A classified catalog and its applications in the TLS

Christian Wittern

Kyoto University Institute for Research in Humanities Center for Informatics in East-Asian Studies

2023-05-12

Classified Catalog in the HXWD

Classified catalogs: the problem

- Many slightly different versions have been used in Chinese bibliographical treatises and catalogs of libraries.
- The catalog in the TLS is based on the catalog of the Kanseki repository
- However, texts not in the Kanseki repository are not included / not treated properly
- Different ways to classify texts should be possible:
  - by content
  - by creation period
  - by region

Kanseki Repository catalog

- KR1 經部 Jing: Confucian Classics (incl. music, dictionaries and elementary learning)
- KR2 史部 Shi: Historiography and politics
- KR3 子部 Zi: Masters, philosophers and treatises
- KR4 集部 Ji: Anthologies (Poetry and Collected Writings)
- KR5 道部 Dao: Daoist texts
- KR6 佛部 Fo: Buddhist texts
- Below these top categories is exactly one level of subcategories

Changes to consider for the TLS

- Currently, the catalog is completely tesselating, e.g.
  - no overlap among categories
  - every text is member of exactly one category
- Allow texts to be member of multiple categories?
- Add additional subcategories for large subcategories

Text list

- New text list in the TLS

Facets

- One obvious application is to use this for analysing the texts
- For search results, we can see, how many results fall in each category
- Since the categories are in a hierarchical taxonomy, higher level categories encompass the lower ones
- The analytical cut can be made at different levels, depending on the purpose

Expected and unexpected results

- The number of characters for each text has been registered
- These numbers are also aggregated to the higher levels
- On every level, this can be used to form expectations on how many result should fall into each level, assuming an even distribution
- The distribution is rarely even -> unexpected results
- The ratio of given results / expected results is calculated on each level

Example of ratio calculus

- A category A has two subcategories B and C, both have a text with 50 characters
  - the ratio of expected results is 0.5 (50 / 100) for both B and C
- A search results in 10 hits for category A, with 3 in B and 7 in C
  - B : 3 / 10 = 0.3
  - C : 7 / 10 = 0.7
- Ratio

Example of ratio calculus

- A category A has two subcategories B and C, both have a text with 50 characters
  - the ratio of expected results is 0.5 (50 / 100) for both B and C
- A search results in 10 hits for category A, with 3 in B and 7 in C
  - B : 3 / 10 = 0.3
  - C : 7 / 10 = 0.7
- Ratio
  - B : 0.3 / 0.5 = 0.6
  - C : 0.7 / 0.5 = 1.4

Classified Catalog in the HXWD

Hierarchical taxonomy for dates

- Currently, texts have a date set for them, usually a smaller or larger span,
  - e.g. 17 BC or 200 BC to 300 BC
- This allows to approximately sort the results in chronological order
- Given a hierarchical taxonomy, e.g
  - Eastern Han / Han / Old Chinese / Ancient Chinese / pre-modern Chinese
- we can still associate a text with a precise date, but also harness the higher levels were appropriate
- for texts that are not precisely datable, we do not have to lie

Problems and questions

- How to we construct the taxonomy
  - do we want/need to allow users to use their own?
- Anthologies contain text from different periods
  - Can we (do we need to) date with more granularity
    - e.g. sections, paragraphs, voices in a text?
- For regional differences see the next section

Classified Catalog in the HXWD

Regional aspects

- The region of origin for a text is significant for many questions
- How can we incorporate this into the TLS
- Hierarchical taxonomy of regions
- Based on traditional cataloging
- Needs expansion and reworking

Traditional cataloging

- A typical byline in a catalog would look like this
  - 陶淵明全集　晉　陶潛　撰
  - Tao Qian of the Jin dynasty authored "The complete works of Tao Yuanming"
  - Jin is the era and region of Tao Qian's origin
- Or another example
  - 入唐新求聖教目錄　日本　圓仁　編 A catalog of newly obtained scriptures from the Tang, compiled by Ennin of Japan
  - In this case the regional aspect is even more obvious