Representing Aggregate Works in the Digital Library

George Buchanan[1], Jeremy Gow[2], Ann Blandford[2],
Jon Rimmer[3] and Claire Warwick[3]

[1] University of Wales, Swansea g.r.buchanan@swansea.ac.uk
[2] UCL Interaction Centre, London j.gow,a.blandford@ucl.ac.uk
[3] UCL SLAIS, London j.rimmer,c.warwick@ucl.ac.uk

Abstract. This paper studies the challenge of representing aggregate works such as encyclopaedia, collected poems and journals in digital libraries. Reflecting on materials used by humanities academics, it demonstrates the complex range of aggregate types and the problems of representing this heterogeneity in the digital library interface. We demonstrate that aggregates are complex and pervasive, challenge many common assumptions and confuse the boundaries between organisational levels within the library. The challenge is amplified by concrete examples.

Keywords: Digital Libraries, Architecture, Collection Building

1 Introduction

As more pre-digital humanities material is made available digitally, many collections now deal with aggregate works which associate a single identity with a set of atomic documents. But whilst these historic items are being digitised, historic forms of reference may be neglected. Locating an item within an aggregate requires searching and browsing to accurately reflect its structure.

One common and simple aggregate is the journal. If a collection is built of individual journal articles, then one document consistently represents one article, a journal issue is a set of articles, a volume a set of issues. It would appear logical that a similar approach should be effective for other aggregates. However, that is not the case. If a work is bound in two separate volumes, then it would make sense to separate between the two. However, that means that we now have two separate 'documents' in the library, which need to be linked for the purposes of browsing and searching. Counter-examples can also be found where multiple books are bound in one volume. An effective library will support retrieval under either criteria.

In addressing aggregate works, we presuppose the existence of an atomic 'document unit'. Aggregate works are defined as ordered trees with documents units at the leaves. This paper continues with an enumeration of aggregate features, followed with a review of problematic cases. We close with a discussion of related literature and the course for future research.

2 Aggregate Structures in Practice

Here we enumerate some significant features of aggregate works. Note that these features are not all mutually exclusive:

Homogenous Aggregation Each aggregated unit is of the same type.

Heterogenous Digital Forms Though an aggregate work may be logically homogenous, its digital form may vary internally. e.g. digitisation occured over a period in which practice shifted.

Serial Aggregation Aggregation from a series of related publications. e.g. journals or larger works that are published over many years.

Binding Aggregation A work was printed and released as one item, but bound in separate volumes.

Composite Aggregation When a work is published in parts, as with *serial aggregation*, but each part is itself bound within a different aggregate. e.g. 19th Century novels serialised in magazines.

Containing Aggregation A work may be small and unavailable in its own right, but available contained within larger works which are not themselves aggregates, e.g. a poem within a work of fiction.

Heterogenous Aggregation A work is created from units of diverse types. For instance, newspapers and journals contain articles of different types that may need to be distinguished in the DL interface.

Supplementary Aggregation Where an original work is supplemented by further material, possibly by another author.

Incomplete Aggregation Some aggregates are incomplete, either because they were not fully published or because a collection is only partial.

Variable Aggregation Different versions of an aggregate work may bring together different material, or different versions of the same material.

Furthermore, the boundary between external and internal document structure is not fixed, and many of the issues above may also occur *within* a document. What is important, from the view of a DL system, is that the treatment of internal and external aggregation are treated consistently in the DL architecture and also in the user interface, to ease the task of readers and librarians alike.

3 Difficult Cases

Our own experience on realising aggregates is based on the Greenstone DL system [8], and DSpace [6]. Simple cases such as journal collections result in few problems. However, beyond such regular structures, problems rapidly multiply. In a collection of literature the scale of items varies from a short stories to a multi-volume "epics". If we faithfully replicate the physical text, some items will be multi-volume, whilst a single volume may contain several works. The concept of 'volume' thus becomes problematic.

Indexing a collection by volume conflates works that share the same volume, whilst indexing by works only will conflate volumes of the same work. Clearly, neither solution is optimal: the natural conclusion is to index by the smallest unit (work, volume) and aggregate upwards to unify elements of the same item. This underlying storage can be represented in different ways in the library interface: e.g., matches against a single search for separate volumes of the same work can be unified in the search result list. This option is already available in Greenstone [8], and can be achieved in DSpace with careful configuration and effort.

During browsing, however, the contradictory use of volume (as a part or as an aggregate) will still emerge in some form or other. One can distinguish the part-of and aggregate-of styles of volume by introducing a three-level hierarchy and using disciminating labels for the top and bottom levels. Many items are represented by only one item at each level, and as reported in [7] such simple single-child relationships should be pruned so that unnecessary interaction is minimised. Thus to improve the interactional efficiency, the experienced hierarchy becomes irregular. The issues of unifying hierarchy nodes in search result lists remains a problem (though this can be achieved in Greenstone).

We now focus on complex cases with increasing degrees of difficulty, particularly *containing aggregation* and *composite aggregation*. Composite aggregates represent particularly problematic structures. Serialised fiction such as Conan Doyles *Gang of Four* disrupts DL assumptions in its original form. If each newspaper in a collection is stored as a single document, then a reader will need to map the Gang of Four in its original context to particular editions of the correct publication – which they may not know!

An alternative approach would be to extract and record the elements of the story as one DL document, tidily avoiding the problem for a searcher specifically looking for the Gang of Four, but conversely divorcing it from its original context - to connect each article with its context in the original magazine, the user must in fact engage in the 'hunting' of articles we apparently just avoided. Such contextual interpretation is the knub of many items of humanities research. Clearly, an optimal approach allows both the recovery of the original composited piece, and the magazines of which it was part.

4 Related Work

The difficulties of the representation of aggregate works in digital libraries has already received attention: e.g. Hickey and O'Neill [1] note problems encountered in applying FRBR [3]. O'Neill proposes treating aggregates as published volumes of more than one work, and to avoid recording aggregates as works in their own right. This introduces an inconsistency with the accepted FRBR model where every published volume (*manifestation*) is an instance of a single work.

Two electronic document standards support aggregate works: TEI [4] and METS [2]. In both cases, aggregates are achieved by pointers to their parts, to create a whole. TEI primarily uses pointers between parts of the aggregate, whereas in METS a central document contains references to part or whole other METS documents. Aggregates have been poorly represented in DL systems: e.g. DSpace [6] and Greenstone [8] focus on treating collections as sets of objects, with a hierarchical classification structure. Aggregates can be represented using the classification structure, but at the loss of consistent treatment of aggregates across both searching and browsing. In library science, the need to find and recover texts via bound volumes has emphasised the same approaches we see in DL systems. Aggregates are generally indexed by part where the parts are discrete works: e.g. the British Library binds brief tracts together in volumes, but each tract in a volume is indexed separately. Conversely, multi-volume works are usually, but not universally, indexed by one entry.

Svenonius [5], p. 103, notes that there are two potential routes to relating aggregates with their constituent parts: first, formal linkage structures; second, providing descriptive aggregation (meta–)data for each item. The latter approach, though informal and easy to apply, leaves much of the retrieval work with the user, and greater room for mismatches between the descriptive data and the corresponding description of the part or aggregate in the catalogue index.

5 Conclusion

We described above a number of different forms of aggregate work found in the humanities. Simple forms may be supported in DLs with only small shortcomings in representation. However, more complex forms of aggregation which occur frequently in historic literature map less readily to existing DL architectures and interfaces. In our research, we wish to investigate further the appropriate interactions to support the occurrence of aggregates in search result lists, and the location of desired aggregates in the course of information seeking.

Acknowledgements This research is supported by EPSRC Grant GR/S84798.

References

1. T. B. Hickey and E. T. O'Neill. Frbrizing oclc's worldcat. *Cataloging and Classification Quarterly*, 39:239–251, 2005.
2. Library of Congress. *Metadata Encoding and Transmission Standard (METS).*
3. S. G. on the Functional Requirements for Bibliographic Records. *Functional requirements for bibliographic records.* K.G. Saur, 1998.
4. C. Sperberg-McQueen and L. Burnard, editors. *Guidelines for Electronic Text Encoding and Interchange.* TEI P3 Text Encoding Initiative, Oxford, 1999.
5. E. Svenonius. *The Intellectual Foundation of Information Organization.* Digital Libraries and Electronic Publishing. MIT Press, 2000.
6. R. Tansley, M. Smith, and J. H. Walker. The dspace open source digital asset management system: Challenges and opportunities. In *Procs. European Conference on Digital Libraries*, pages 242–253. Springer, 2005.
7. Y. L. Theng, E. Duncker, N. Mohd-Nasir, G. Buchanan, and H. Thimbleby. Design guidelines and user-centred digital libraries. In *Proc. 3rd European Conf. for Digital Libraries, ECDL*, pages 125–134. Springer-Verlag, 1999.
8. I. H. Witten, S. J. Boddie, D. Bainbridge, and R. J. McNab. Greenstone: a comprehensive open-source digital library software system. In *Proc. ACM conference on Digital libraries*, pages 113–121. ACM Press, 2000.