WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

Open Data Hopes and Fears

Determining the barriers of Open Data

Martin Beno

Kathrin Figl

Jürgen Umbrich

Axel Polleres

Open Data defined

3 Crucial principles

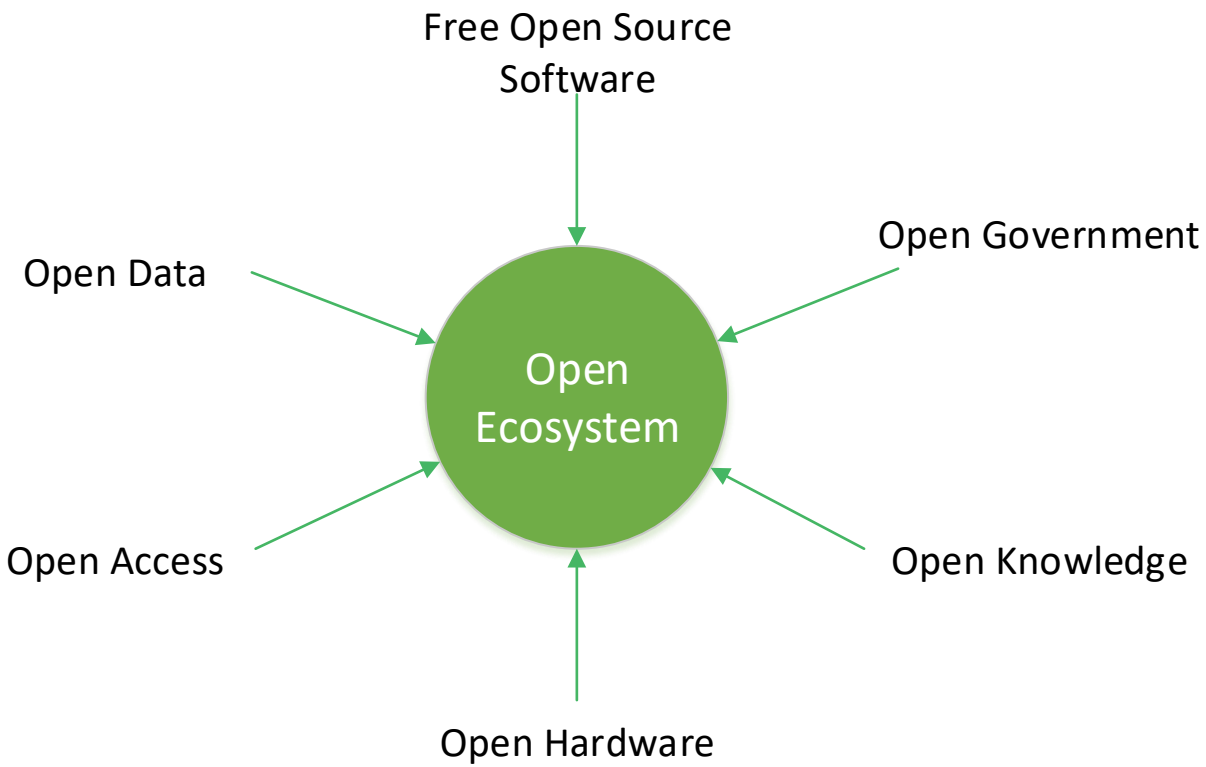
- **Availability and access**: The work must be provided as a whole and at no more than a reasonable one-time reproduction cost, and should be downloadable via the Internet without charge. Any additional information necessary for license compliance (such as names of contributors required for compliance with attribution requirements) must also accompany the work.
- **Re-use and Redistribution**: The data must be provided under terms that permit re-use and redistribution including the intermixing with other datasets. The work **must** be published in a **machine-readable** form.
- **Universal Participation**: Everyone must be able to use, re-use and redistribute.

- Definition by OKFN: http://opendatahandbook.org/guide/en/what-is-open-data/

Déjà vu?

- The concept of sharing, re-use and redistribution is a long-established philosophy in IT.
- The Open Data definition is comparable to Free and Open Source Software as defined by the Free Software Foundation:
  - The freedom to run the program as you wish, for any purpose
  - The freedom to study how the program works, and change it so it does your computing as you wish
  - The freedom to redistribute copies
  - The freedom to distribute copies of your modified versions to others

Open Ecosystem



Success stories (FOSS)

Most of IT infrastructure today would not work without free software.

Even Microsoft relies on free software in certain parts of their organization.

The server-market, which has once been dominated by proprietary software now mostly runs on free software.

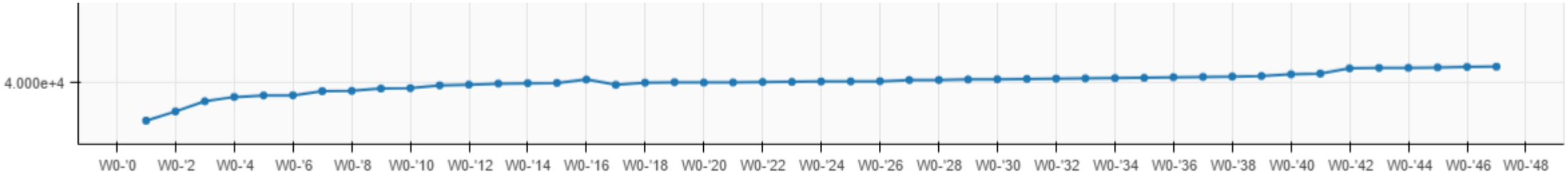
Hope: Will Open Data enjoy success stories of such magnitude in the future?

Motivation

- Transparency in government
  - More openness, reduced corruption, increased trust and relationship between the citizens and the government, improved and streamlined services.
- Enabling independent developers to develop value-added services and applications.
  - Creating applications with social and economic value
  - Stimulating economic growth.
- Promoting collaboration of citizens with the government
  - Citizen participation is the foundation of democracy

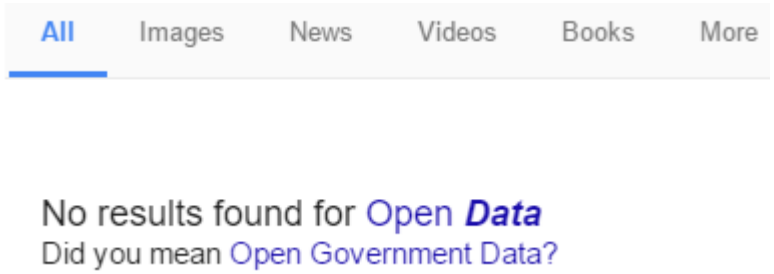
Open Data is a rising trend

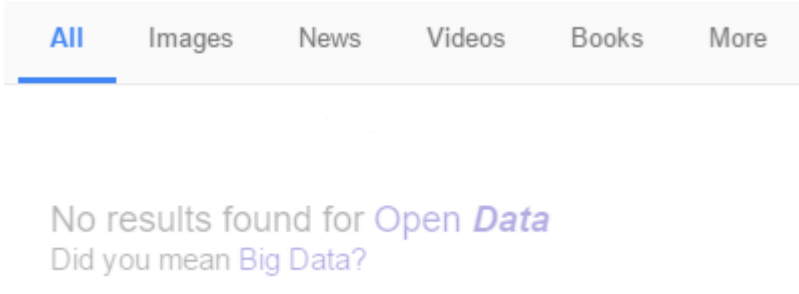
- Every single EU member country has an Open Data platform available in some form.
- Open Data Portalwatch currently tracking 259 from all over the world.
- Number of published datasets has been steadily rising:
  - Data.gov.uk:
  - June 2016 = 32766 datasets
  - May 2017 = 43065 datasets



Why are we talking about barriers again?

- Mostly focused on Open **Government** Data. Participation from private entities has been rather lacking.
- Data is being used, but not enough.
- There is still a general unawareness about what Open Data actually is.





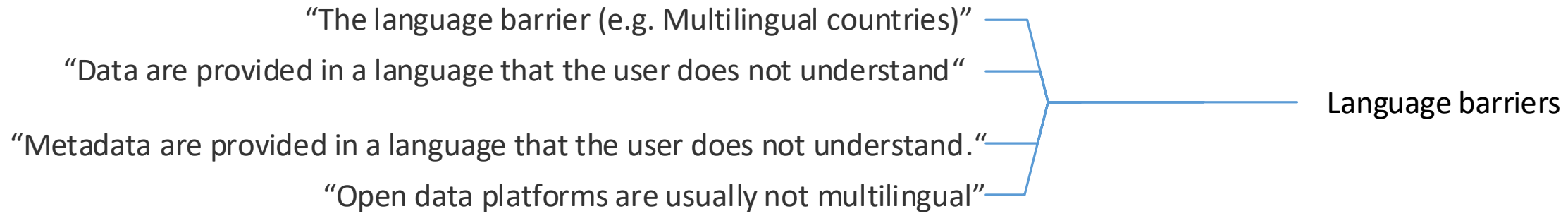
Disclaimer: Not actual search results

Barriers in the context of Open Data

- "**A circumstance or obstacle that** keeps people or things apart or **prevents** communication or **progress**." – Oxford Dictionary (Emphasis mine).
- We therefore see barriers merely as obstacles in the successful progress of Open Data. Obstacles, which we can overcome.
- "*Indeed the challenges and constraints faced by re-users of public data differ from the ones encountered by the public data providers*."
  - – Risk Analysis to Overcome Barriers to Open Data (Martin, Foulonneau, Turki and Ihadjadene)
- Barriers have been split into 3 categories:
  - User specific (Open Data portals, data quality, user legal constraints)
  - Provider specific (Strategic and business, privacy and security, provider legal constraints, technical barriers)
  - Both user and provider (Knowledge and experience)

Determining the barriers

- Barriers have been identified by various studies, yet their relevance and importance has not been researched. This leads to inconsistencies whether an issue really presents a barrier for most Open Data users and publishers.
- Initially, nearly a hundred potential barriers were found and documented.
- We identified duplicates/analogous barriers and merged them into one.



- List of barriers was then further reduced to the most significant barriers. Including all barriers in the questionnaire would result in a smaller response rate

User Barriers – Data portals

Data portals as the main source of open datasets.

However, many portals are making the use and re-use of data difficult for the users:

- Registration needed
- Data is behind a paywall
- No information about the quality or content of the datasets
- Difficult browsing and searching

Data retrieval itself is often problematic due to slow or non-existing APIs

- Although the prevalence of APIs has been increasing ever since standardized platforms such as CKAN became the norm.
- CKAN is the most widely used Open Data portal platform according to the Open Data Portalwatch

Software	Count
CKAN	151
OpenDataS...	12
Socrata	101
XMLDCAT	2

User Barriers – Data quality

Data is still being published in proprietary formats (.docx, xlsx)

Even worse in non-machine-readable formats (PDF)

Even if machine readable, often not processable by a machine without fixing the dataset first.

- This results in a frustrated data user

```
ERROR:  invalid input syntax
CONTEXT:  COPY dataset, line 3, column sequence
```

User Barriers – Legal constraints

The **open definition** is simple enough. Everyone can reuse, modify and redistribute. Therefore, the users do not fear any legal consequences right?

Wrong! Data is occasionally still published with restrictive licenses or no license at all making the rights of use and reuse unclear.

Parallel: A complex licensing situation is nothing new in Open Source Software

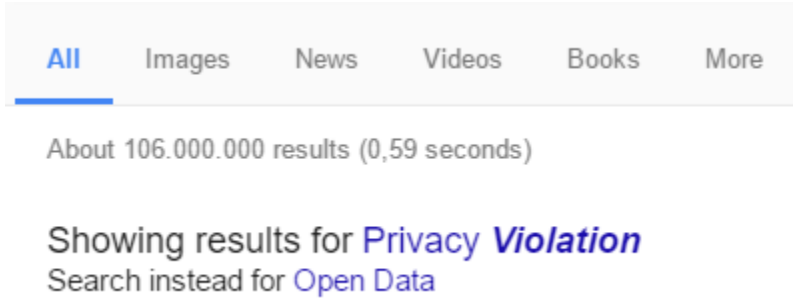
- The Open Source Initiative currently lists 78 different licenses.
- Published software projects often do not specify a license.

Finally: Threat of lawsuits.

- Open Data users are often independent developers who do not have the resources to fight a lawsuit.

The rights of the users need to be **clearly defined**!

Provider Barriers – Privacy and security



Private and sensitive data need to be protected.

Loss of control about the released information is a threat to potential publishers.

Data release might also hurt 3rd parties (Data about policy plans of a city released => as a result property value may decrease)

Security vulnerabilities of the publishing platforms, such as critical security bugs in CKAN.

In certain cases, the APIs might also be used to perform a denial of service attack.

Provider Barriers – Strategic and Business

Open Data is often not a strategic priority.

Data publishing is seen as a potential loss of profit.

- Current business model relies on the sale of data
- Similarly in Open Source Software, for many companies, the source code enables value capture if it is used in products, which they can sell.

Open Data needs to be updated and maintained, which implies increased costs

Especially difficult to implement Open Data publishing in the private sector, where the companies do not see direct monetary benefits and thus, lose interest in Open Data.

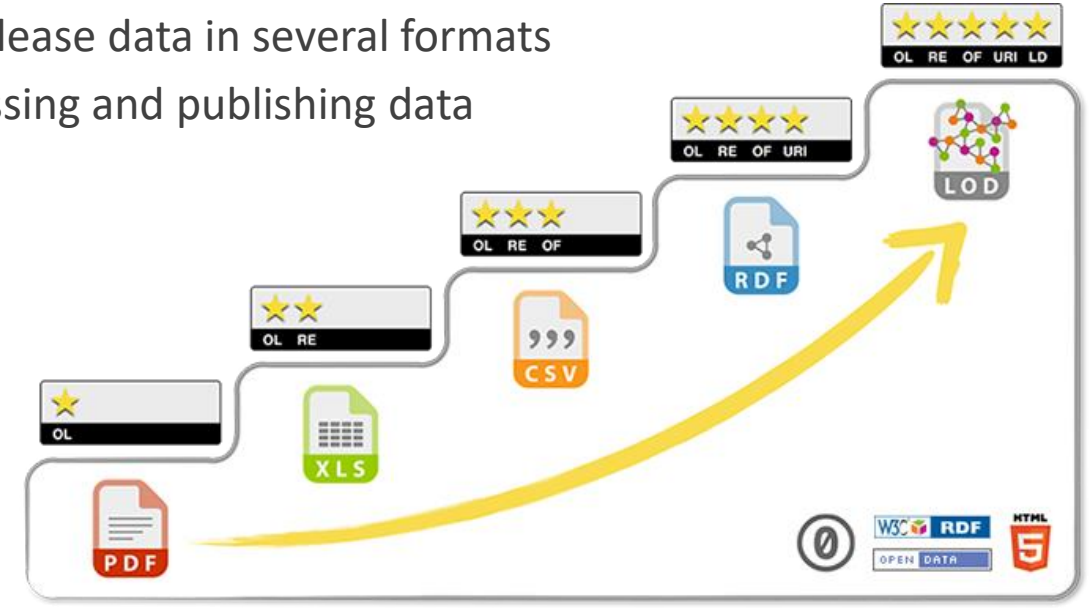
Provider Barriers – Legal constraints and technical barriers

Legal constraints:

- Providers face the threat of lawsuits when private sensitive information is released.
- Unclear data ownership.

Technical barriers:

- It is often expected of the provider to release data in several formats
- There is no software standard for processing and publishing data



User and provider

Open Source Software is often rejected, because in contrast to proprietary software, no helpdesk or commercial support is available (Although, there are many exceptions).

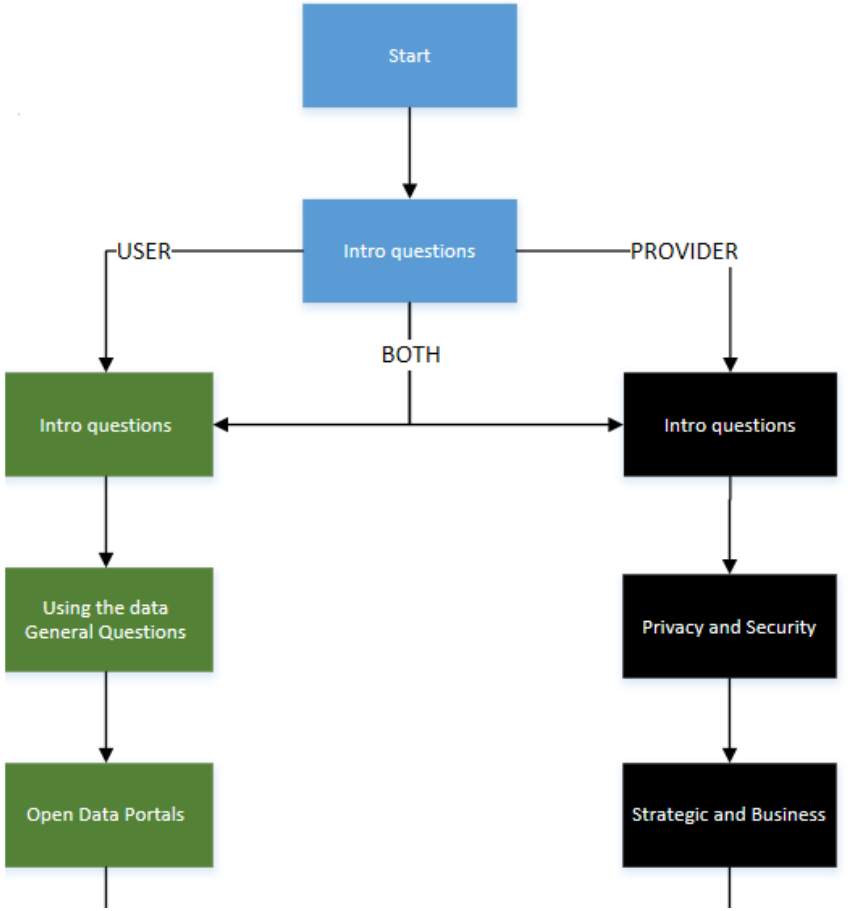
Similarly, Open Data providers publish the data “as-is” and it is often left to the user to figure out how to use the data and for what purpose.

There is a general lack of documentation and guiding principles.

Standardized guidelines would not only help users use the data, but would also ease the publishing process for the provider.

Survey

- Online only
- Modular structure for users and providers of data
- Importance of barriers measured using a 5 –point Likert-scale
  - 1 - Not a barrier
  - 2 – Somewhat of a barrier (Still possible to use/publish the data)
  - 3 – Moderate barrier (Made it difficult to use/publish the data)
  - 4 – Serious barrier (Made it extremely difficult to use/publish the data)
  - 5 – Extreme barrier (Completely prevented me from using/publishing the data)
  - + An open text field for each category of the barriers



Participation

Survey was launched in November 2015, active until March 2016.

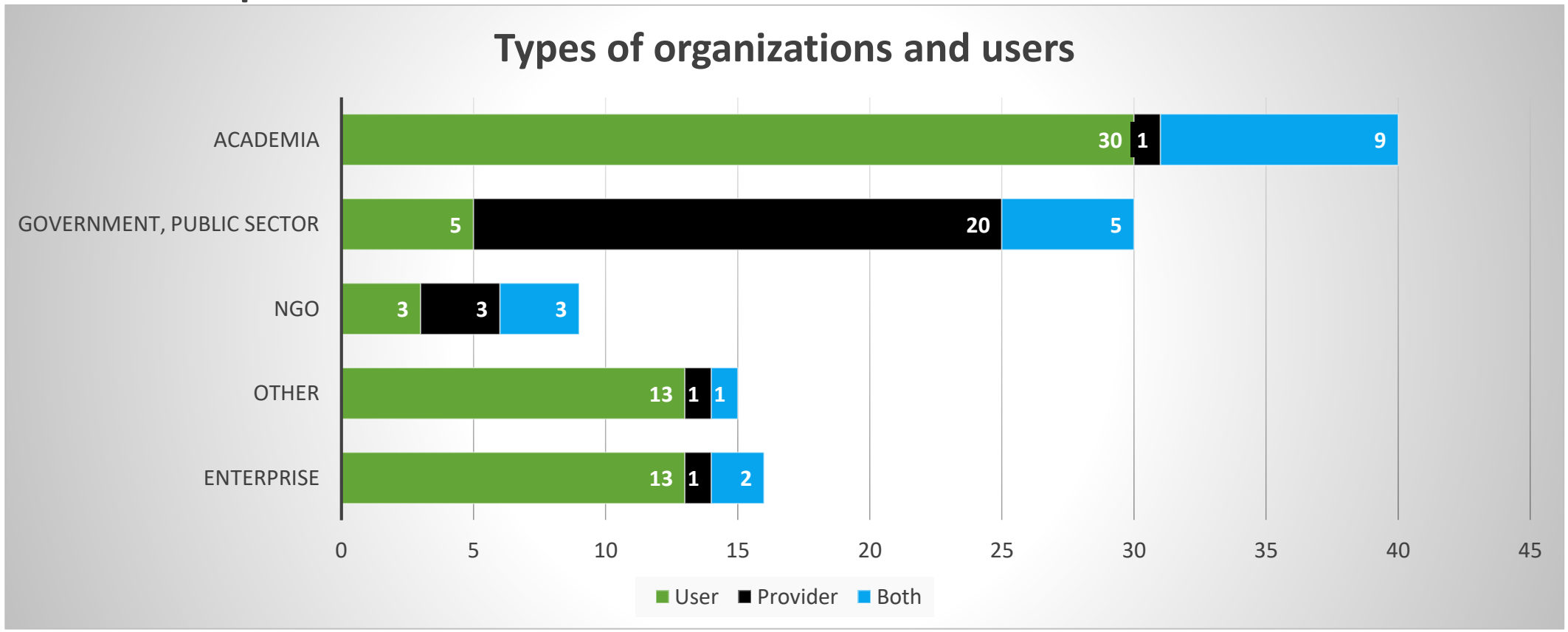
Limited to Austria

Distribution channels:

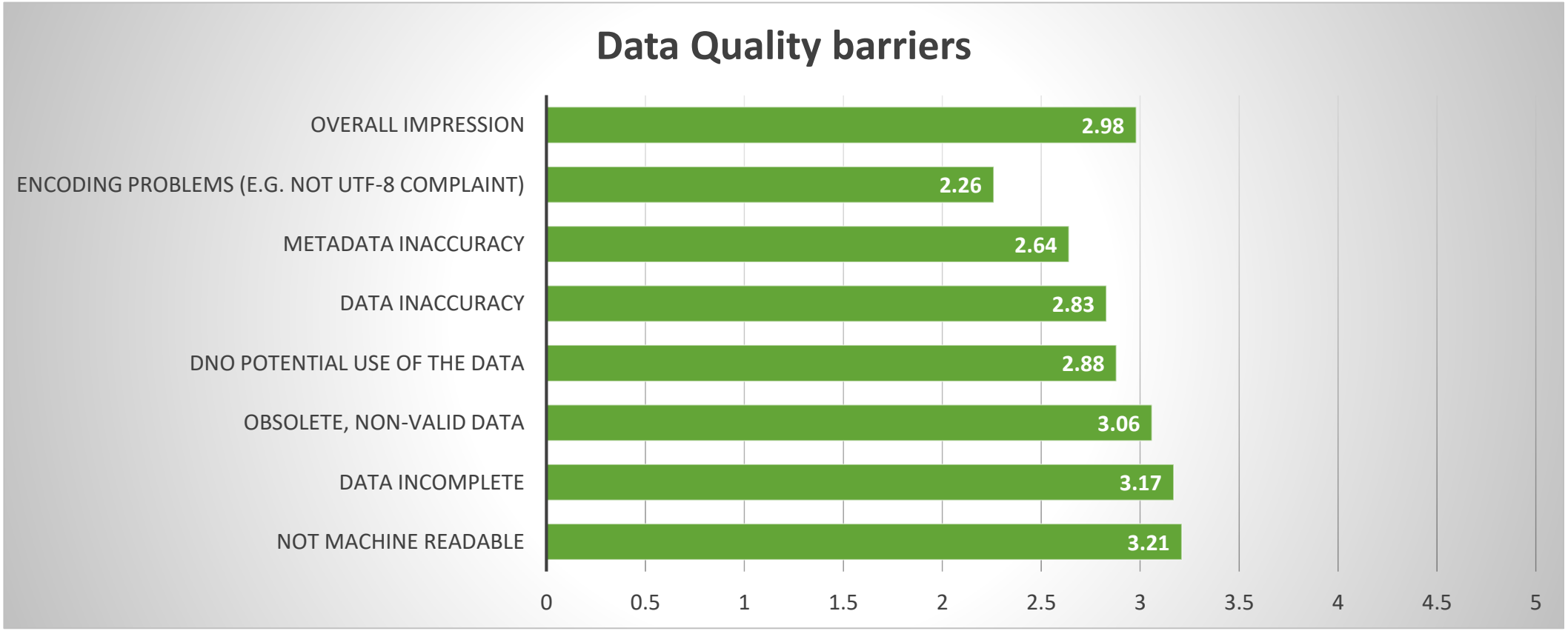
- Social media
- Meetups
- Austrian Open Government Data cooperation
- Directly contacting data-users and publishers.

Full responses	Incomplete/Empty responses	∑
110	200	310

Participants



Results – Data quality (Users)

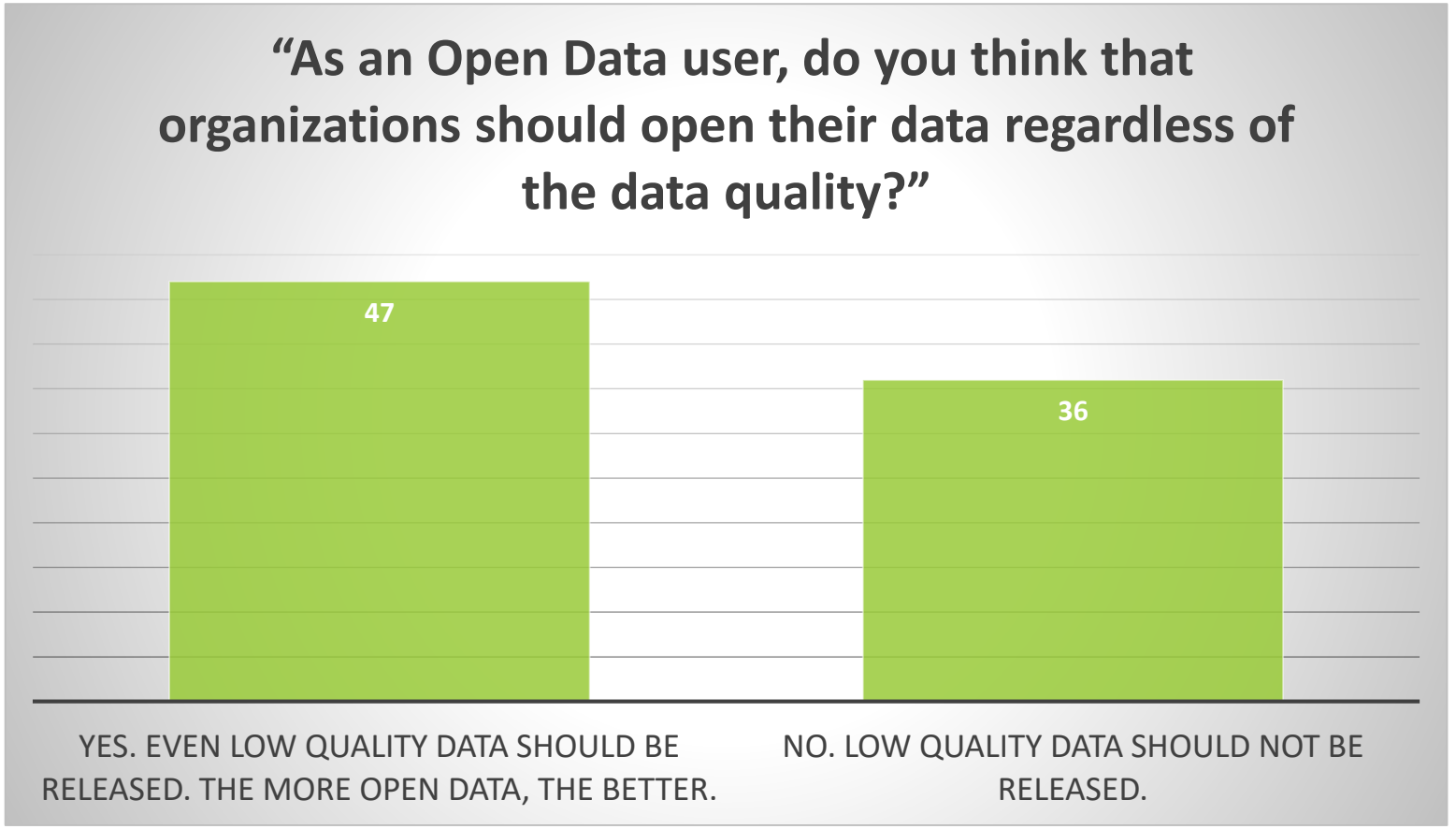


Results – Data quality (User comments)

“Some federal states publish excellent datasets on individual topics, but not all of them publish the same data. There is no Austria-wide comparability. In addition, the data is still being published in pdf format”

“Metadata is often no help at all to learn more about the dataset. Either they are incomplete or missing completely”

Results – Data quality (Users)



By Comparison: Opening Source Code

It could be argued that many projects published on Github may not be of the highest quality or even usable at all

And others might not be particularly useful…

README.md

Hello, World!

Hello world in every programming language.

Add 'Hello, World!' in Hodor. #330

Merged leachim6 merged 1 commit into `leachim6:master` from `TeHMoroS:new/hodor` on Feb 12

Conversation 2 Commits 1 Files changed 1

Changes from all commits 1 file +1 −0

1 h/hodor.hd

```
@@ -0,0 +1 @@
+hodor.hod('Hhodor? Hodor!? Hodor!? o, Hooodorrhodor orHodor!? d!');
```

Yet, they are being released. Is this the right approach for Open Data as well?

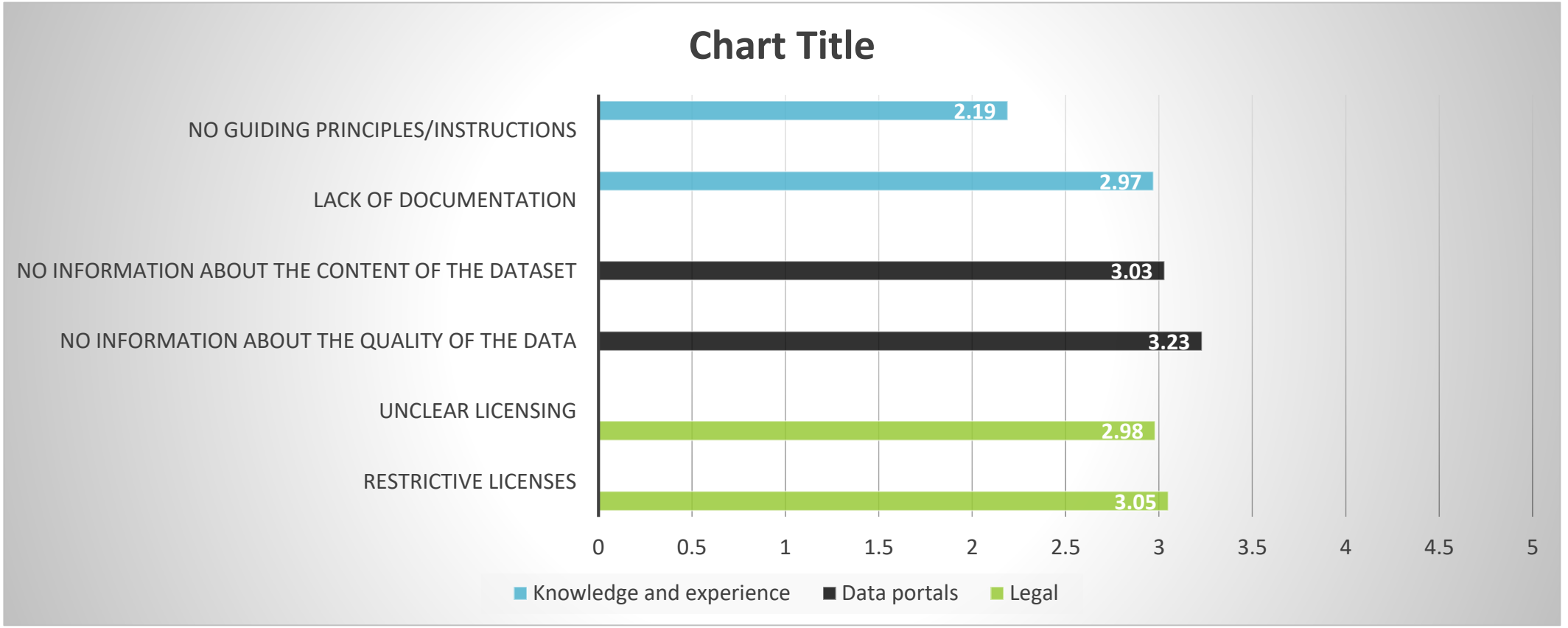
Results – Data quality (Users)

Interesting to note that more than half of the participants (65%) indicated that organizations publishing Open Data should adhere to a certain set of quality standards and rules.

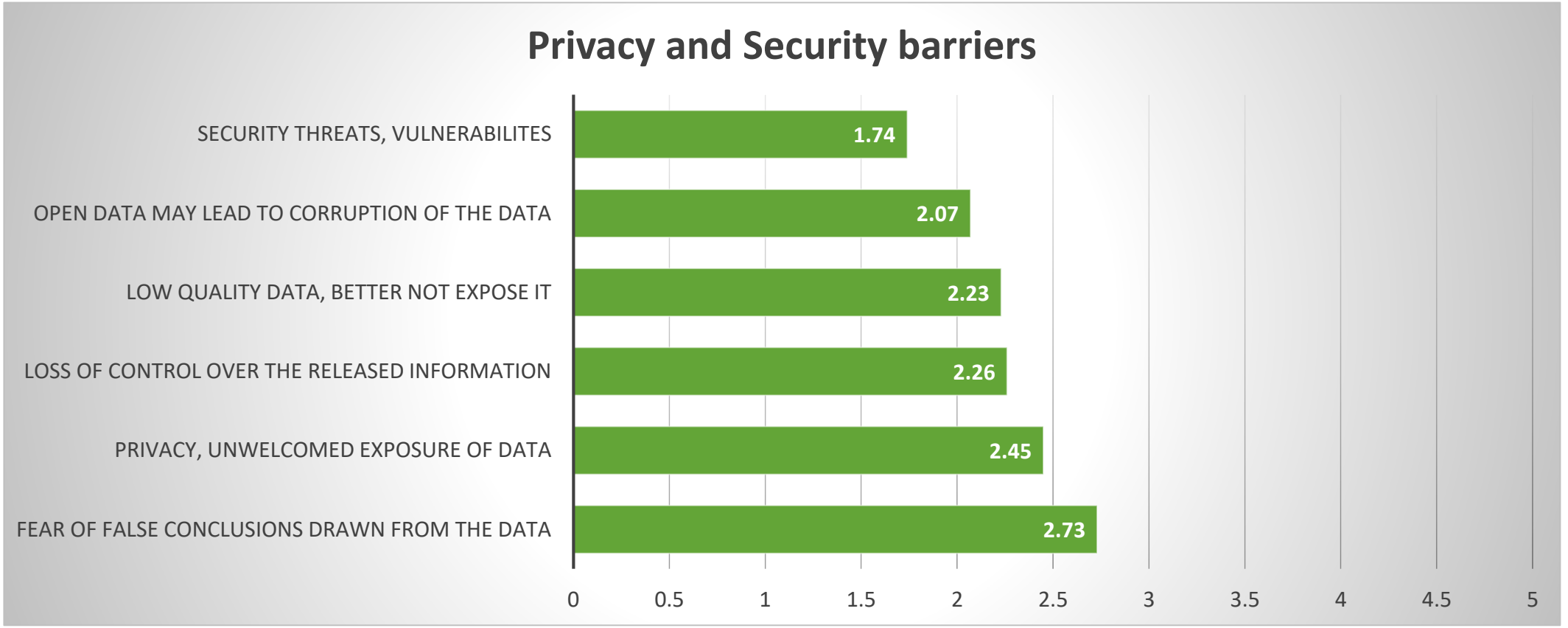
We asked the participants to provide some examples:

- “Minimum standard format for all data. If the data is available only in ‘exotic’ formats, especially when these are not open formats, a version in a minimum standard format must be available as well”
- “Data should always be published in a machine-readable format. That means no PDFs for instance.”
- “There should not be any rules, rather recommendations, such as ‘use utf-8’”

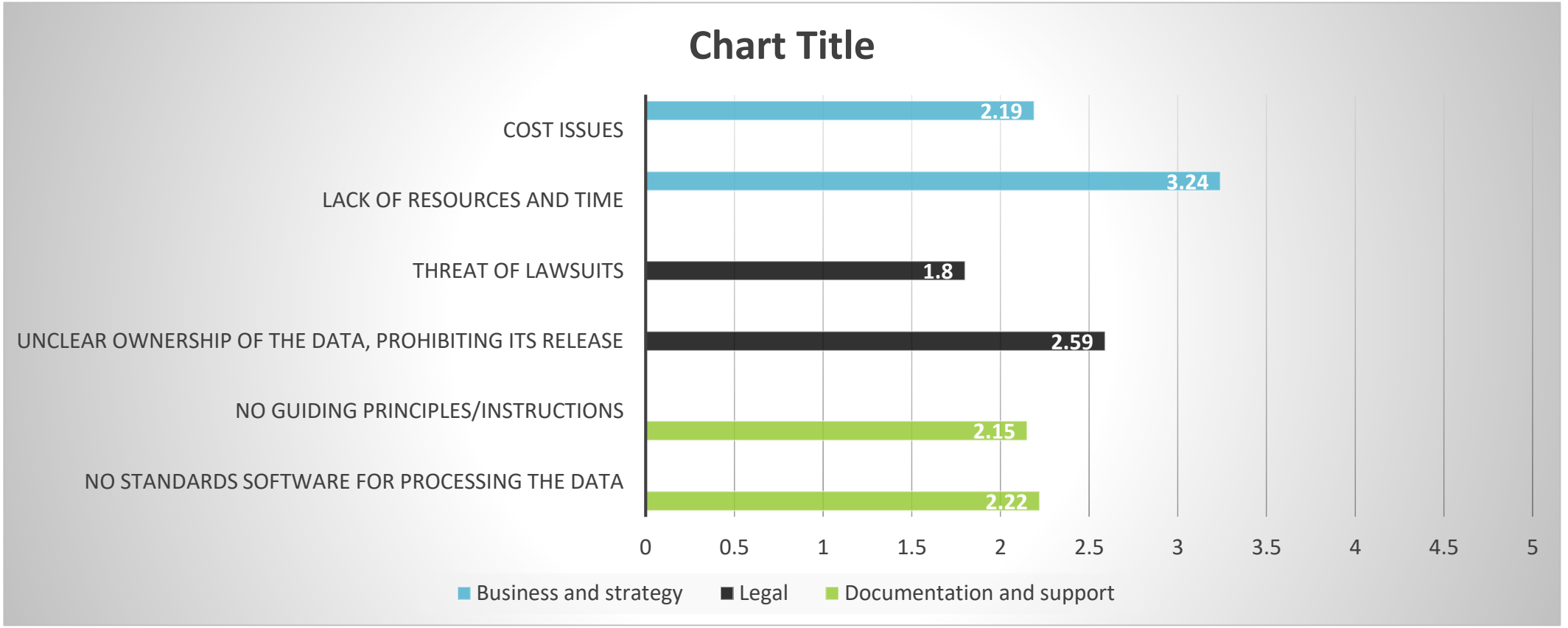
Results – Other top barriers (Users)



Results – Privacy and Security (Providers)



Results – Other top barriers (Providers)



Conclusion and discussion

Most significant barriers reported in the literature have been confirmed in our survey.

We were also able to identify previously unreported barriers such as:

- Slow responsiveness of Open Data portals
- Lack of data harmonization
- Encoding issues (Data not being UTF-8 complaint)
- Licenses are not machine-readable

Limitations

Potential bias – Many respondents already experienced Open Data users/publishers.

- We were unable to discover any new barriers faced by newcomers to Open Data

Questionnaire was sent to Austrian respondents only.

- Our findings might not be transferable to other countries.

The majority of respondents are from the public sector.

- Future research: Open Data in the private sector.

Other possible future research:

- Relationship between Open Source and Open Data. Can any mitigation strategies used in Open Source be applied to Open Data?

Online results

Available at https://odsurvey.ai.wu.ac.at/results

Other projects: http://data.wu.ac.at

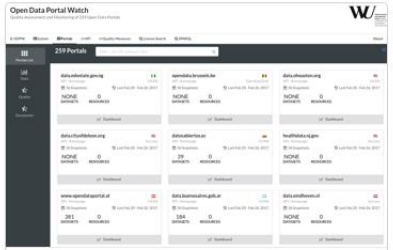
Projects

WU Open Data Portal

WU lectures, rooms and organizations

data.wu.ac.at is an Open Data portal where you can find data about lectures, rooms and organizations at WU.

121 datasets

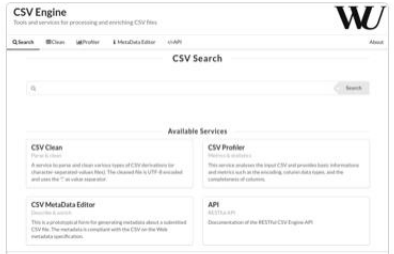


Open Data Portal Watch

Monitoring & exposing portals' metadata

Open Data Portal Watch assesses the evolution of the (meta) data quality of about 260 Open Data portals over since September 2014.

259 portals



CSV Engine

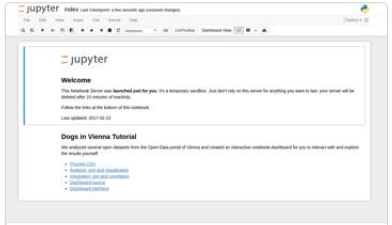
Search & enrich CSVs

The CSV Engine is a collection of tools and services for processing and enriching CSV files.

DBpedia Wayback Machine

Extract past DBpedia versions

The DBpedia Wayback Machine aims at providing the wayback functionality for DBpedia based on the revisions of their Wikipedia article.



Jupyter Notebook Server

Programming & Documentation

Notebook documents are documents which contain both computer code (e.g. python) and human-readable rich text elements.

Only available within local WU Vienna network

Open Data Assistant

Multilingual search chatbot for Open Data

The assistant supports cross-lingual search by integrating metadata of 7 open data portals in 7 different languages.

The way forward

Mitigation of barriers is crucial for the success of Open Data.

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

EFMD EQUIS ACCREDITED

Thank You for Your attention

Questions? Discussion?