A Quality Framework for On-line Job Vacancy (OJV) Data

Nigel Swier
To be or not to be, that is the question.

**English → Lithuanian.**
Būti ar negali būti, kad yra klausimas.

**Lithuanian → Chinese.**
是或不, 這是個問題。

**Chinese → Czech.**
Nebo ne, to je problém.

**Czech → Greek.**
Η όχι, είναι ένα πρόβλημα.

**Greek → English.**
Or rather, it is a problem.
Selected Quality Issues with OJV data

- Not all job vacancies are advertised on-line.
- Some types of jobs are more likely to be advertised on-line than others.
- The penetration of on-line job markets varies by country and changes over time.
- In most countries, the on-line job market is comprised of multiple actors with different business models and there is usually no definitive source of on-line job ads.
- A single vacancy may be advertised multiple times on different platforms (or even the same platform)
- A single advertisement may contain more than one vacancy.
- Variables such as occupation, industry and geography classifications need to be derived from other variables in the job ad. This may incur processing errors.
- Different levels of information will be available for different vacancies (for example, vacancies advertised through agencies will usually not have the name of the employing business, which will affect the quality of industry coding.
- Information extracted from the unstructured parts of job advertisement may be subject to errors
- On-line job ads are a stock of advertised vacancies, which does not correspond directly to the concept of an unfilled job vacancy as measured by the JVS.
## UNECE Big Data Quality Framework

<table>
<thead>
<tr>
<th>Hyper-dimension</th>
<th>Source</th>
<th>Metadata</th>
<th>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Input</strong></td>
<td>Institutional/Business environment</td>
<td>Complexity</td>
<td>Accuracy and Selectivity</td>
</tr>
<tr>
<td></td>
<td>Privacy and Security</td>
<td>Completeness</td>
<td>Linkability</td>
</tr>
<tr>
<td><strong>Throughput</strong></td>
<td></td>
<td>Usability</td>
<td>Coherence / Consistency</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Time-related factors</td>
<td>Validity</td>
</tr>
<tr>
<td><strong>Output</strong></td>
<td></td>
<td>Linkability</td>
<td>Accuracy and Selectivity</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Coherence / Consistency</td>
<td>Validity</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Validity</td>
<td>Coherence / Linkability</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Time-related factors</td>
<td>Coherence / Consistency</td>
</tr>
</tbody>
</table>

**Phase of Business Process**

- **Processing Principles**:
  1. System Independence
  2. Steady States
  3. Application of quality gates

- **Complexity**
- **Accessibility and Clarity**
- **Relevance**
Two-Phase Quality Framework for Data Integration

Phase 1

- Measurement (variables):
  - Target concept
  - Target measure
  - Obtained measure
  - Edited measure
  - Single-source (primary) micro data

- Representation (objects):
  - Target set
  - Accessible set
  - Accessed set
  - Observed set
  - Frame error
  - Selection error
  - Missing redundancy

Phase 2

- Measurement (variables):
  - Target concept
  - Harmonized measures
  - Reclassified measures
  - Adjusted measures

- Representation (units):
  - Target population
  - Linked sets
  - Aligned sets
  - Statistical units
  - Integrated (secondary) micro data

Li-Chun Zhang, 2012
Points for Discussion

- Does the CEDEFOP web scraping system have an overall system of quality management?
- Could this framework, or elements of it, be incorporated into the overall system?
- What metadata will be available to users?
- Is there an established set of use cases?
- What scope is there for NSIs assist with validation?
  - Job vacancy survey
  - Alternative OJV sources
  - Practicalities (e.g. sharing data)
Input Phase

- Much of this is already well documented and so detailed discussion is probably not necessary.
Throughput Phase – Processing Principles

• System Independence
• Steady States
• Quality Gates
Throughput Phase (Checklist)

- What steps taken to check that all target data is collected?
- What backup systems are in place for when scraping fails?
- Are there processes in place to establish if data has been missed and if so, is it possible to estimate how much data has been missed?
- What changes are made to data during pre-processing (e.g. cleaning/standardising variables). Are metrics available to determine the quality of this processing?
- What steps is the CEDEFOP system taking to identify multiple jobs per job ad?
- Is it possible to identify and remove ghost vacancies?
- What methods are used for deduplicating data? How are the deduplicated results tested? Can the quality of deduplication be measured (e.g. confusion matrices)?
- What methods are used for extracting data from unstructured text? What metrics are available to measuring the quality of this processing?
- What methods are used for coding variables and deriving other data (e.g. for occupation, industry, geography)? Are there metrics available to measure quality?
Output Phase

- Are there any institutional barriers to publishing data from the Cedefop system?
- What metadata will be available?
- Will any specific quality measures be published? If so, how frequently?
- What guidance will be available to users, particularly in terms of what the data can be used for?