ERACEP: IDENTIFYING EMERGING RESEARCH AREAS IN THE ERC RESEARCH PROPOSALS

Final Workshop DBF & ERACEP CSAs
20/21 February 2013, Brussels

Dr. Thomas Reiss
Etienne Vignola-Gagné
Piret Kukk
Fraunhofer ISI Karlsruhe

Prof. Dr. Wolfgang Glänzel
Dr. Bart Thijs
KU Leuven
Objectives:

- Identification of topically emerging research areas and analysis to what extent the activities supported by the ERC cover and contribute to these research areas

Concept:

- Combination of two perspectives:
  - From the ‘landscape of science’
  - From ERC funding procedures
Identification of emerging fields

Expert validation

In-depth analysis of 10 emerging clusters

Pilot with 6 fields composed of 46 clusters

Selection of 20 most dynamic fields for fine mapping

40 dynamic fields

Landscape of Science

Matching of grant applications

Identification of emerging fields

Expert validation

In-depth analysis of 10 emerging clusters

Pilot with 6 fields composed of 46 clusters

Selection of 20 most dynamic fields for fine mapping

40 dynamic fields

Landscape of Science

Matching of grant applications
Aims:

- Explore how ERC grant applications can be mapped to dynamic fields and clusters identified via bibliometric analyses and expert validation.
- Analyse to which extent emerging topics are addressed by ERC grant applications.
- Find which proposals from the 2009 Starting Grants Call (932 proposals, 165 successful) can be classified within the emerging topics previously identified.

These emerging topics should be present in the proposals made to ERC calls.
Matching sample

2009 Starting Grants Call

• 932 proposals, 165 successful
• 888 could be used for analysis
Matching approach

Two approaches:

1. Identify journals mentioned in grant applications and map to dynamic fields defined by WoS subject categories
2. Text matching between proposals and thematic clusters via text mining

Possible value for users

- Ex ante:
  - Classify applications into structure inherent to landscape of science
  - Preselect proposals: emerging – established
  - Provide bibliometric mirror for qualitative expert assessment
- Ex post:
  - Reflection on selection procedure
Matching 1: WoS Subject Categories assignment to proposals procedure

Fictive example

Proposal 99999

Text Mining

Journal of Optics

1. Optics

2. Physics, applied

3. Physics, Atomic, Molecular & Chemical

4. Chemistry, Physical

5. Chemistry, Inorganic & Nuclear

Angewandte Chemie Int.
Matching 1 – opportunities and limitations

Opportunities:
- Allows automated construction of a sample of proposals belonging to selected WoS Subject Categories (and eventually, emerging topics)

Limitations:
- False positives means additional qualitative work is required to finalise the sample (mitigated by low rate of false negatives);
- Text mining also captures keywords contained in journal titles but that may not refer to articles in the proposal;
- Previous publications in the proposals do not always indicate the content of the project presented therein;
- Publications cited in principle might tend towards mainstream

➤ **Focus on second matching approach**
The text matching is built in a vector space model. This ‘Vector Space Model’ (VSM) is most appropriate for representing the content in documents (Salton, 1975).

Each document is a vector with its own position in a space. Each vector has as many dimensions as there are terms. If a term occurs in a document, its value is not equal to zero.

These values are defined as

\[ \text{TF-iDF} = \text{term frequency} \times \text{inverse document frequency} \]

This allows down-weighting of common words as they will appear in many documents.

Similarities between proposals and papers are calculated as the cosine of the angles between the vectors representing each document in the term-based vector space and expressed by TF-iDF.
The mapping of the proposals to topics or detected clusters is based on the average similarity that the applications have to all the linked papers in one cluster. A paper is linked to a proposal if they share at least one term. So the average is taken over all non-zero similarities.

The threshold to define a mapping between an application and a cluster is set at least 0.025.

298 of all processed applications (888) are matched with at least one cluster. This is one third. 173 applications have multiple assignments.

Example: Application 241161 entitled: ‘Intergenerational correlations of schooling, income and health: an investigation of the underlying mechanisms is mapped with 4 clusters:

- Quality of Life
- Health Policy
- Tobacco
- Gender and Family
Matching 2 – results

- 1/3 of all proposals matched at least to one cluster within 6 selected dynamic fields (out of 20)
Ex ante:
- Emerging topics (*) addressed well by proposals
- More conventional topics (e.g. “oil”, “battery”, “coal”) less frequent in proposals

Ex post:
- Highest success rate for “renewables”
- However, also proposals in emerging topics* highly successful
Matching 2 results – Biomedical Engineering

<table>
<thead>
<tr>
<th>Matched Applications</th>
<th>Cluster Label</th>
<th>Absolute</th>
<th>Relative (%)</th>
<th>Share of Success</th>
</tr>
</thead>
<tbody>
<tr>
<td>Catheter Ablation</td>
<td>3</td>
<td>1,5</td>
<td>0%</td>
<td></td>
</tr>
<tr>
<td>Brain Computer *</td>
<td>4</td>
<td>2,0</td>
<td>0%</td>
<td></td>
</tr>
<tr>
<td>Nanostructured</td>
<td>45</td>
<td>22,3</td>
<td>18%</td>
<td></td>
</tr>
<tr>
<td>Cartilage</td>
<td>18</td>
<td>8,9</td>
<td>17%</td>
<td></td>
</tr>
<tr>
<td>Tomography</td>
<td>32</td>
<td>15,8</td>
<td>28%</td>
<td></td>
</tr>
<tr>
<td>Scaffolds</td>
<td>60</td>
<td>29,7</td>
<td>15%</td>
<td></td>
</tr>
<tr>
<td>Bone Cement</td>
<td>22</td>
<td>10,9</td>
<td>9%</td>
<td></td>
</tr>
<tr>
<td>Cardiac</td>
<td>13</td>
<td>6,4</td>
<td>15%</td>
<td></td>
</tr>
<tr>
<td>Kinematics *</td>
<td>5</td>
<td>2,5</td>
<td>0%</td>
<td></td>
</tr>
</tbody>
</table>

Ex ante:
- Strong focus on “scaffolds” and “nanostructuring”
- Emerging topics (*) addressed rarely

Ex post:
- By far highest success rate for imaging approaches
- Proposal in emerging topics* not successful, small number effect?
Matching 2 results – Environmental Sciences

<table>
<thead>
<tr>
<th>Cluster Label</th>
<th>Matched Applications</th>
<th>Absolute</th>
<th>Relative (%)</th>
<th>Share of Success</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diversity</td>
<td></td>
<td>17</td>
<td>15,6</td>
<td>35%</td>
</tr>
<tr>
<td><strong>Radiation</strong> *</td>
<td></td>
<td>5</td>
<td>4,6</td>
<td>80%</td>
</tr>
<tr>
<td><strong>Nanomaterials</strong> *</td>
<td></td>
<td>4</td>
<td>3,7</td>
<td>0%</td>
</tr>
<tr>
<td>Waste Water</td>
<td></td>
<td>20</td>
<td>18,3</td>
<td>10%</td>
</tr>
<tr>
<td>Bentic Zone</td>
<td></td>
<td>23</td>
<td>21,1</td>
<td>17%</td>
</tr>
<tr>
<td>Sediments</td>
<td></td>
<td>16</td>
<td>14,7</td>
<td>13%</td>
</tr>
<tr>
<td>Air Pollution</td>
<td></td>
<td>10</td>
<td>9,2</td>
<td>10%</td>
</tr>
<tr>
<td>Soil</td>
<td></td>
<td>14</td>
<td>12,8</td>
<td>21%</td>
</tr>
</tbody>
</table>

Ex ante:
- Strong focus on waste water and bentic zone issues
- Emerging topics (*) addressed rarely

Ex post:
- Proposals for emerging topic “radiation” highly successful
### Matching 2 results – Geography

<table>
<thead>
<tr>
<th>Cluster Label</th>
<th>Absolute</th>
<th>Relative (%)</th>
<th>Share of Success</th>
</tr>
</thead>
<tbody>
<tr>
<td>Living</td>
<td>13</td>
<td>10,0</td>
<td>31%</td>
</tr>
<tr>
<td>Natural Resources</td>
<td>23</td>
<td>17,7</td>
<td>35%</td>
</tr>
<tr>
<td>Community</td>
<td>26</td>
<td>20,0</td>
<td>31%</td>
</tr>
<tr>
<td><strong>State-Region</strong></td>
<td><strong>43</strong></td>
<td><strong>33,1</strong></td>
<td><strong>28%</strong></td>
</tr>
<tr>
<td>Diversity</td>
<td>25</td>
<td>19,2</td>
<td>20%</td>
</tr>
</tbody>
</table>

**Ex ante:**
- Strong focus on emerging topic “State-Region”

**Ex post:**
- High success rate in all topics
- Proposals in emerging topic “State-Region” highly successful
Matching 2 results – Obstetrics

<table>
<thead>
<tr>
<th>Cluster Label</th>
<th>Absolute</th>
<th>Relative (%)</th>
<th>Share of Success</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prenatal Diagnosis *</td>
<td>1</td>
<td>2,4</td>
<td>0%</td>
</tr>
<tr>
<td>Cesarean</td>
<td>4</td>
<td>9,5</td>
<td>25%</td>
</tr>
<tr>
<td>Pregnancy Complications</td>
<td>1</td>
<td>2,4</td>
<td>0%</td>
</tr>
<tr>
<td>Menopause</td>
<td>3</td>
<td>7,1</td>
<td>33%</td>
</tr>
<tr>
<td>Surgery: Incontinence</td>
<td>0</td>
<td>0,0</td>
<td></td>
</tr>
<tr>
<td>Infertility</td>
<td>5</td>
<td>11,9</td>
<td>0%</td>
</tr>
<tr>
<td>Cancer</td>
<td>28</td>
<td>66,7</td>
<td>18%</td>
</tr>
</tbody>
</table>

Ex ante:
- Applications focusing largely on cancer issues
- Emerging topic (*) not covered by applications

Ex post:
- Numbers too low
Matching 2 results – Public Health

<table>
<thead>
<tr>
<th>Cluster Label</th>
<th>Absolute</th>
<th>Relative (%)</th>
<th>Share of Success</th>
</tr>
</thead>
<tbody>
<tr>
<td>Living Environment *</td>
<td>1</td>
<td>2,0</td>
<td>0%</td>
</tr>
<tr>
<td>Aids, Africa *</td>
<td>6</td>
<td>11,8</td>
<td>17%</td>
</tr>
<tr>
<td>Quality of Life</td>
<td>8</td>
<td>15,7</td>
<td>13%</td>
</tr>
<tr>
<td>Health Policy</td>
<td>17</td>
<td>33,3</td>
<td>12%</td>
</tr>
<tr>
<td>Tobacco</td>
<td>7</td>
<td>13,7</td>
<td>14%</td>
</tr>
<tr>
<td>Gender and Family</td>
<td>12</td>
<td>23,5</td>
<td>17%</td>
</tr>
</tbody>
</table>

Ex ante:
- Emerging topics (*) not addressed very well by applications

Ex post:
- Rather high success rate for emerging topic “AIDS, Africa”, however only low absolute numbers
Advantage of the applied methodology:

- Text-based links do not require parsing of documents to journal references or individual papers.
- Links can be quantified.
- Links can be calculated as soon as the applications are submitted.

Possible problems:

- At this moment the proposal contains all the text that has been submitted: Summary, CV of PI, publication list, detailed description.
- Some common terms (e.g. cell, migration) can distort classification.
- In public health we found several proposal that were dealing with a disease from a purely medical perspective while the topic deals with the societal impact of the disease. Of course, they share a lot of common terms and are even in the same context.

Possible improvements:

- Separated sections or even files for the proposals.
- Increase the weight of phrases in the calculation of similarities.
- Manual validation cannot be eliminated.
Conclusions from the user perspective

Ex ante

1. ERC grant applications can be classified according to a dynamically evolving inherent structure of the landscape of science.
2. Evaluation panels can be organised along such structures and adjusted dynamically.
3. Thereby the structuring of grant applications and setting up of evaluation panels are driven mainly by inherent dynamics of the science landscape and not by static external classifications. This may allow better adjustments of procedures to the evolution of science.
4. ERC grant applications can be mapped to dynamic fields within the landscape of sciences.
5. ERC grant applications can be matched with thematic clusters (emerging or established) within dynamic fields. Thereby proposals can be identified that map to emerging topics.
6. This could be used for supporting a pre-selection of applications and the assignment of proposals to evaluation panels.
Ex post

1. Self-evaluation of selections procedures, not in a sense of simple “good” or “bad” judgments. Rather,
   - facilitates the identification of differences in procedures between scientific fields based on coverage of emerging topics and success rates for all topics;
   - points to topics, where reasons for selection/non-selection may need additional consideration.

2. Provides hints for improving procedures, e.g. organization of structure and content of proposals in a way that facilitates routine bibliometric analysis (separate files for scientific/technical descriptions and information related to PI).
# Self evaluation - example

<table>
<thead>
<tr>
<th>Coverage of emerging topic</th>
<th>Success rate in emerging topics</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>high</td>
<td>high</td>
<td>energy and fuels, geography</td>
</tr>
<tr>
<td>high</td>
<td>low</td>
<td>?</td>
</tr>
<tr>
<td>low</td>
<td>high</td>
<td>environmental sciences</td>
</tr>
<tr>
<td>low</td>
<td>low</td>
<td>biomedical engineering</td>
</tr>
</tbody>
</table>

**Observations:**
- Differences between scientific fields concerning coverage and success rates, reasons?
- If emerging topics are addressed well, success rates of those applications are high (reverse not correct)