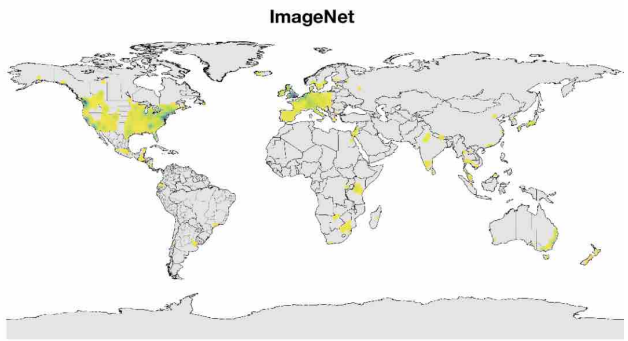
Data Mining and Landscape Architecting: On Creating Multimodal Resources

Desmond Elliott

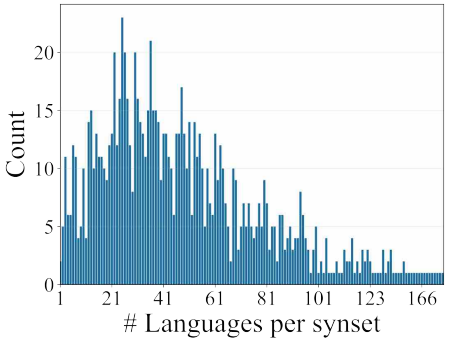
Language and Multimodal Processing Group
Department of Computer Science
University of Copenhagen

The Rise of Culture x Multimodality

- It is important to work with resources that represent the diversity of the cultures for reasons of heritage, inclusion, and equal evaluation



Density map of geographical distribution of images in ImageNet (DeVries+, 2019)



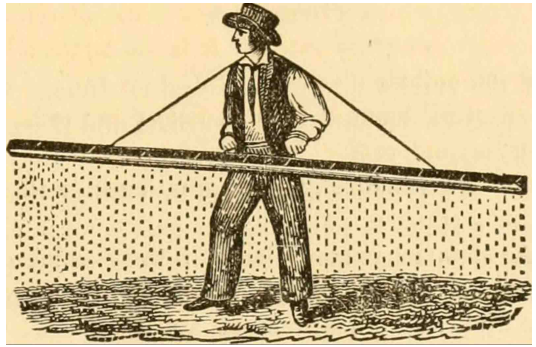
Lack of WordNet concept coverage in multilingual Wikipedia articles (Liu+ 2021)

So how should we collect resources?

Two Basic Types of Resource

1. Mined from existing data sources

2. Created from scratch



Four Collector Stereotypes

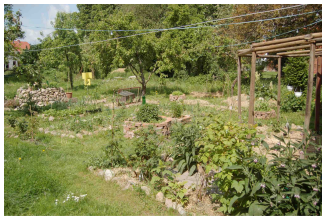
Data Miner

Large-scale extraction

Detectorist

Focused search for artefacts

Permaculturist



Minimal intervention

Baroque

Painstaking design

Five Key Factors

1. Scope

   What type of data are you aiming to collect?
2. Annotator Relationship

   How are you working together?
3. Images

   What type of images?
4. Texts

   What type of texts?
5. Tooling

Use Cases

Multi30K (2015)

1. The two men on the scaffolding are helping to build a red brick wall.
2. Zwei Mauerer mauern ein Haus zusammen.

MaRVL (2021)

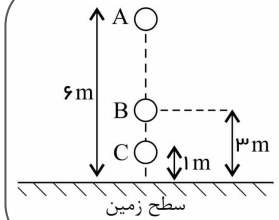
(b) *Görsellerden birinde dizlerinde kanun bulunan birden çok insan var.* ("In one of the images, there are multiple people with qanuns on their knees.", concept: *Kanun (çalgı)* (QANUN, a popular instrument in Turkey), label: TRUE)

FoodieQA (2024)

以下菜品是哪个地区的特色菜？
Which region is this food a specialty?

A 江苏 (Jiangsu)
B 京津 (Beijing & Tianjin)
C 香港 (Hong Kong)
D 广西 (Guangxi)

Kaleidoscope (2025)



University Entrance Exam | Physics

Question:
گوله‌ای مسیری مطابق شکل را طی می‌کند. کار نیروی وزن از A تا B چند برابر کار نیروی وزن از B تا C است؟

Options:
A. 1.5
B. 1.25
C. 1.2
D. 2

1. Scope

- What type of data are you aiming to collect?
- Does your dataset come from a broad collection of concepts / domains, or is it a deep dive into a specific subject matter?

LAION
Multi30K
WiT

MaRVL
CulturalVQA
Kaleidoscope
FoodieQA

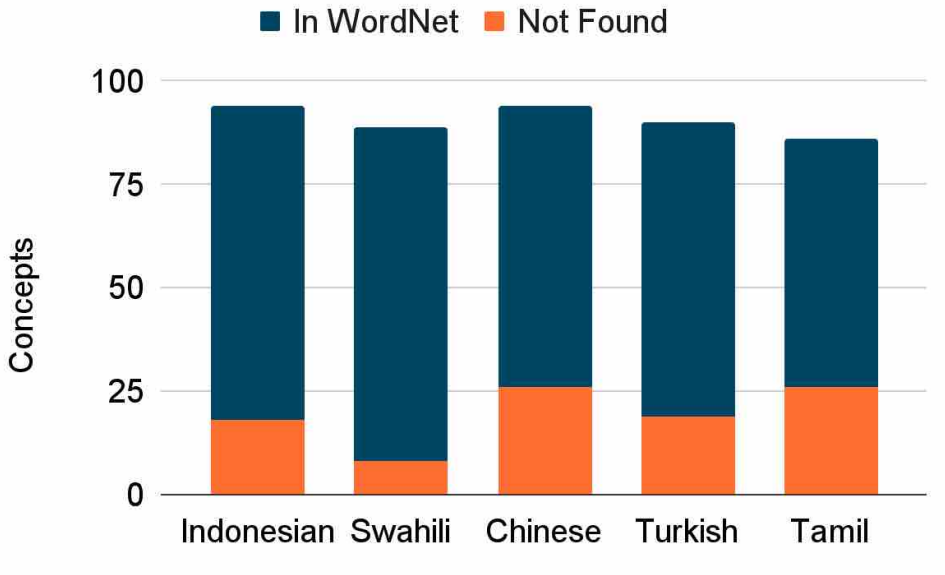
Broad
Deep

Multi30K: Replicate

A brown dog is running after the black dog.

Ein schwarzer und ein brauner Hund rennen auf steinigem Boden aufeinander zu

MaRVL: Concepts beyond ImageNet



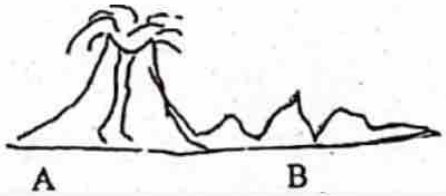
FoodieQA: Fine-grained Chinese Food

Kaleidoscope: Exams

Geography

Question #5357

'A' চিত্রের প্রদর্শিত ঘটনাটি কী?



1) সঞ্চয়

2) অগ্ন্যুৎপাত

3) ক্ষয়

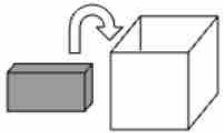
4) ভূমিগল

Social Sciences | Bengali | Bangladesh

Mathematics

Question #9326

Alice a beaucoup de blocs, tous de dimensions 1 cm x 2 cm x 4 cm. Elle en met le plus possible dans une boîte de dimensions 4 cm x 4 cm x 4 cm sans qu'aucun ne dépasse. Combien a-t-elle rangé de blocs dans la boîte ?



1) 6

2) 7

3) 8

4) 9

STEM | French | France

2. Annotator Relationship

- What is the relationship between the collectors and the team?
- Are you working together towards a shared goal, or are they hired?

CulturalVQA

MaRVL

FoodieQA

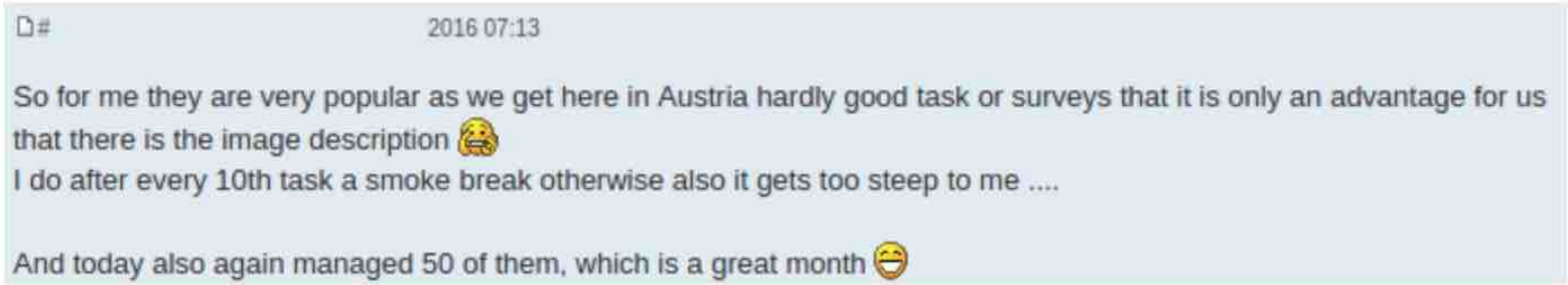
Multi30K

Kaleidoscope

LAION

Hired | Volunteers | Collaborators | Indirect

Multi30K: Hired



2016 07:13

So for me they are very popular as we get here in Austria hardly good task or surveys that it is only an advantage for us that there is the image description
I do after every 10th task a smoke break otherwise also it gets too steep to me ....

And today also again managed 50 of them, which is a great month

Source: machine translated extract from crowdworker discussion forum about our crowdsourcing task.

Multi30K: Speculations

2016 07:14

> **I think it's about how different people perceive images ...** *... so what is the thing that strikes nearly all or where the focus is. Possibly even something for an AI ... laughing Since many photos which you describe I have for example, the woman at the end of the stairs lying etc.*

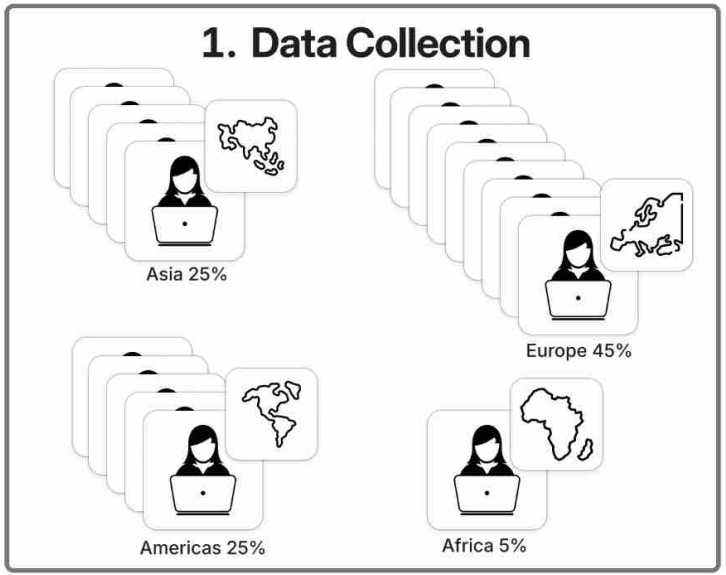
Yes good chance when I look so what he has been previously employed. I find something fascinating

way, I have just the absolutely is a dark image of a man apparently in a halfpipe and jumps, but you can not identify absolutely with what if skates or board or something else ... I hate something xD

Source: machine translated extract from crowdworker discussion forum about our crowdsourcing task.

Kaleidoscope: Collaboration

- Open-science collaboration with an incredible community of early-career scholars
- Co-authorship offered in exchange for collecting data above a threshold



3. Images

- What type of images are in your resource?
- From which sources are you collecting them?
- What are the licenses of the images?
- How are you protecting PII?

Kaleidoscope

Not Photos

Photos

Multi30K

MaRVL

CulturalVQA

Curated

MaRVL

Private Data

FoodieQA

Web Scrape

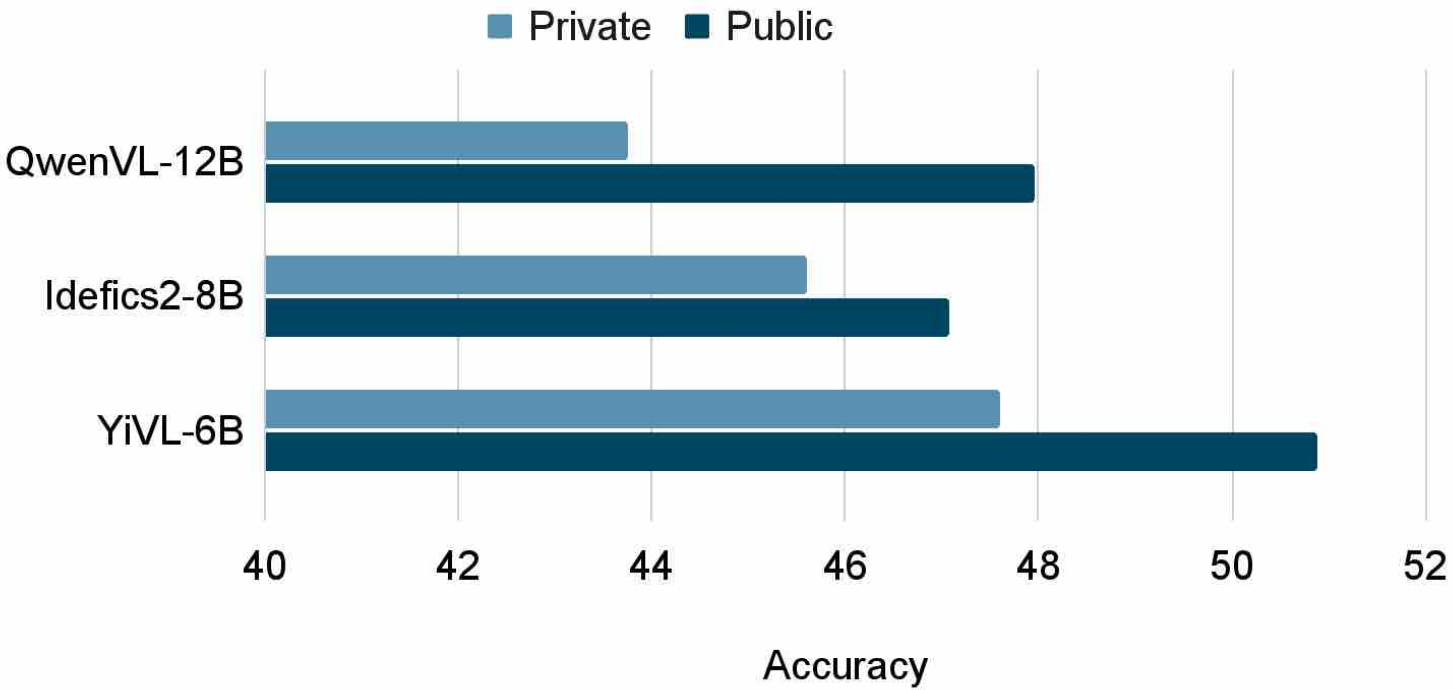
LAION

Multi30K

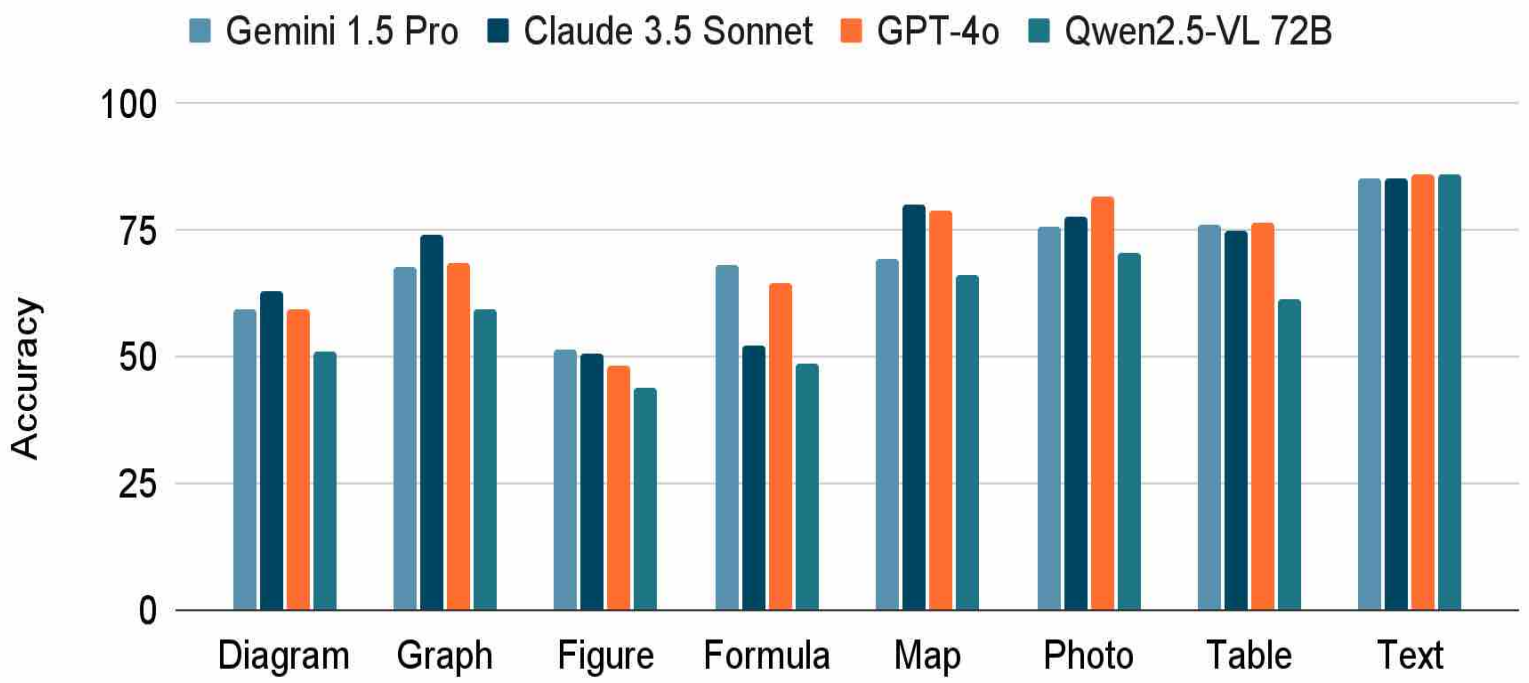
Types

Sources

FoodieQA: Private Images



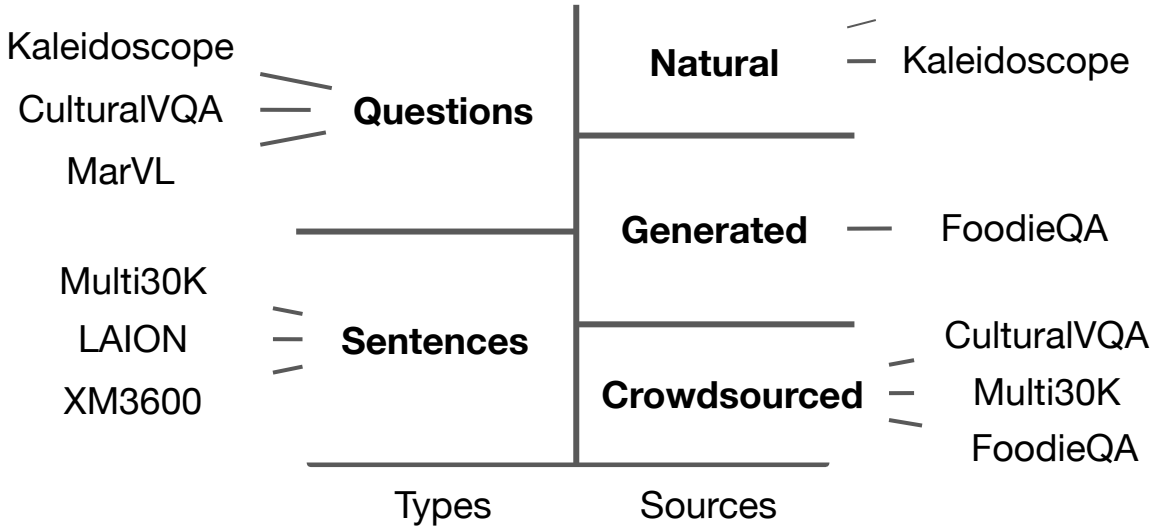
Kaleidoscope: Eight Image Types



4. Texts

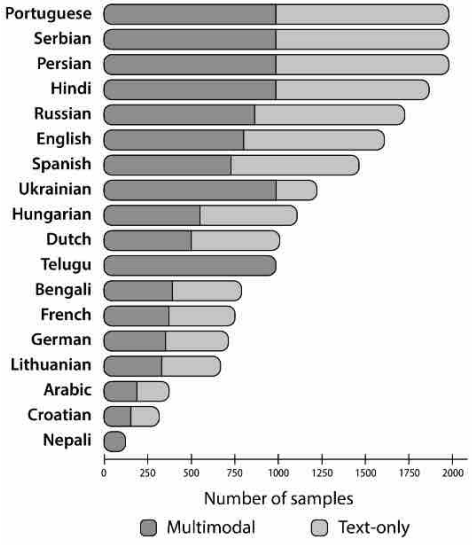
- What type of text are you collecting?
- Where are you getting the text from?
- Which languages are you covering and why?

LAION

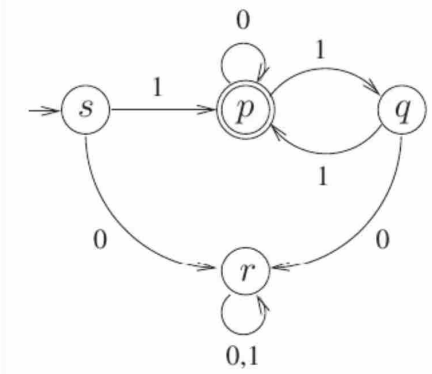


Kaleidoscope: Natural Questions

- The texts were created by exam-board professionals for educational purposes



Consider the Deterministic Finite-state Automaton (DFA) A shown below. The DFA runs on the alphabet {0, 1}, and has the set of states {s, p, q, r}, with s being the start state and p being the only final state. Which one of the following regular expressions correctly describes the language accepted by A?



FoodieQA: Procedural Single-Image QA

- The fine-grained taxonomy and careful labelled meant that we could automatically create questions using templates

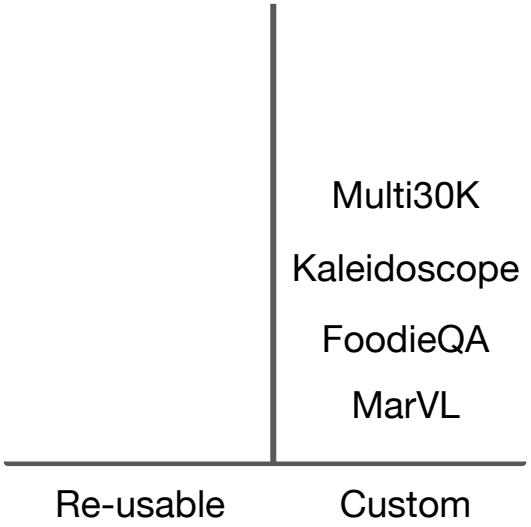
<dish>是哪个地区的特色菜? (What region is <dish> a specialty dish of?)
<dish>是哪个地区的特色美食? (In which region that <dish> is a local specialty?)
去哪个地方游玩时应该品尝当地的特色美食<dish>? Which place should you visit to taste the local specialty food <dish>?

以下菜品是哪个地区的特色菜？
Which region is this food a specialty?

A 江苏 (Jiangsu)
B 京津 (Beijing & Tianjin)
C 香港 (Hong Kong)
D 广西 (Guangxi)

5. Tooling

- Which platforms are you using to create your resource?
- Are you relying on re-usable tools or is your pipeline custom-built?



Acknowledgements


E. Bugliarello


R. Ramos


A. Kádár


L. Specia


L. Barrault


W. Li


N. Collier


G. Chrupała


C. Qui


D. Oneață


S. Frank


R. Sanabria

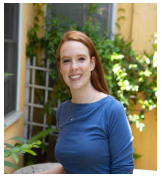

M. Fadaee


F. Liu


B. Martins

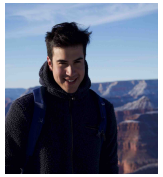

E. Hasler


E. Ponti


S. Hooker


S. Reddy


A. Alishahi


I. Salazar

Conclusions

- Two basic types of data in Culture x Multimodality
  - Data that is **mined** from something that already exists
  - Data that is **created** from scratch given a few ingredients and tools
- Lots of factors to keep-in-mind when we are creating a high-quality resource that will be re-used by hundreds–thousands of people
- Remember: not all data should be extracted (Bird, 2021)

References

- I. Salazar*, M. Fernández Burda*, S. Bin Islam*, A. Soltani Moakhar*, S. Singh*, F. Farestam*, A. Romanou*, et al. **Kaleidoscope: In-language Exams for Massively Multilingual Vision Evaluation**. arXiv:2504.07072.

- W. Li, C. Zhang, J. Li, Q. Peng, R. Tang, L. Zhou, W. Zhang, G. Hu, Y. Yuan, A. Søgaard, D. Hershcovich, and D. Elliott. **FoodieQA: A Multimodal Dataset for Fine-Grained Understanding of Chinese Food Culture**. EMNLP '24.

- F. Liu*, E. Bugliarello*, E. M. Ponti, S. Reddy, N. Collier, and D. Elliott. **Visually Grounded Reasoning across Languages and Cultures**. EMNLP '21.

- D. Elliott, S. Frank, K. Sima'an, and L. Specia. **Multi30K: Multilingual English-German Image Descriptions**. Workshop on Vision and Language at ACL 2016.