Hierarchically structure Cultural Heritage objects

Shenghui Wang, Rob Koopman
OCLC Research, Leiden, Netherlands

1 What is the problem?

More and more Cultural Heritage (CH) content is being digitised and made available through digital libraries and aggregators such as Europeana.eu and the new Digital Public Library of America (dp.la). These aggregators provide access to large numbers of heterogeneous Cultural Heritage objects (CHOs), however, also bring challenges to users to explore such large aggregations.

Aggregating metadata from heterogeneous collections raises quality issues such as uneven granularity of the descriptions, ambiguity between original and derivative versions of the same object, even duplication if different providers give access to a same object. Also, simple, common-denominator vocabularies such as Europeana Semantic Elements (ESE) are inappropriate for capturing internal semantic links between objects (e.g., parts of an object, adaptations of a work, objects representing others) or external links to contextual entities (*e.g.*, places or persons related to an object). Many data providers do not have resources to provide richer and interoperable metadata as instructed in the CIDOC-CRM[1] and the new Europeana Data Model.[2]

Traditional usage of a search box and a query-response mode of interaction are no more sufficient when users do not have clearly defined information needs, or when they want to gain an overview over collections. More and more browsing and exploration functionalities based on thesaurus, facets, or clustering have been proposed to improve user search experiences. However, overviewing and exploratory browsing have still not been investigated much [1].

2 Hierarchically structuring CHOs

We propose a bottom-up approach: finding related Europeana objects at different levels of similarity, which potentially reflect different semantic relations between them. As shown in Fig. 1, after calculating the clusters at level 80 (with the similarity of 80%), we generate an artificial record from each cluster, gathering in each metadata field its values for all clustered records. These artificial records, together with all the records which could not be clustered at

---

[1] http://www.cidoc-crm.org
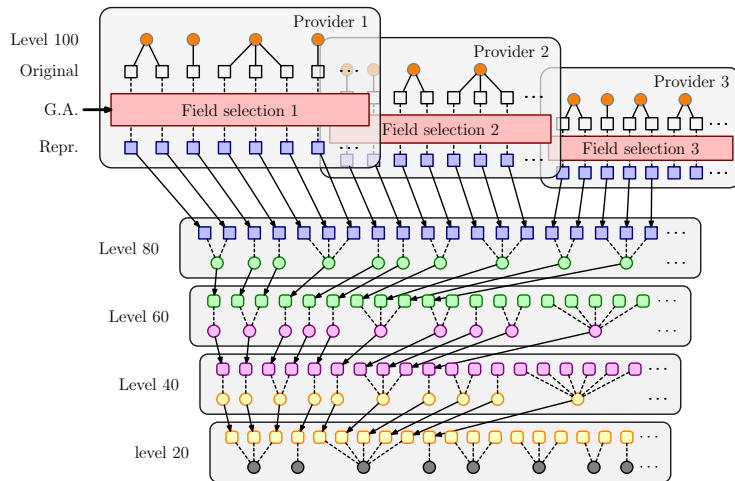[2] http://pro.europeana.eu/edm-documentation

Figure 1: Hierarchical structuring of CHOs at different similarity levels.

level 80, will join the clustering process at level 60. We again cluster at level 40 and 20 in the same way.[3] In the end, hierarchies of records are generated, so that one can have some structural information about these records, instead of quickly getting drowned in the sheer amount of data.

3 Future work

Once the hierarchy is constructed, we can apply the map-based visualisation [1] to build a map for Europeana. Users can zoom out to see the overview of the whole collections and zoom in to see more closely related objects. For each cluster at each level, we need to find a suitable label to best describe the content of the clustered objects. One possible way is to map the artificial objects to certain external semantic structure, such as Wikipedia, and use the matched concepts to represent the clusters.

References

[1] M. Hall and P. Clough. Exploring large digital library collections using a map-based visualisation. In *Proceedings of 17th International Conference on Theory and Practice of Digital Libraries*, pages 216–227, Valletta, 2013.

[2] S. Wang, A. Isaac, V. Charles, R. Koopman, A. Agoropoulou, and T. van der Werf. Hierarchical structuring of cultural heritage objects within large aggregations. In *Proceedings of 17th International Conference on Theory and Practice of Digital Libraries*, pages 247–259, Valletta, 2013.

---

[3]See more technical details in [2]