection (data input) template.</td>
</tr>
<tr>
<td></td>
<td>o Assessment of currently available evaluation reports (pdf files) from existing databases (INNO-Appraisal and NESTA) and migration to new database location.</td>
</tr>
</tbody>
</table>
| | o Migration of existing template information to new location. (Not possible due to
Template changes)

- Re-analysis of existing reports to comply with new data input template. (Piloted – awaiting development of SIPER FC functionality – October 2015)

**Re Task 6:**
- Preparation of report, containing a full documentation of the dataset (taking into account the changes introduced through the collective work) and explaining conditions of ‘relevant’ use.

**Activities engaged (and participants):**

**Re Task 1:**
- Report prepared (PC)

**Re Task 2:**
- Assessment of I categorisations for potential harmonisation; no clear possibilities identified at this stage.

**Re Task 4:**
- Work on principles and conditions of access: Preparation of standardised text to be used outlining conditions of use of public sourced evaluation reports, including options for requests for withdrawal of material.

**Re Task 5:**
- Administrative (internal) Portal development completed – allows upload of evaluation reports at present: (ongoing – internal SIPER Basic version complete and tested).
- Reconfiguration of database information collection (data input) template – completed April 2015.
- Formulation of new Support Instrument Typology (Completed March 2015)
- Assessment of evaluation reports (pdf files) from existing databases (INNO-Appraisal and NESTA) and migration to new database location. (ongoing - c.300 reports migrated).
- Design and construction of Factual Characterisation element of database (completed August 2015).
- Design and construction of Judgemental Characterisation element of database (ongoing – estimated completion date September 2015).
- Design and construction of Policy maker interface element (ongoing – target date October/November 2015)
- Data base opening – January 2015 target.

**Re Task 6:**
- No report prepared – completed as Technical Specifications document, agreed with UoM IT Services, June 2014.

**Outputs:**
- Report on (intended) content and technical structure of IPER dataset
- Evaluation upload facility (internal portal) – tested
- Revised template for data collection/report characterisation
- Support Instrument Typology – integrated into data collection template
- Initiated trial uploading of evaluation reports

**Challenges:**
- IT development task was larger than foreseen and mobilisation of IT services took longer than anticipated (due to internal University restructuring) – now developed timetable for phased delivery of database and functionalities.
- Uncertainty over responsibilities for Web design due to internal restructuring issues.

**Programme for next year (with milestones anticipated):**
- Continued development of portal functionality to allow on-line input of evaluation report characterisation (to proceed via internal team access, external team access and external policy maker access and input)
- Re-categorisation of existing evaluation reports to comply with new data input template.
Annex to WP7 - Demonstrating advanced data integration with the SMS platform: Grant selection bias as case

Peter van den Besselaar, Ali Khalili Loizou, Al Iddrissou, Frank van Harmelen
Network Institute - VU University Amsterdam

Introduction

Up to now, science and innovation studies have relied either on small scale rich data (e.g. case study data), or on large scale relatively ‘poor’ data (Publication or Patent data, surveys) – in the latter case only covering a small part of the relevant dimensions of the questions to be answered and the problems to be solved. Knowledge progress asks for both: rich and heterogeneous data but on a large scale. Data infrastructures are needed to provide this.

Consequently, the core of relevant data infrastructures is to deliver data integration – more than just data. Data are increasingly everywhere, so the social scientists’ role becomes less the collection of data than the production of research data out of the data available – to a large extent the space of linked open data, but also proprietary data that can be made available for research. Advanced technologies can be deployed to make this possible. Advanced tools are needed to do the job – tools that together construct a workflow supporting the researcher. As these tools remain rather complex, an infrastructure requires next to the tools additional support by e.g., programmers and data scientists.

SMS platform

If RISIS wants to play that role in science and innovation studies, we need to show how the project will contribute to this long-term perspective. More than the datasets, the platforms may have to do that job. The data sets are of course crucial for research, but increasingly the linked open data space will provide relevant data. And the RISIS dataset will be some among many others available.

This demonstrator is an example of how it may work. The demonstrator shows how the SMS-RISIS platform enables and supports answering complex research questions with high policy relevance. Figure 1 shows the basic architecture of SMS.

SMS consists of various datasets, which may have different access levels (regulated through an access control service). The data will be preprocessed and then converted into RDF format (see below for the preprocessing). The metadata are stored in a search tool, enabling potential users to identify what they need. Data integration runs through named entity recognition (what kind of entities are in the data?) and disambiguation (the identity resolution service). Data enrichment runs through e.g., geo-location services, and through combining the data with other databases such as DBpedia that store additional data about the entities in the various databases. Through a range of queries, the researcher can retrieve out of the integrated data a set of research data needed for answering the research question. That data can go into analysis and visualization tools.
The case
The research question for the demonstrator is whether the different success-rates between male and female applicants are an effect of gender bias, or of different performance levels of male and female researchers. And in case it is bias, where is this bias occurring?
This question is theoretically challenging, as literature is clearly undecided on whether gender bias exists, and if so, what the causes are. It is also politically relevant, as it is a core topic in H2020, e.g., in the GERI program. The innovative capacity of science and society depends on the mobilization of resources – and gender bias should not block this. And the issue of gender bias in science is permanently in the public and political debate.
The demonstrator study aims at explaining success in terms of a series of independent variables such as performance variables (publications, citations, previous grants), the quality of the network (based on ranking of institutions the applicant is affiliated with), personal characteristics (age, gender), and the panel characteristics (gender stereotyping, etc). The model in in figure 2.
The first problem is that whether gender bias exists can only be solved by taking past performance into consideration – and most studies do not do this: only by controlling for past performance, one can decide if gender differences in success rates are gender bias.
A second problem is what counts as past performance. This is highly debated. In this study we use several publication and citation based indicators, grants acquired (grants are a resource and – because of the symbolic value an output); and the realized collaboration network. Among the latter we also count the host organization.
Thirdly, several personal characteristics may impact the decision-making process, such as gender, age, nationality, university of the PhD.
Fourthly, several attributes of the panels may play a role, as there the decisions are made. One may think of the (partly discipline based) level of gender stereotyping, the research topics covered by the panel members, the country, language, age and gender distribution of panelists, the workload of the panel, the gender distribution of the applicants.
The SMS platform demonstrator

In order to answer the question of the existence and causes of gender bias, we need a whole range of data – in order to cover the many variables involved - and a big case. Both we have: 3030 applicants, and their applications, the CVs and the review reports. About 95% of the applicants did participate, so we may assume that we do not have self-selection bias. We do have a variety of data sets for the project. The facet browser of the infrastructure shows the data available for the demonstrator (Fig 3).

In fact the system already includes much more data sources. But we show here only those that will be used for the demonstrator. Several of the datasets are secret (the administrative data about the applications, the applications, the CVs and the evaluation forms). Some other are proprietary, but relatively easy accessible if one has the appropriate licenses (e.g., Web of Science). Other data are open, such as the Leiden Ranking, ETER, and also non-RISIS data such as the CORDIS data.

In figure 4, the datasets used for the demonstrator are represented ‘in’ the infrastructure. Also the tools that are deployed in preprocessing and integration are shown. As is clear, it is not yet the full-scale version of the SMS platform. Of course, preprocessing and coupling require quality control – but we skip the discussion about how to do this as this is still ongoing work. This is also the case for the disambiguation – a very crucial task under development.
Data for the demonstrator

Figure 3. The SMS facet browser

Figure 4. The gender study data in the platform

RISIS – First Reporting Period – Strategic Report
The workflow

**Phase 1: preprocessing**
The raw data need preprocessing in order to be integrated and used in the analysis. For example, in this demonstrator we needed to extract from the CVs all organization names in order to be able to calculate the median rank order of the organizations mentioned in a CV as a proxy for the quality of the network of the applicant. The same holds for the number of grants and the amounts of those grants, in order to calculate a measure for past success in grant acquisition. Also the review reports are preprocessed for term extraction (using Cortext) and these terms are then again preprocessed to construct our linguistic categories and the scores of each applicant on those categories.

Figure 5 gives some preprocessing examples:
(i) Review reports -> linguistic variables measuring gender stereotyping (using Cortext and LIWC)
(ii) CVs -> extraction of received grants
(iii) CVs -> extraction of organizations for determining the applicants’ network quality (extracting entities and entity recognition)
(iv) WoS records are passed and converted into the RDF format.
(v) Disambiguation then leads to indicators of output and impact
(vi) ETER and the Leiden Ranking are converted in RDF

![Preprocessing Diagram](image)

**Fig 5. Preprocessing**

**Phase 2: Integration**
All data are converted into RDF. For all datasets entity types (people, organizations, projects) are identified and are connected. The entity scheme for this demonstrator is shown in figure 6. Disambiguation connects the same entities in the different databases: e.g., persons in the project database and in the publication database; organizations in the CV database and in the Leiden Ranking.

Attributes are related too, using category services. For example, research fields may be included in the different datasets in different classifications.

RISIS – First Reporting Period – Strategic Report
Phase 3: Querying the linked data to retrieve a required dataset

The integrated data platform is rather large, as it of course has much more data than only those needed for answering a specific research question – as the platform has a general function for supporting research.

---

Figure 6. Data schema

So one needs to query the data platform to get those data one needs for a specific research project. In order to support the users that have not the skills of writing their own queries, and to prevent too many too complex queries making the system break down, a series of standard queries is provided. These are based on the kind of questions researchers in the field have. However, support is given for specific needs – that is the peoples part of the infrastructure.

An example of a query is in figure 7: asking for the reviewers’ scores for the applicants, which can be answered through the integration of the applicants’ database with the publication database. Another query would be “give for all applicants the reviewers’ scores, the number of publications an citations, and the amount of grants reported in the CVs”.

Phase 4: analysis.

After retrieving the data, one may import the data in analysis and visualization tools. One may also retrieve data for constructing variables that later on go into an analysis.

An example of the latter: construction of variables. This is a preliminary analysis. For the demonstration study we need a variable at the panel level: the degree of gender stereotyping. We need this in order to find out whether (i) gender stereotyping exists in panels, and (ii) whether gender stereotyping influences the success of female applicants.
How is this done? Preprocessing the review reports leads to word frequencies for each review for each applicant. There words are then aggregated into linguistic categories, such as negation words, negative evaluation, positive evaluation, agency words, etc. Analyzing correlations between the categories and between categories and application success leads to the selection of negating words as best indicator. The data are aggregated to the panel level and that gives two variables: The relative frequency of negating words in reviews of female researchers (compared to male applicants) at the panel level and (ii) the relative success rate male applicants compared to female, again at the panel level. The relation between the two variables is shown in figure 8.

What we see is a positive relation: the more negating words are used in a panel for female applicants compared to male applicants, the stronger the relative success rate of male applicants. That is expected. However, the effect is not symmetrical. The top-right quadrant is much more populated and has more extreme values than the bottom-left quadrant:

1- When men get 10% more negating words, men score about 2.5% (average) lower than female applicants. However, if female applicants get 10% more negating words than male applicants, females score 7.5% lower in terms of success.

2- Women receive much more negating words than male applicants: There are many panels scoring above 1.10, but hardly panels scoring below 0.9.

3- Finally, in all but two life science panels, men get less negating words than women. And in all but one life science panels, men have equal or better success rate than women.
This is the first step in the variable construction. The next step will question: does this pattern remain if we control for differences in scholarly performance (publications, citations, grants)? In other words, are differences in success rate only gender differences or is it gender bias? That is for the next step.

**The results expected from the demonstrator**
The demonstrator will illustrate (i) that combining data leads to more exiting and relevant studies than single datasets, (ii) that advanced platforms for data integration are very helpful in making this possible, and (iii) that always manual work and different supporting expertise (e.g. data science) is needed to do this – but at a higher level than before.