Paths

Personalised access to cultural heritage spaces

Aggregating Cultural Heritage Collections using Automatically Generated Topic Hierarchies

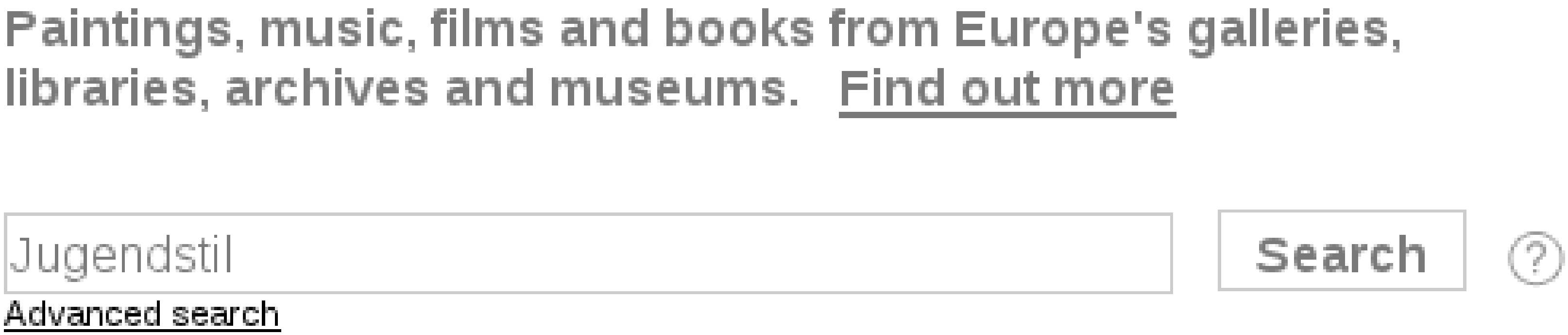
Mark M Hall
Information School / Computer Science
Sheffield University
Sheffield, UK

EuropeanaTech 2011, Vienna, 4th - 6th October 2011

European Commission
Information Society and Media

Accessing aggregated collections

- Searching works



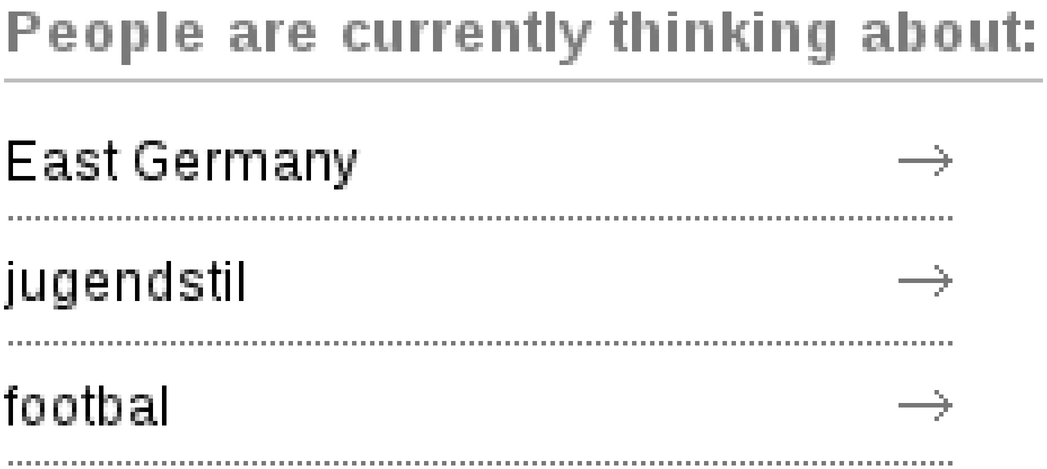
Paintings, music, films and books from Europe's galleries, libraries, archives and museums. Find out more

Jugendstil

- If you know what you are looking for

Accessing aggregated collections

- Recommendation works



People are currently thinking about:

East Germany →

jugendstil →

footbal →

- If you want to have an area of investigation suggested

Accessing aggregated collections

What if you want neither of those options?

What if you just want to browse around?

What if you want to know what is available in the collection?

Providing collection overviews

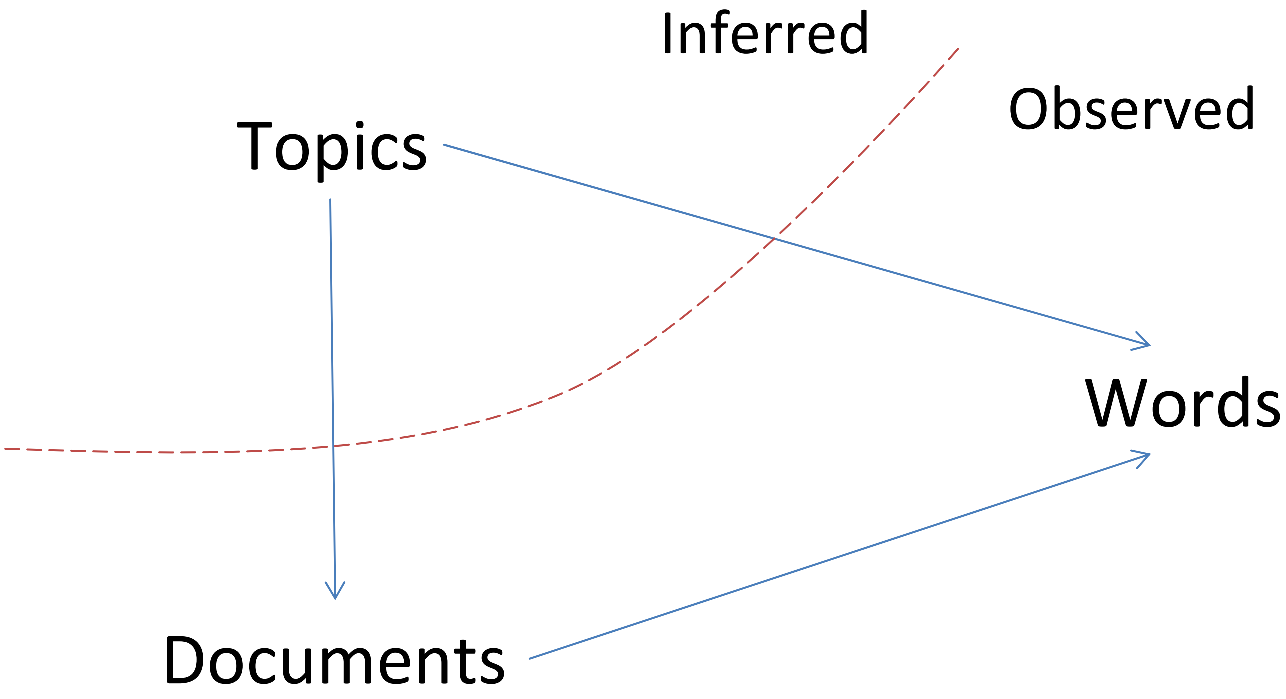
- Use an existing thesaurus
  - – Only parts of the aggregated collection will refer to it
- Use multiple thesauri
  - – No unified overview
- Create a manual overview
  - – Time/Resource consuming
- Use an automated approach
  - – Not perfect, but good enough

Automatic overview approaches

- Using statistical topic models
  - – Creates an entirely new, custom hierarchy
  - – Latent Dirichlet Allocation
    - Flat
    - Hierarchical

- Using a unifying thesaurus
  - – Uses a known thesaurus
  - – Links items to the thesaurus

Latent Dirichlet Allocation (LDA)



Inferred

Observed

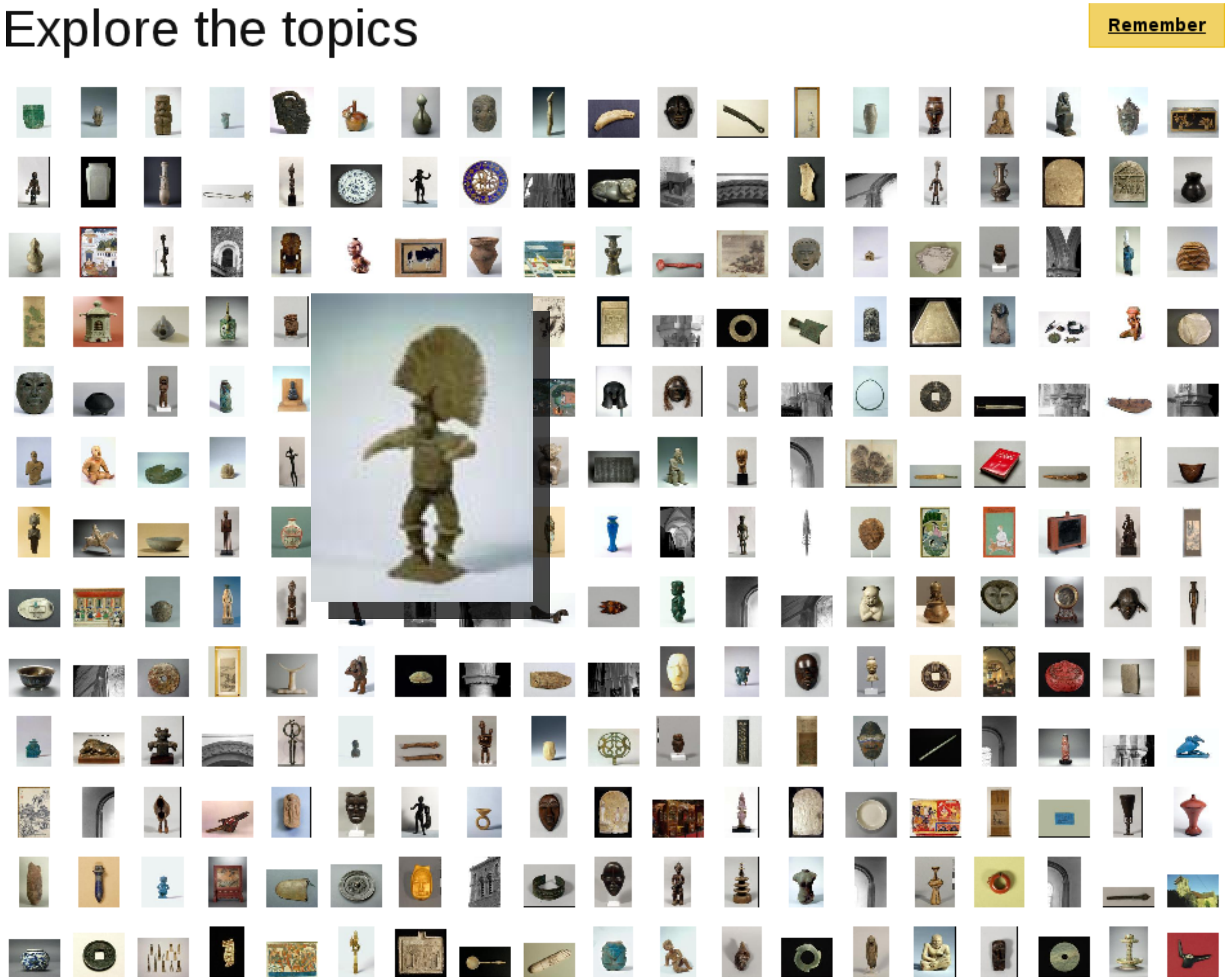
Topics

Words

Documents

Flat LDA

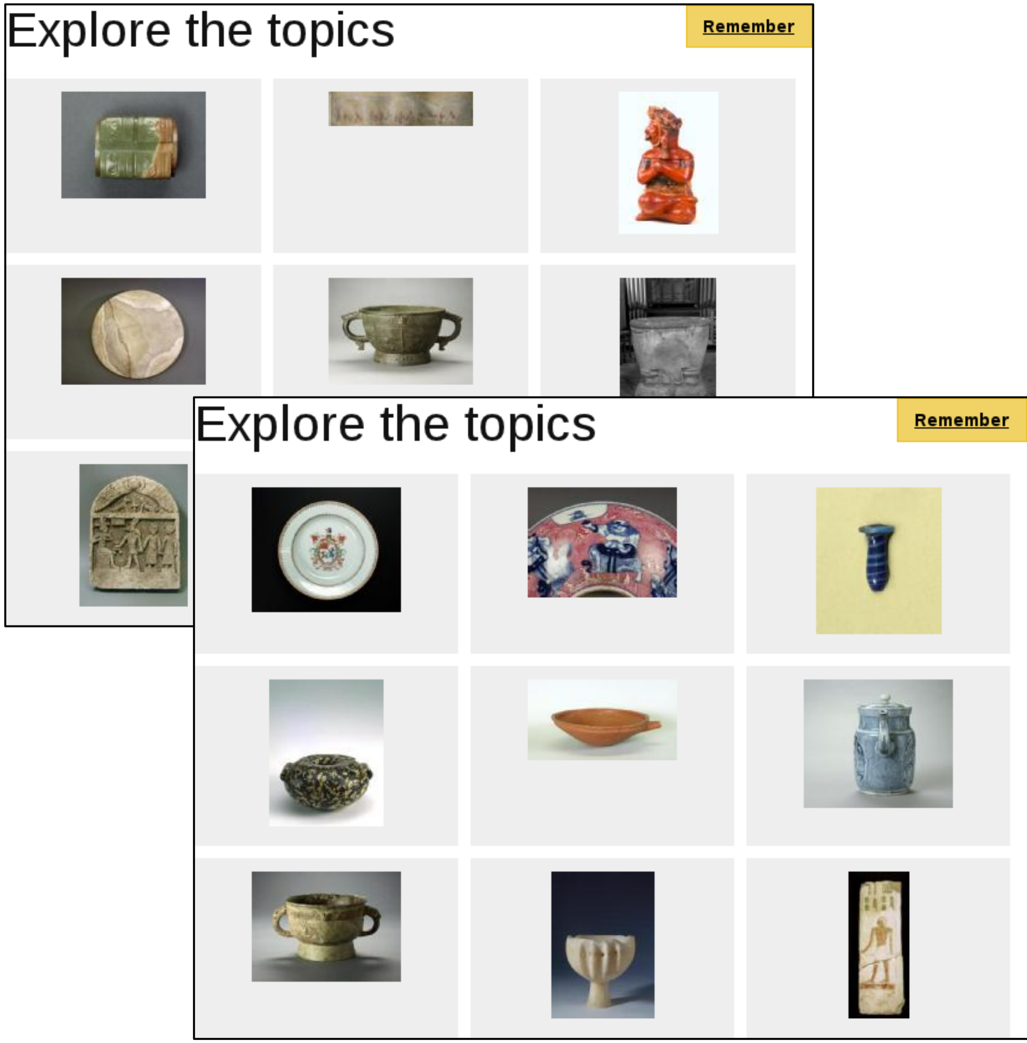
1. Generate topics
2. Select image for each topic
3. Place in grid



Explore the topics

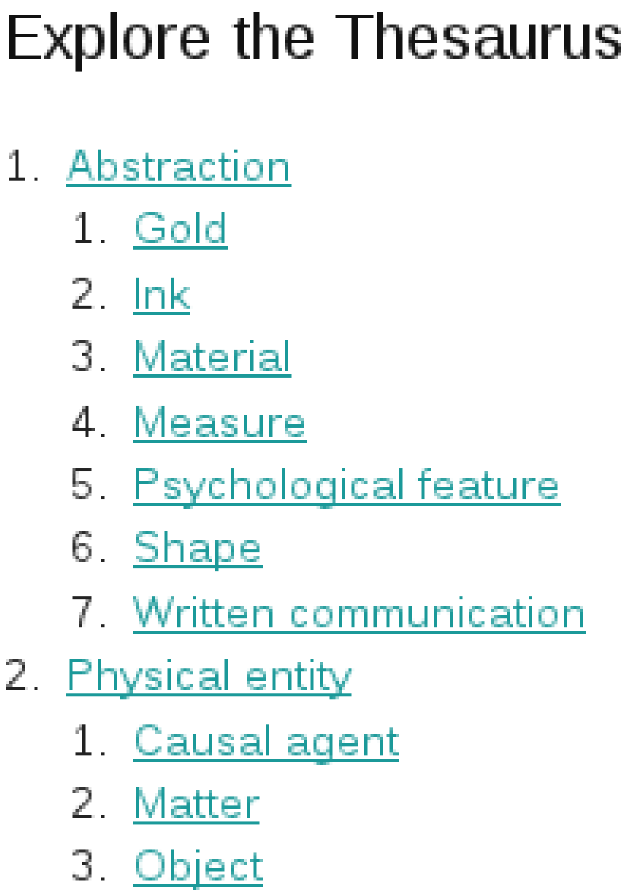
LDA Hierarchies

- Generate a small set of topics
- Assign each item to one topic
- For each of the topics generate a new set of sub-topics using the items assigned to that topic
- Repeat
- Display as a navigational grid



Unifying Thesaurus

- Pick a thesaurus that covers all / most subjects in the aggregated collection
- Use Flat LDA topics to identify the most important keywords in the collection
- Map these keywords into the thesaurus
- Use the mapped keywords to link all items to thesaurus entries



Explore the Thesaurus

1. Abstraction
    1. Gold
    2. Ink
    3. Material
    4. Measure
    5. Psychological feature
    6. Shape
    7. Written communication
2. Physical entity
    1. Causal agent
    2. Matter
    3. Object

Conclusion

- It works
- It's flexible
- It can deal with large collection sizes
- It is language neutral (except for the thesaurus work)
- It is not as good as a manual classification
  - But it is a lot faster and cheaper

Contact

Thank you for your attention.

m.mhall@sheffield.ac.uk

See how it could work in our Hackathon contribution!

I have leaflets if you are interested in our work.