

Helen Lippell

Taxonomy consultant

EIGHT CAUSES OF BAD DATA QUALITY

(And what you can do about it)



THE EUROPEAN DATA CONFERENCE ON REFERENCE DATA AND SEMANTICS

Cause 1: Lack of agreed definitions for concepts and categories

Organisation:

Public sector government body dealing with consumer financial services issues

Key lines of attack:

- Buy-in from senior management
- Securing resources
- Working across the whole organisation



Cause 2:

Lack of business processes to manage vocabularies and definitions

Organisation:

Large global company in the entertainment sector

Key lines of attack:

- Working 'bottom-up'
- Working with content creators and taggers
- Documentation improvements



Cause 3: Different audiences use different terminology for the same concepts

Organisation:

UK healthcare body who manage education and training for health professionals and students

Key lines of attack:

- Organisation-wide taxonomy
- Researching different user needs
- Capturing full spectrum of technical and common vocabulary used in the domain



Cause 4:

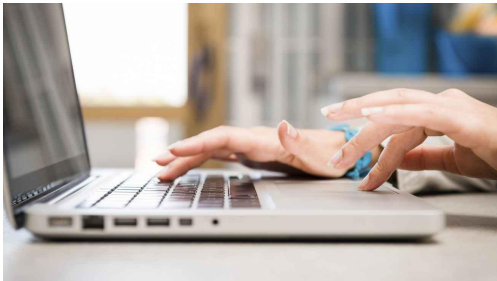
Domain is not fully understood by the organisation and its customers

Organisation:

Government organisation responsible for producing digital information for the general public

Key lines of attack:

- Learning about the domain yourself
- Working with content producers to understand audience needs
- Improving information architecture and user journeys
- Mapping technical jargon to the language the general public use



Cause 5: Language is ambiguous

Organisation:

Business information news publisher

Key lines of attack:

- Disambiguating homonyms in the taxonomies
- Teaching automatic classification tools to better identify the context of ambiguous words



Cause 6: Poor quality source data

Organisation:

News agency supplying wide range of publications and broadcasters with syndicated content

Key lines of attack:

- Acknowledging limitations of public linked open data
- Training journalists to check and update data
- Improving data editing user interfaces



Cause 7: Data entropy

Organisation:

Vendor of bespoke information services covering the UK public sector

Key lines of attack:

- Making business aware of the impact of their data entropy
- Training staff to tag, manage and organise data more accurately
- Auditing data regularly



Cause 8: No data!

Organisation:

Supermarket with physical stores, online and ecommerce services

Key lines of attack:

- Evaluating what is important for the organisation to know about its digital assets
- Using automated classification to get basic metadata
- Investing in people to manage the asset repository long-term



Thank you!



- <https://twitter.com/octodude>
- [Helen Lippell | LinkedIn](#)
- <https://thesearchnetwork.com/>
- <https://www.taxonomybootcamp.com/London/2021/>

ENDORSE.

THE EUROPEAN DATA CONFERENCE ON REFERENCE DATA AND SEMANTICS



Publications Office
of the European Union

