RISIS
Research infrastructure for research and innovation policy studies

SEVENTH FRAMEWORK PROGRAMME

AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY

RISIS – WP8 Demography

Research infrastructures for the assessment of science, technology and innovation policy

Barbara Heller-Schuh, Michael Barber, Marlies Züger, Thomas Scherngell

European Commission, DG RTD, 01/14-01/17
Grant Agreement no: 313082
RISIS Annual Week, 26-29 January 2015, Rome
Group session B

Agenda

I. Demographic events in the EUPRO Database

II. Tracking demographic changes in EUPRO: a 4-stage approach

III. Expected outcome

Background

- The EUPRO database currently covers a period of more than 30 years during which organisation names have changed due to mergers, splits, take-overs and spin-offs.

- EUPRO is intended to be used as information source to track demographic events of public research organisations.

- EUPRO is expected to deliver a good coverage of demographic events of public research organisations; however full coverage cannot be guaranteed.

Demographic events in the EUPRO Database

- At the moment name changes of organisations are not tracked in the EUPRO database. Organisations are labeled by their most recent valid name.
- This facilitates analysing dynamic participation patterns of individual organisations in the EU FPs.
- On the other hand, important data about organisations are only accessible through the underlying standardisation table and directly searching for the applicants in former EU FPs is not possible.
- Thus, the field *stApplicant* will be implemented in the EUPRO database, where the standardised original organisation name is stored.

Examples

- Simple name changes
  - In 2012, the **Facultés universitaires Notre-Dame de la Paix** changed its name to **Université de Namur** (in 2012)
  - The University of Hasselt was established in 1971 as **Limburgs Universitair Centrum**; in 2005 it changed its name to **Universiteit Hasselt**.
- Mergers/Take-overs
  - **Karlsruhe Institute of Technology (KIT)** was created in 2009 when the **University of Karlsruhe** merged with the **Karlsruhe Research Centre**.
  - In 2010, TKK was merged with **Helsinki School of Economics** and **University of Art and Design Helsinki** into **Aalto University**.
- Splits/Spin-offs
  - **Fachhochschule Oldenburg/Ostfriesland/Wilhelmshaven** was split in 2009 into **Jade Hochschule** and **Hochschule Emden/Leer**
  - **Link Campus University** used to be the Italian branch of the **University of Malta** and was recognized as university by the Italian Government in 2011

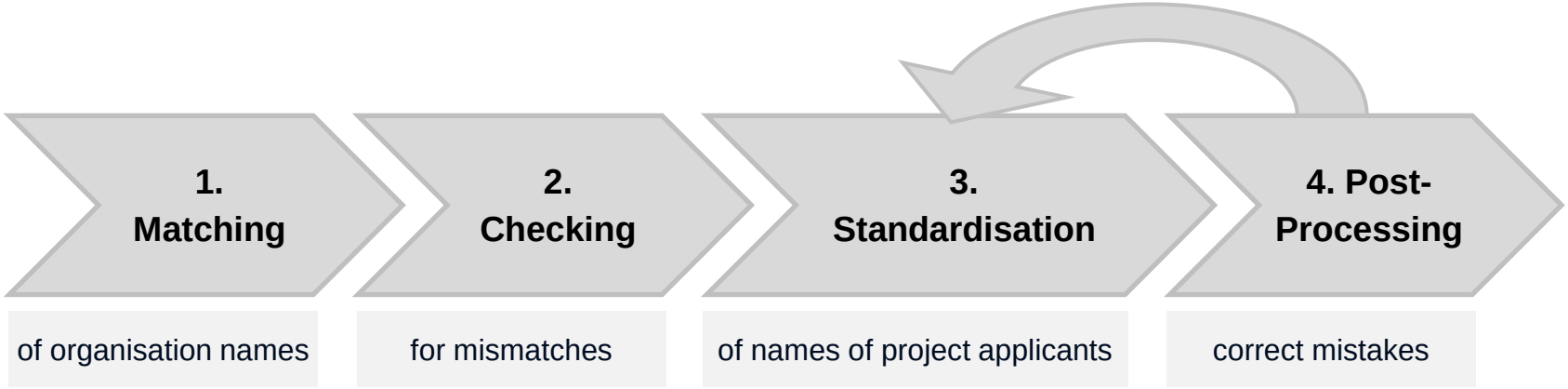
Tracking demographic changes: a 4-stage approach

I. Matching

II. Checking

III. Standardisation

IV. Post-Processing



Step 1: Matching

- The matching-algorithm is based on a weighted 2-gram model and the TF-IDF measure (Term Frequency – Inverse Document Frequency).
  - Parse: Université de Namur → universite de namur
  - 2-grams: universite de namur → un, ni, iv, ve, er, ... mu, ur
  - TF-IDF: treat 2-grams as the terms and the set of names as the corpus of documents. Similarity increases when names have more 2-grams in common, but 2-grams frequent in the set of names contribute less than rare ones.
- Input is a list of all organisation names including the original as well as the standardised version (stored in *stOrg*).
- The comparison of names is made individually for each organisation (those with identical standardised organisation names).
- The output is a score between 0 and 1, where 0 means no match and 1 means perfect match in the weighted 2-gram model (but not necessarily between the two organisation names).

Example for a sorted matching output

original organisation name	standardised org. name	score
HASSELT UNIVERSITY	Universiteit Hasselt	0.657651946169
TRANSPORT RESEARCH INSTITUTE - HASSELT UNIVERSITY	Universiteit Hasselt	0.424294586939
HASSELT UNIVERSITY - TRANSPORTATION RESEARCH INSTITUTE	Universiteit Hasselt	0.286880936889
UNIVERSOTÉ DE LIMBURG	Universiteit Hasselt	0.056492594007
Limburg University	Universiteit Hasselt	0.041552208712
LIMBURGS UNIVERSITAIR CENTRUM / STUDIECENTRUM VOOR MULTIMEDIAAL EN INTERACTIEF LEREN	Universiteit Hasselt	0.040106436661
University of Limburg Faculty of Economics and Business Administration	Universiteit Hasselt	0.035748760115
LIMBURGS UNIVERSITAIR CENTRUM	Universiteit Hasselt	0.028501765416
LIMBURG UNIVERSITAIR CENTRUM	Universiteit Hasselt	0.027784174203
Dr. L. Willems-Instituut vzw	Universiteit Hasselt	0.027415411742
Limburgs Universitaire Centrum	Universiteit Hasselt	0.025250899474
Dr L.Willems-Instituut v.z.w.	Universiteit Hasselt	0.024471556638
LIMBURGS UNIVESITAIR CENTRUM	Universiteit Hasselt	0.017365295690
LIMBURGS UNIVERSITAIR CENTRUM (LUC)	Universiteit Hasselt	0.017183103724
LIMBURG UNIVERSITY CENTRE (EDM-SMILE)	Universiteit Hasselt	0.013056135097
UNIV CENTRUM LIMBURG	Universiteit Hasselt	0.005427124495
LIMBURGS CENTRUM ONDERWIJS	Universiteit Hasselt	0.001613835185

Step 2: Checking

- Recommendations in the matching procedure have to be checked manually, as it usually produces a small number of perfect matches (score 1) but a wide range of close matches (score 0.9-0.7).

- In this case matches with a low score are of special interest as they indicate name changes of identical standardised organisation names.

- Each of these low-score recommendations needs to be checked manually and tagged as either a spelling variant of the recent valid organisation name (**variant**) or as a name change of the organisation (**change**).

- The tagging should proceed in a simple and quick process without additional investigations in the web.

Example for checking procedure

spelling variant of "Universiteit Hasselt"

original organisation name	standardised org. name	score	variant	change
HASSELT UNIVERSITY	Universiteit Hasselt	0.657651946169	✓	
TRANSPORT RESEARCH INSTITUTE - HASSELT UNIVERSITY	Universiteit Hasselt	0.424294586939	✓	
HASSELT UNIVERSITY - TRANSPORTATION RESEARCH INSTITUTE	Universiteit Hasselt	0.286880936889	✓	
UNIVERSOTÉ DE LIMBURG	Universiteit Hasselt	0.056492594007		✓
Limburg University	Universiteit Hasselt	0.041552208712		✓
LIMBURGS UNIVERSITAIR CENTRUM / STUDIECENTRUM VOOR MULTIMEDIAAL EN …	Universiteit Hasselt	0.040106436661		✓
University of Limburg Faculty of Economics and Business Administration	Universiteit Hasselt	0.035748760115		✓
LIMBURGS UNIVERSITAIR CENTRUM	Universiteit Hasselt	0.028501765416		✓
LIMBURG UNIVERSITAIR CENTRUM	Universiteit Hasselt	0.027784174203		✓
Dr. L. Willems-Instituut vzw	Universiteit Hasselt	0.027415411742		✓
Limburgs Universitaire Centrum	Universiteit Hasselt	0.025250899474		✓
Dr L.Willems-Instituut v.z.w.	Universiteit Hasselt	0.024471556638		✓
LIMBURGS UNIVESITAIR CENTRUM	Universiteit Hasselt	0.017365295690		✓
LIMBURGS UNIVERSITAIR CENTRUM (LUC)	Universiteit Hasselt	0.017183103724		✓
LIMBURG UNIVERSITY CENTRE (EDM-SMILE)	Universiteit Hasselt	0.013056135097		✓
UNIV CENTRUM LIMBURG	Universiteit Hasselt	0.005427124495		✓
LIMBURGS CENTRUM ONDERWIJS	Universiteit Hasselt	0.001613835185		✓

name change from "Limburgs Universitair Centrum" to "Universiteit Hasselt"

Step 3: Standardisation

- In case that the original organisation name is a spelling **variant** of the standardised organisation name,
  - the standardised organisation name in *stOrg* is copied automatically to the *stApplicant* field (“Universiteit Hasselt”).

- All entries tagged as **changes** need to be examined more closely
  - a standardised version of the original organisation name has to be stored in the *stApplicant* field (“Limburgs Universitair Centrum”)
  - name changes and demographic events will be documented in a separate table in the data base (standardised original name, recent valid standardised organisation name, type of change, effective date)

- Standardised organisation names in the fields *stOrg* and *stApplicant* are assigned to the respective IDs in the Register of Public-Sector Research and Higher Education Organizations (OrgReg).

Step 4: Post-Processing

- The final step is to determine and correct mistakes, which may occur in the following cases:
  - a demographic event was missed:
    a relevant mismatch was not identified in step 2 (Checking)
  - the underlying standardisation is (still) incorrect:
    the wrong standardised name was assigned to an organisation in former standardisation processes
  - the underlying standardisation is (still) incomplete:
    two or more names still exist for the same organisation

- It is not to be expected that all mistakes are detected and eliminated, since there is no way to particularly search for them. However, while proceeding Step 3 (Standardisation) one should be aware of those kind of mistakes.

Expected outcome

- Documentation of demographic events
- Standardisation of organisation names over time
- Input to the Register of Public-Sector Research and Higher Education Organizations (OrgReg)

- Implementation plan
  - AIT will start with a first pilot on HEIs (perimeter will be identical to that of ETER) in February 2015
  - 4-stage approach will be applied to PROs as soon as a list of PROs (according to the definition in RISIS) is given