Representing Aggregate Works in the Digital Library

George Buchanan¹, Jeremy Gow², Ann Blandford², Jon Rimmer³, and Claire Warwick³

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 occurred 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,