Extracting Dewey Decimal Classifications from Dublin Core Metadata Records With the DISTIL Project: Preliminary Findings and Observations.

Michael Khoo
The iSchool, Drexel University,
3141 Chestnut St, Philadelphia, PA 19104, USA
khoo@drexel.edu

Douglas Tudhope, Ceri Binding
University of Glamorgan,
Pontypridd, CF37 1DL, Wales, UK
{dstudhop, cbinding}@glam.ac.uk

Despite the visions of large-scale information infrastructures such as cyberinfrastructure and E-science, achieving efficient federated search and discovery across heterogeneous digital repositories remains an elusive goal. The DISTIL (Document Indexing & Semantic Tagging Interface for Libraries) project is pursuing one strategy that seeks to this address this goal in the humanities and the social sciences. The approach involves (1) taking individual Dublin Core metadata records from four digital libraries (the Internet Public Library (ipl.org), the National Science Digital Library (nsdl.org), the Digital Library for Earth Systems Education (dlese.org), and Intute (www.intute.ac.uk); (2) performing text analyses on selected fields from each record (currently *title*, *description*, *subject* and *keywords*) in order to generate terms that summarize the content of each record; and (3) submitting these terms to PERTAINS (PERsonlisation Tagging interface INformation in Services), a tool that then generates DDC tags from keyword input [1, 4]. These DDC tags will then be used to support interoperable browsing across repositories.

A significant underlying assumption of DISTIL is that a metadata record, as a human-crafted description of a digital resource, contains, and can be analyzed for, an inherent sense of 'aboutness' with regard to that digital resource (in other words, that catalogers have the skill to generate accurate descriptions of resources). Extracting this aboutness is however a complex task. Initial work with DISTIL has compared two different approaches to the text analysis of metadata records (step 2 above) [2]. In the first analysis, a basic count is being used to identify frequently-occurring nouns and verbs. In the second analysis, an online text-parsing tool, the U.K. National Centre for Text Mining's TerMine [3], is being used to identify compound noun phrases. Both processes are being applied to a small sample of records randomly selected from NSDL, DLESE, and Intute, and each process is producing different results. The frequency analysis produces a clear hierarchy in which a few terms occur frequently, and many terms occur infrequently. However, the top terms are general and not necessarily useful in producing precise DDC tags. The TerMine noun phrase analysis produces many precise results, but in a flat hierarchy, with many noun phrases occurring once only, making it difficult to decide which result is the most significant.

In the workshop we will report on progress with these text analyses. Some sort of combination of the approaches would appear to be useful, and the next goal is to refine these initial steps to see how the different approaches might be combined and reconciled, as well as to experiment with further approaches to analyzing metadata records. We will also introduce issues that arise when the input terms and phrases fed into DISTIL result in multiple hits across a large DDC entry vocabulary. (A physical analogy might be where to place, for example, an anthropological monograph in a library – by discipline, geographical area, social theoretical approach, etc.) There are various potential options available here for making sense of multiple hits, such as clustering and ranking based on analyses of the distributed nature of hits, if 'topics' within the DDC hierarchies are sorted not by low-level topics, but (for example) by discipline.

References

[1] Binding, C., and Tudhope, D. Terminology Web Services. Knowledge Organization 37(4), 287-298 (2010)
[2] Khoo, M., Tudhope, D., Binding, C., Abels, E., Lin, X., & Massam, D. (2012). 'Towards Digital Repository Interoperability: The Document Indexing and Semantic Tagging Interface for Libraries (DISTIL). Theory and Practice of Digital Libraries (TPDL) 2012, Paphos, Cyprus, September 23-27, 2012.
[3] TerMine: http://www.nactem.ac.uk/software/termine/
[4] Tudhope, Douglas and Binding, Ceri. 2008. Faceted Thesauri. Axiomathes, 18(2): 211–222.