The definition of Open Data [1] can be summed up in the following statement:

**Freely available data, published in a machine readable format which allows everybody to do everything without restrictions at anytime.**

Open Data is mainly published on **Open Data Portals**:

- **Datasets** hold metadata (i.e., basic information) of resources, e.g., authorship or licensing information.
- **Resources** are available for access or download in one or more formats, e.g., CSV, PDF, and Microsoft Excel.
- Data portals offer an API to retrieve datasets.

**Motivation:**

Poor quality and high heterogeneity of metadata and data sources in Open Data portals is a serious risk that could disrupt the success of the Open Data movement.

**Problem Statement:**

There exist no objective and comprehensive reports about the actual quality and evolution of portals.

### Findings

#### Usage and Completeness:

- Extra keys (green) partially unused: Bottom left portals have low values
- Indication that there are redundant metadata keys

#### Evolution:

- Number of datasets is steadily increasing
- Quality Metrics change with new datasets
- License openness (yellow line) is high but decreasing

#### Heterogeneity:

- Non-machine-readable formats (e.g., PDF) are widely used
- 218 different format descriptions and 1516 different license IDs

#### Resource Overlap:

- 3 large harvesting portals (e.g., publicdata.eu)
- Also active cross-publishing

### Further Work

**Metadata homogenization:** Mapping of different metadata formats to integrate other portal software.

**Automated Quality Improvement:** Use acquired data for automated correction.

### References

- Jürgen Umbrich, Sebastian Neumaier, and Axel Polleres. Best paper award.

### Poor quality

Data portals offer an assessment and evolution of (meta-)data quality in the Open Data Landscape. Sebastian Neumaier