Subject classification of collection-level descriptions using DDC for information landscaping

Gordon Dunsire[1*], George Macgregor[2] & Ritchie Thomson[3]

[1] Centre for Digital Library Research, Department of Computer & Information Sciences, University of Strathclyde, Livingstone Tower, 26 Richmond Street, Glasgow G1 1XH. UK.

[2] Information Strategy Group, Liverpool Business School, Liverpool John Moores University, John Foster Building, 98 Mount Pleasant Street, Liverpool L3 5UZ. UK.

[3] Head of Cataloguing, Queen Margaret University Library, Clerwood Terrace, Corstorphine, Edinburgh EH12 8TS. UK.

Unpublished working paper
Date: 01/12/2006

Abstract: Collection-level description (CLD) has emerged as an important tool for facilitating user access to large heterogeneous collections within digital library and hybrid information environments. Such metadata enables "information landscaping" techniques to be deployed, thereby allowing users to survey, discover and identify relevant collections. This can aid the precision of item-level queries by eliminating collections which may produce a significant number of false-drops or may contain no relevant items. The ability to provide suitable subject indexing and subject-based organization within such collection-level environments is an increasingly important user requirement, particularly for landscaping; yet it remains highly problematic owing to, for example, the broad subject coverage of many collections and the item-level nature of controlled vocabularies. In this paper we propose a methodology for the subject designation of collections using the Dewey Decimal Classification (DDC). The proposed approach allows the establishment of reliable, consistent and meaningful DDC class numbers to facilitate improved user browsing and searching tools within CLD systems. The methodology will be demonstrated using the Scottish Collections Network (SCONE) and alternative techniques to facilitate general subject analysis will also discussed.

Keywords: Dewey Decimal Classification (DDC), collection-level description, subject analysis, information retrieval, information landscaping, information environments

1. Introduction

The use of collection-level description (CLD) has emerged as a valuable means of facilitating user access to heterogeneous collections within large digital and hybrid library environments (Chapman, 2005). CLD can support "information landscaping", a method for users to survey, discover and identify collections of items having the potential to satisfy their information needs (Chapman, 2004). This helps the user to improve the precision of item-level queries by avoiding collections which are likely to produce a high proportion of false-drops and to save time by not searching collections which are likely to contain no relevant items. Digital library expansion and the growth of cross-domain virtual collections are such that users increasingly require CLD to successfully navigate growing numbers of distributed and heterogeneous collections (Macgregor, 2003).

Although there may be several parameters for landscaping collections (such as access conditions or item format), the ability to provide suitable subject indexing and subject-based organization within collection-level environments is an increasingly important requirement. Belkin (1982) has noted that when a user feels compelled to use any information retrieval system it is because they are experiencing a gap in their knowledge. This gap generally can not be filled by retrieval strategies for known items. Instead, the user needs to find resources whose details are unknown at the beginning of their search, employing strategies that often involve searching or browsing for relevant subject information irrespective of who may have authored the information or published it (Garshol, 2004). This suggests that the ability of users to landscape collections by subject is of increasing importance. However, the assignment of accurate subject heading or classification notation at the collection level can be problematic. Collections can span multiple subjects across several disciplines, making it difficult or impossible to apply conventional procedures associated with item-level subject content analysis. The lack of any recognized research or guidance on subject analysis in CLD therefore limits the opportunities for offering reliable subject-based retrieval or landscaping techniques. Given the relative importance attached to providing such tools, the need to develop suitable methodologies for the consistent and meaningful subject analysis of collections is essential.

In this paper we propose a methodology for the subject designation of collections using the Dewey Decimal Classification (DDC). The proposed approach allows the establishment of reliable, consistent and meaningful DDC class numbers to facilitate improved user browsing and searching tools within CLD systems. The methodology will be demonstrated using the Scottish Collections Network (SCONE) as a test bed (http://scone.strath.ac.uk/service/). Although the focus is on the use of DDC, the techniques employed will be useful for general collection-level subject analysis involving verbal subject heading or other classification schemes. Alternative techniques to facilitate general subject analysis will also be discussed.

Since the use of formalized CLD in the library domain remains a relatively new area of development, section 2 briefly introduces CLD and its role in landscaping techniques. Section 3 describes the use of DDC within collection-level environments and the rationale behind, and the problems traditionally inherent in, the subject analysis of collections. These sections also briefly review related literature. Section 4 provides an exposition of the proposed methodology and demonstrates practical examples for illustrative purposes. Alternative approaches pertaining to functional granularity, collection de-composition and the Conspectus approach are addressed in section 5. Conclusions and further work are included in section 6.

2. Collection-level description and landscaping

CLD is a structured, open, standardized and machine-readable form of metadata providing a high-level description of an aggregation of individual items in both digital and physical environments (Chapman, 2004). Such metadata contain information about the existence, characteristics, and availability of specific collections and associated item-level finding-aids. A user can utilize a CLD service to identify collections with a common characteristic such as subject, collector, or location, across multiple information domains such as libraries, archives or museums (Dunsire & Macgregor, 2003). Mere identification of relevant collections may be sufficient for some needs, for example planning research strategies or consortial access agreements. CLD services can also enable a user to identify super-, sub- and other related collections, and to access collection catalogues, indexes and other finding-aids (Fig. 1). This process is more commonly termed "landscaping" and has been defined as allowing users to (Dunsire, 2004a):

- identify collections sharing common characteristics;
- describe and relate collections in a coherent and consistent manner;
- and, traverse levels of granularity in collections

CLDs may also be used for collection management purposes and to discharge an institution's curatorial responsibilities (Heaney, 2000).

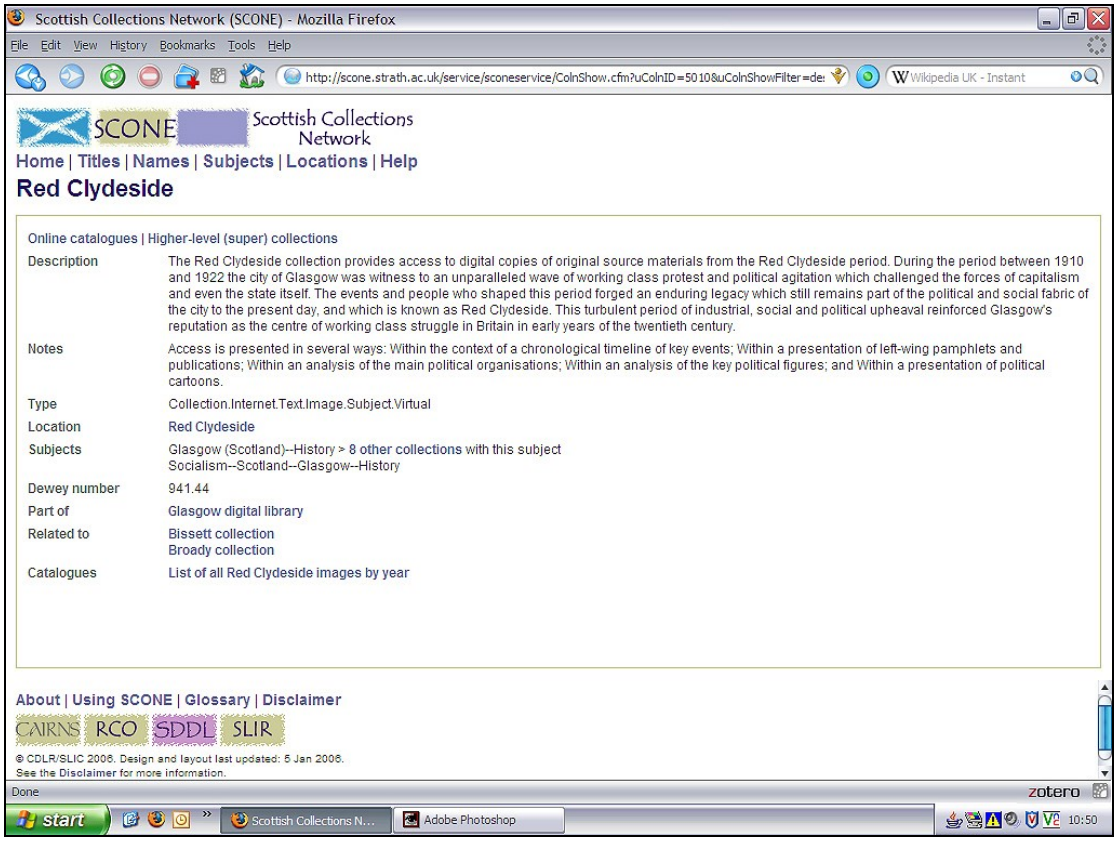

Fig.1. Example of a CLD, displayed via the SCONE user interface.

It is essential that the common properties of collections, in their various permutations, are described using a standardized schema to enable optimum landscaping and information retrieval functionality across distributed networked services. Several schemas have been proposed.

The UK Research Support Libraries Programme (RSLP) funded and developed the RSLP Collection Description (CD) schema (Powell, 2000; Powell, Heaney & Dempsey, 2000). This schema has subsequently become a de facto standard and enjoys wide international use within a variety of applications (Geisler, Giersch, McArthur & McClelland, 2002; Shreeves & Cole, 2003; AJLSM, 2004; Apps, 2004; Chapman, 2005; Foulonneau, Cole, Habing & Shreeves, 2005). The RSLP CD Schema has also informed subsequent approaches or schemas, such as those proposed by the Dublin Core Metadata Initiative (Dublin Core Collection Description Working Group, 2006) and the National Information Standards Organization (NISO, 2005). An alternative but compatible approach has been proposed by the RSLP-funded Scottish Collection Network (SCONE) (Dunsire, 2002). SCONE was developed in parallel with the RSLP CD schema and provides CLDs for over 5500 Scottish collections, both digital and physical. 1300 of these CLDs include DDC numbers.

Both the RSLP CD schema and the SCONE schema are based on an analytic model of collections and their catalogues (Heaney, 2000); the SCONE approach uses a fuller implementation of the model (Chapman, 2004). On the basis of extensive practical testing and research, SCONE has also proposed and implemented numerous extensions not specified in the original model (Dunsire, 2002). Heaney has subsequently extended his model to services mediating access to collections, taking into account extensions proposed by SCONE (Heaney, 2005). Even so, Heaney has acknowledged that the original model and subsequent extensions do not provide a comprehensive model of collections and that their purpose is merely to provide sufficient clarification regarding the essential attributes of collections and to expose

the process of resource discovery by users in such environments. It is also worth noting that the model deliberately neglects the analysis of subject in facilitating collection access and is therefore not modelled as a separate entity within the model. The need to consider subject access is required in order to better understand how Heaney's model can better be deployed. This need motivates much of the work documented in this paper.

Since there is no single established format for CLDs, SCONE has been developed to interoperate and output CLDs in a variety of formats (e.g. RSLP CD schema, Dublin Core Collection Description schema, UK Information Environment Services Registry (IESR), etc.) (Fig.2) (Dunsire, 2004b). SCONE is used as the test bed for the methodologies proposed in section 4.4 of this paper.

```
<?xml version="1.0"?>
<iesrd:iesrDescription
 xmlns:dc="http://purl.org/dc/elements/1.1/"
 xmlns:dcterms="http://purl.org/dc/terms/"
 xmlns:dcmitype="http://purl.org/dc/terms/dcmitype/"
 xmlns:iesr="http://iesr.ac.uk/terms/#"
 xmlns:iesrd="http://iesr.ac.uk/"
 xmlns:rslpcld="http://purl.org/rslp/terms#"
 xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
 xsi:schemaLocation="http://iesr.ac.uk/ http://iesr.ac.uk/schemas/xsd/iesr.xsd">
 <dcmitype:Collection>
  <dc:identifier xsi:type="dcterms:URI">
   http://scone.strath.ac.uk/coln/7952
  </dc:identifier>
  <dc:title>
   Dept. of Computing Science and Mathematics eTheses
  </dc:title>
  <dcterms:abstract xml:lang="en">
   Electronic copies of theses produced by students from the Department of Computing
Science and Mathematics of the University of Stirling.
  </dcterms:abstract>
  <dc:type xsi:type="dcterms:DCMIType">Collection</dc:type>
  <dc:type xsi:type="rslpcd:CLDT">
   Collection.Internet.Text.Image.Special.Form.Virtual
  </dc:type>
  <iesr:hasService xsi:type="dcterms:URI">
   http://scone.strath.ac.uk/coln/7953
  </iesr:hasService>
  <iesr:hasService xsi:type="dcterms:URI">
   http://dspace.stir.ac.uk/dspace/handle/1893/36
  </iesr:hasService>
  <dc:subject xsi:type="dcterms:LCSH">
   Computer science
  </dc:subject>
  <dc:subject xsi:type="dcterms:LCSH">
   Mathematics
  </dc:subject>
  <dc:subject xsi:type="dcterms:DDC">
   004
  </dc:subject>
  <dc:subject xsi:type="dcterms:DDC">
   510
  </dc:subject>
  <rslpcd:owner xsi:type="dcterms:URI">
   http://scone.strath.ac.uk/agnt/5393
  </rslpcd:owner>
  <dcterms:isPartOf xsi:type="URI">
   http://scone.strath.ac.uk/coln/7911
  </dcterms:isPartOf>
 </dcmitype:Collection>
</iesrd:iesrDescription>
```

Fig.2. SCONE CLD output in IESR format for collection entity.

3. Collection subject analysis and rationale of DDC within CLD environments

3.1 Collection subject analysis: problems

Lee (2000) has noted the increasing movement towards the "collection" as a vehicle for information delivery within digital environments. Hill, Janée, Dolin, Frew and Largaard (1999) also comment on the increasing heterogeneity of items comprising such collections. If users are to landscape and navigate these growing heterogeneous collections using current metadata solutions (i.e. CLDs), appropriate subject-based tools have to be made available. Not only do such tools have to be made available in conjunction with CLDs, they have to be accurate, consistent and reliable if CLD services are to provide meaningful access and maintain user confidence.

Assigning accurate and reliable subject headings or classifications can pose problems within collection-level environments. Many of these difficulties are encountered because theories on subject analysis are traditionally based on item-level activities. Lancaster (2003) notes that preparing a *representation* of the subject matter of information entities (e.g. subject indexing, indexing for classification, etc.) involves two related stages: conceptual analysis; and, translation. In essence, conceptual analysis involves determining the "aboutness" of an item by examining various attributes, such as the item title, sub-title, scanning or reading portions of the item contents, etc. Translation involves converting the identified concepts and representing them using the chosen classification scheme, subject heading list, thesaurus or other controlled terminology.

Such methodologies are often impractical or difficult to apply at the collection level. For example, it is unfeasible to expect cataloguers to scan, read or absorb portions of all the items contained within a collection to determine concepts for translation. Even if this were feasible, the subject representation scheme may not formally allow the translation of large numbers of concepts identified in a collection encompassing a wide span of subjects into a useful set of one or a few classifications or terms. There are also problems in using other elements of CLD to inform conceptual analyses. Collection titles and summary descriptions often say very little about the subject nature of the items they contain. The collection-level difficulties inherent in traditional subject analysis theory are also compounded by the tools themselves. Most controlled vocabularies are optimized for item-level resource discovery and guidance for their implementation generally reflects traditional indexing theory.

4. Literature review

Attention to the problems inherent in collection-level subject analysis (i.e. conceptual analysis and translation) has been limited in the library and information science domain. This is perhaps attributable to the relative infancy of CLDs as a tool for user resource discovery or landscaping, and the historical emphasis on item-level management within libraries. Archives, by contrast, have always engaged in a *type* of CLD by virtue of managing archival fonds (Chapman, 2004).

The idea of provenance traditionally underpins the description and organization of archival materials. Archival provenance dictates that, irrespective of their physical or digital manifestation or whether they were created by individuals or corporate bodies, the integrity of the archival fond should be maintained (Beattie, 1997). Respecting the integrity of the archival unit ordinarily precedes the assumption that archival materials can be dispersed into various subject or temporal taxonomic systems. To do so would be to remove the contextual background considered necessary if the items within the collection are to be correctly interpreted. The importance attributed to provenance is therefore reflected in archival description schemes such as ISAD(G) and Encoded Archival Description (EAD) which are

creator- or document-oriented and which do not traditionally accommodate the provision of subject-based access points. The problem of collection-level subject analysis and concept representation has therefore attracted most attention in the archival community where issues pertaining to subject analysis have gradually emerged as archival description has been forced to converge with library systems (Gabriel, 2002).

Despite this, progress on developing appropriate methodologies for collection-level subject analysis within archives has been inadequate. Most attempts have identified the need to facilitate subject-based access to archival collections, but few propose any particular methods or guidance. Beattie (1997) summarizes the desirability of subject-based access to archival collections but concedes that the archivist's ability to do so is currently compromised by several factors, including over-emphasis of provenance-based retrieval systems, the nature of archival collections themselves, lack of appropriate controlled vocabularies, and outdated assumptions about user requirements. Beattie concludes that archivists have to broaden their view of subject access and accept that provenance-based systems are now unviable. In lieu of any codified guidance on implementing subject-based access, Beattie proposes various enhancements to archival description to increase alternative access points and to save time for the user when engaging in content analysis. Gilmore (1988) also notes the need to provide subject access to archival collections via online library catalogues and highlights some potential methods, such as call-number searching using a combination of item and collection-level cataloguing. This approach is predicated on the ability to analyze collections proficiently and Gilmore acknowledges that archivists are frequently forced to assign broad subject headings, resulting in materials "disappearing into a void".

While investigating the merits of controlled and uncontrolled vocabularies for subject retrieval within archives, Ribeiro (1996) noted that information retrieval techniques within archival collections were relatively undeveloped and those standards that existed were generally not applied. Ribeiro consequently took concept analysis and translation guidance provided by the International Organization for Standardization (ISO) 5963:1985 (ISO, 1985) and proposed a series of revised elements that should be observed during the analysis and translation of archival collections. According to Ribeiro such analysis of archival collections should entail the examination of five attributes: series title (when it exists); provenance statement; indexes (i.e. original indexes or contents tables accompanying volumes or bundles); type of documents; and, elements within each document or record. Commenting on the latter, Ribeiro noted that small collections within restricted subject areas were conducive to concept analysis and translation. However, Ribeiro conceded that even when document types within an archival collection are largely homogeneous, the items will often encompass a diverse distribution of subjects thus inhibiting effective translation. Ribeiro notes that this evades concept analysis and makes it "impossible to establish any objective criterion for content analysis or to identify the concepts". Ribeiro concludes by stating that, in such scenarios, the elements within each document or record are omitted; that is, not indexed or considered in the indexing process.

These issues have implications for the accurate and useful representation of subjects in CLD to support landscaping and navigation, and there exists a real need to research and develop suitable methodologies for subject analysis in CLD environments.

4.1. Role of DDC in landscaping

The effectiveness of using the Dewey Decimal Classification (DDC) for knowledge organization and resource discovery within digital environments has been demonstrated widely (see for example, Saeed & Chaudhry, 2002; Vizine-Goetz, 2002; Chowdhury & Chowdhury, 2004; Nicholson, Dawson & Shiri, 2006; OCLC, 2006; Vizine-Goetz, 2006).

The use of DDC numbers in CLDs is of wider relevance, particularly within the UK. National services such as the Information Environment Services Registry (IESR) require DDC numbers to be assigned to collection descriptions to facilitate the use of a terminology server within the JISC Information Environment (Apps, 2006). The proposed terminology service utilizes a DDC spine for terminology mapping between disparate subject schemes (Nicholson & McCulloch, 2006). Although there are other tools which lend themselves to the subject description of collections (ARE THERE ANY BETTER, REALLY? WHAT DO YOU THINK, GORDON?), the use of DDC notation enables a higher degree of language independent machine manipulation and interoperability. A user-supplied subject term, typically intended for item-level searching, is mapped to a DDC number which can then be used to automatically landscape collections with co-extensive or broader subject coverage. Matching to a broader subject is carried out by successive truncation of the DDC number, equivalent to traversing the notational hierarchy to identify superordinate classes, until collections are found; for example, the item-level number 530.1433 (Quantum electrodynamics) eventually landscapes collections classified at 530 (Physics).

A methodology for using DDC to augment subject-based collection landscaping techniques is therefore worthy of investigation. Relevant functionality offered by DDC includes the browsability of its expressive numerical notation, hierarchical arrangement of captions and scope notes, and its relative index of subject and discipline relationships, along with keyword searching of captions, scope notes and the relative index. It is also a suitable scheme for implementing innovative browsing tools that require structured notation or semantic hierarchies, such as topic maps or so-called "metabrowsing" (Wiesman, Van den Herik & Hasman, 2004) and is often commended for its mnemonic qualities (Maltby, 1975).

5. Contextual collection subject analysis using DDC: procedures

6.1 Item-level rules vs. collection-level rules

Most well established controlled terminologies are optimized for item-level resource discovery and guidance for their use is influenced by traditional indexing theory. DDC is no exception to this. The method proposed for using DDC requires that some specific implementation rules be ignored or modified, and that a "contextual analytical approach" be adopted. By ignoring those item-level rules which are essentially redundant within collection-level environments, it is possible to reconcile the numerous concepts and subject strengths of collections and translate these into a single (or manageable few) DDC numbers. In this respect the method borrows archival techniques since emphasis is placed on the collection as a contextual unit (similar to an archival fond). It could be suggested that the scheme is being applied in an invalid way if basic rules of applying DDC are dropped or amended. However, given users' familiarity with established schemes such as DDC and the absence of guidance on application in increasingly common multi-granular digital information environments, the need to modify and dispose of inapplicable rules appears necessary and justifiable. Before proposing the methodology, it is first necessary to study the rules of applying DDC in relation to collections.

4.2. Application of the DDC approach to classification

Although optimized for item-level subject analysis, DDC provides accommodation for the subject coverage of collections in several places in its semantic hierarchy. However, the treatment is inconsistent. Collections are variously treated on the basis of:

- the form of their constituent items, for example in class 709 (Photographs);

- their use as a research tool in specific disciplines, for example the standard subdivision –074 (Museums, collections, exhibits);
- their curatorial environment, for example class 708 (Galleries, museums, private collections of fine and decorative arts);
- DDC also accommodates other non-subject characteristics of collections, such as geographical location and age or educational level; however, these are given their own attributes in most CLD schemes and it is not necessary to resort to DDC as a source of values.

The most complete and coherent treatment of collection subjects is applied to library and archive collections using class 026. The caption for notation 026 is "Libraries, archives, information centers devoted to specific subjects and disciplines", with the scope note "Class here information organizations and library departments and collections in specific disciplines and subjects; comprehensive works on special libraries. For special libraries not devoted to specific disciplines and subjects, see the kind of library in 027.6, e.g., general museum libraries 027.68, general libraries in newspaper offices 027.69" (Dewey, 2005). The classification number for a subject-specific collection is built by adding the specific notation for the subject to the base notation 026. But this base notation is redundant when the scheme is applied exclusively to subject-specific collections, as every class number will begin with the same three digits. Dropping the base notation will have no impact on the utility of DDC for subject landscaping in CLD services; the result is the same as if the collection is treated as a single item, and the semantic hierarchy is preserved.

This suggests that the special accommodation for collections in DDC can safely be ignored and concepts can be directly translated to DDC numbers, provided the collection and item level retrieval functions of an information environment are kept separate. If it is necessary to preserve the integrity of DDC numbers in a system using cross-searching of collections and items in a single set of metadata, the "correct" DDC numbers for subject-based library and archive collections can be derived automatically by prefixing the base notation 026. Specific mappings from collection-level DDC to item-level DDC would be required for other types of subject-based collections.

The introduction to DDC states "Classifying a work with the DDC requires determining the subject, the disciplinary focus, and, if applicable, the approach or form" (Dewey, 2005). DDC notates "recurring physical form" as a standard subdivision which can be added to most DDC subject numbers, as in the -074 example above. However, the physical form of a collection as a whole is not a recognised attribute in most CLD schemes although the physical form of constituent items may be (Dunsire, 2002b), exemplified by the "Size" and "Item format" attributes proposed by DCMI (Dublin Core Collection Description Working Group, 2006) and the "Extent" attribute proposed by NISO (NISO, 2005). Use of DDC subdivisions for physical form is therefore not required and can be ignored.

A key principal underpinning the DDC is that a work be "classed in the discipline for which it is intended, rather than the discipline from which the work derives" (Dewey, 2005). This implies that the curatorial environment of a collection should not be a factor in assigning subjects to its CLD, so that archive, library, and museum collections about the same topic can be assigned the same classification.

The introduction to DDC explains that a "key element in determining the subject is the author's intent" (Dewey, 2005). At the collection level, the agent with the nearest equivalent role to author is the collector, the person or organisation formulating the scope and selecting the constituent items of the collection. The intent of the creators of the individual items in the collection, including authors, artists, and crafters, may be overridden by the intent of the collector. The subject of a collection is not necessarily derived from the subjects of its

constituent items; as Heaney (2000) notes, "the subject of a Collection need not be the same as the subject of the Contents (e.g. the subject of a Collection of bindings is the binding of the items, not the subject of the Content of the items)". However, subject analysis clearly should take into account the subjects of the items; if all the items in a collection of book bindings happen to have a common subject, then it is useful to assign that subject, as well as 686.3 (Bookbinding), to the collection. But it is the intention of the collector which provides the primary indicator of the context of the collection, and therefore the subject-based significance of the collection.

The introduction to DDC notes that "The title is often a clue to the subject, but should never be the sole source of analysis" (Dewey, 2005). This was found to be true for collections recorded in SCONE. Some collection titles indicate the subject directly, such as "SCRI raspberry literature collection". Others are ambiguous. For example, a title derived from a personal name may indicate a biographical subject, as with "Robert Louis Stevenson collection" which is classified at 928.21 (Writers in literature … in English), or a subject of interest to that person, as with the "Gibson collection" which is classified at 510 (Mathematics).

The introduction to DDC also advises that "Bibliographical references and index entries are sources of subject information" (Dewey, 2005). This can be extended to include finding-aids such as catalogues, archival descriptions, and indexes derived from the textual and visual content of the items in the collection. Bibliographic references consist of works about a collection and works which use the collection as a primary source of information.

4.3 Multiple subjects

Analysis of collections in SCONE shows that many can be assigned multiple concepts which are not closely related. Standard item-level application of DDC is intended to result in the translation of multiple concepts to a single number, to ensure consistency not only in semantic relationships between items but also their physical collocation when the number is used as the basis of a shelfmark (Broughton, 2004). This latter aim is mostly redundant in a collection-level environment; physical collocation of significant numbers of collections is likely to be required only in very large organisations, where sequencing will be by the broadest subject, if at all. Thus many of the DDC rules and instructions governing multiple subjects are not useful when translating collection-level concepts to numbers.

Several rules depend on ascertaining which subjects receive fuller treatment than others, but such quantification can be difficult to ascertain without aggregating the treatments given in the constituent items of the collection. As already noted, this is likely to be impractical for all but the smallest collections. It is more useful to identify subjects receiving significant treatment within the collection, as determined by the needs of the CLD service.

The DDC "first-of-two rule" demands that two subjects within the same discipline and receiving equal treatment in the work should be translated to the number coming first in the schedule. Applying the rule in a landscaping service would significantly impair the functionality of identifying collections likely to contain a specific topic by successively truncating its DDC notation until a match is found. An exact match with the subject coming second in the schedule would never occur, and in many cases truncating the notation would raise the hierarchical level above the subject coming first. It is better to subsume this rule with the approach taken by the "rule of three", which stipulates that if a work is analysed as having three or more subjects that are all subdivisions of a broader subject within a single discipline, the work should be classed in the first higher class that includes all of them. For application to CLD a general "rule of two" is proposed: if a collection can be assigned two or more subjects

within a single discipline, class it in the first higher class with a notation common to the subjects.

It may not be useful to apply such a rule in all cases because the notational hierarchy of DDC does not always match the semantic hierarchy. For example, 941.1 (Scotland) is semantically broader than 941.46 (North Ayrshire, South Ayrshire, East Ayrshire), but the first higher common notation is 941 (British Isles). In the case of the "Local and Scottish collection" of Ardrossan Library it would be better to assign the two DDC numbers to reflect the local (Ayrshire) and Scotland-wide aspects of the collection. The proposed "rule of two" should therefore be used according to "classifier's judgment" and the needs of the CLD service.

DDC provides interdisciplinary numbers in many places in its schedules, but use of such notations has the same issues as the rules governing multiple subjects in a single discipline. It is not always possible, or useful, to extend the proposed "rule of two" to the discipline level. Instead, it is better to ignore the DDC rules and guidance on more than one discipline and assign at least one subject within each discipline represented in the collection (Fig.3).

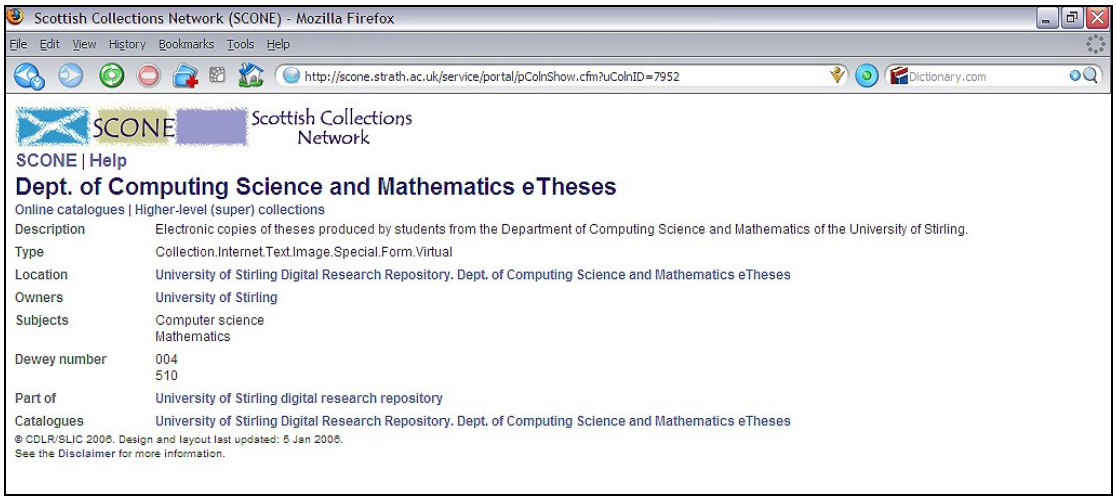

Fig.3. SCONE CLD for a sub-collection of Stirling University's institutional repository. Each of the two subjects is in a different discipline.

Some collections may have such a diffuse subject focus that the number of DDC notations capable of being assigned using this approach is too large to be supported by the resources allocated by the service to the classification process. For small collections, each subject may only be represented by an insignificant number of items within the collection; each item in a collection of ten might cover a different discipline, and there is little utility to be gained by assigning ten DDC notations to the collection. A CLD service should consider issuing specific guidance for such cases, determining at what tipping-point there are too many subjects or too few items to justify assigning classification notations. If there is a requirement for at least one DDC number to be assigned to a CLD, class 000 (Computer science, information, general works) class can be used, but this is discipline-specific and does not support the utility of traversing the notational hierarchy. It is better for the service simply not to assign any subject, and then display those CLDs without DDC numbers as the default landscape when a specific subject landscape fails to identify any relevant collections. That is, collections covering many or all subjects can be used as a landscape of "last resource". Although such a landscape would also include collections with very few items, these would, by definition, not make a significant contribution to false-drops or search duration.

4.4 Proposed methodology

The proposed methodology for applying DDC to CLD is, therefore, to follow the instructions and guidance given for item-level description with the following amendments:

- Ignore the accommodation of collections in DDC, and treat the whole collection as a single item or work.
- Ignore the use of physical form when adding standard subdivisions.
- Ignore the curatorial environment of the collection and determine the discipline for which the collection is intended, rather than the discipline from which it is derived, in the usual way.
- Treat the collector of the collection as the author of the work, and ask "What is the collector's intent?" when determining the subject.
- Treat a finding-aid of the collection as a source of subject information.
- Apply the *rule of application*, which states that a work pertaining to interrelated subjects should be classed with the subject that is being acted upon, in the usual way.
- Ignore the concepts of equal and fuller treatment when more than one subject can be assigned. Instead, apply a concept of significant treatment, relative to the needs of the CLD service.
- Replace the *first-of-two rule* and *rule of three* with a *rule of two* and assign multiple subjects in a single discipline to the next higher inclusive class. Alternatively, treat each subject separately and assign multiple DDC numbers to a single collection.
- Ignore the guidelines for subjects in more than one discipline. Instead, assign at least one DDC number for each discipline identified, irrespective of whether the rule of two has been applied within a discipline.
- Do not assign any subjects if the collection focus is too diffuse or there are too few items.

The decision tree diagram (Fig. 4) can be used in conjunction with the methodology to aid interpretation and implementation.

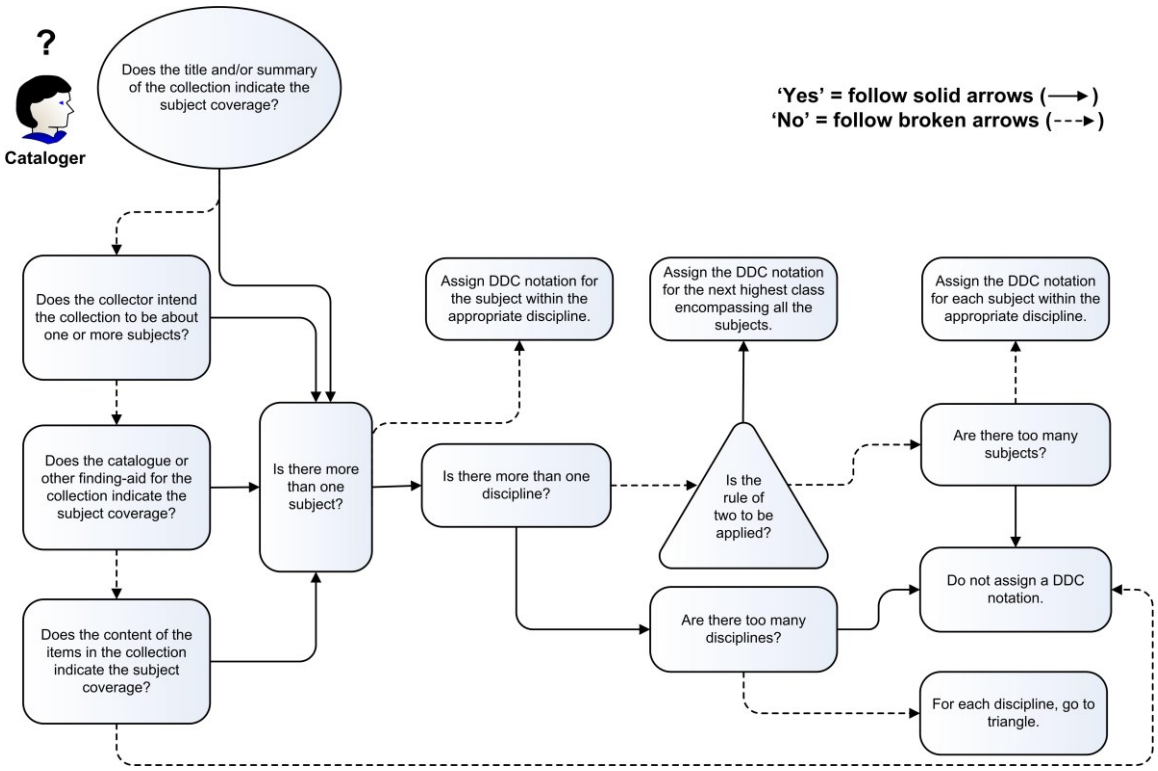

Fig.4. Decision tree diagram to assist in interpretation and application of procedure.

5. Alternative approaches

The proposed methodology facilitates the reconciliation of numerous disparate subject strengths and enables the application of DDC to CLDs. Since it may not be appropriate to assign multiple DDC numbers in all cases, it is possible to apply the concept of "functional granularity". Functional granularity allows a collection to be defined at various levels of aggregation deemed "useful or necessary for the purposes of resources discovery" (Heaney, 2000). A CLD service which wants to assign only one DDC notation per collection can create a set of sub-collections when a collection covers multiple subjects, each based on a specific subject focus. The rule of two can then be applied to each sub-collection. This approach allows the application of the DDC guidelines on using interdisciplinary notations: the interdisciplinary number is assigned to the parent collection, and numbers from the other relevant disciplines are assigned to sub-collections which are defined by specific subjects within those disciplines. For example, a collection of items on the subject of child development can be assigned the interdisciplinary number 305.231 (Child development), which is in the discipline of sociology. If the collection contains a significant set of items about child psychology, a CLD for a functional sub-collection can be created and assigned the number 155.4 (Child psychology). A similar approach can be used to decompose collections with very broad subject coverage into discipline-based sub-collections, with each sub-collection assigned the general DDC class for the discipline. This provides an alternative approach to simply not adding DDC notations to general collections, avoiding the use of a default landscape which might also include collections with too few items. Hierarchical links between sub-collections and their parents can be used by the service to provide seamless and transparent navigation for the user.

Although functional granularity has been applied extensively in SCONE, it has mainly been used to create super-collections described by aggregations of metadata in union catalogues (Dunsire, 2004a). Some proof-of-concept testing for subject retrieval has been carried out.

It is possible to provide subject-based resource discovery of collections with very broad subject coverage by analysing the relative strengths of subject representation as measured by the quantity and scope of items in the collection. Typically, a measure is given against each member of a high-level taxonomy or classification of all disciplines or subjects. One such methodology uses the DDC numbers assigned to the items in a collection (Nicholson, 2002). SCONE itself uses Conspectus, which uses the Library of Congress Classification (LCC) (Dunsire, 2006). A mapping from LCC to DDC would potentially provide a means of integrating resource discovery of both subject-specific and general collections within SCONE and other CLD services using Conspectus.

6. Conclusions

The multi-level granularity nature of traditional information environments has always been recognised: archives are organised by fonds; libraries maintain collections of serials which are collections of issues which are collections of articles; museums divide collections into rooms which contain display cases which contain objects. The digital information environment encourages higher-level granularity by facilitating the aggregation of information objects into collections and collections into super-collections, and lower-level granularity by facilitating the disaggregation of complex digital information objects into simpler components which can be treated as items in their own right.

There is an increasing need to develop tools and techniques to maintain effective resource discovery services with a focus on collections rather than individual works. The aggregation of digital metadata for physical and digital resources is proliferating; such aggregations create functional distributed super-collections. Searching and browsing by subject remains an

important discovery tool at all levels of granularity, and the DDC can successfully be applied at collection-level as well as item-level if the modified approach suggested in the proposed methodology is taken. However, further empirical user-based research is required to test and validate the effectiveness of the proposed method, and we intent to conduct tests to determine whether the classifications meet conventional relevance and precision criteria.

Higher education and research organisations are creating and developing institutional repositories offering metadata to aggregation services via the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). The structure of such repositories is likely to mimic the structure of the organisation, with a breakdown into "communities" equivalent to teaching and research departments. The application of collection-level subject classification may be important to local repositories and aggregation services because such departments are usually defined by subject focus.

The proposed methodology may also be applicable to the discovery of components of complex digital objects. Standards for describing such objects such as Metadata Encoding & Transmission Standard (METS) use hierarchies to group components at multiple levels of granularity in a similar way to the aggregation of items in collections.

Acknowledgements

This work was supported by the Scottish Library & Information Council and the Centre for Digital Library Research, University of Strathclyde.

References

AJLSM. (2004). *MICHAEL / Minerva data model*. Bordeaux, France: AJLSM. Retrieved 28 November, 2006, from http://projets.ajlsm.com/michael/dm-complete/data-model.pdf

Apps, A. (2004). A Registry of Collections and their Services: from Metadata to Implementation. In *DC-2004: Dublin Core and metadata applications (International conference): Vol. 1. 2004 Dublin Core annual conference: metadata across languages and cultures* (pp.67-73). Shanghai: Shanghai Scientific Technological Literature Publishing House. Retrieved 28 November, 2006, from http://epub.mimas.ac.uk/papers/appsdc2004.html

Apps, A. (2006). Disseminating service registry records. *In:* Martens, B., Dobreva, M. (eds): *ELPUB2006: Proceedings of the Tenth International Conference on Electronic Publishing - Digital Spectrum: Integrating Technology and Culture, Bansko, Bulgaria, 14-16 June 2006* (FOI-COMMERCE Sofia), 37-47. Retrieved 28 November, 2006, from http://epub.mimas.ac.uk/papers/elpub2006/apps-elpub2006.pdf

Beattie, D. (1997). Retrieving the Irretrievable: Providing Access to "Hidden Groups" in Archives. *The Reference Librarian, 56*, 83-94.

Belkin, N.J., Oddy, R.N. & Brooks, H.M. (1982). ASK for information retrieval: Part I. background and theory. *Journal of Documentation, 38*(2), 61-71.

Broughton, V. (2004). Essential classification. London: Facet Publishing.

Chapman, A. (2004). Collection-level description: joining up the domains. *Journal of the Society of Archivists, 25*(2), 149-155.

Chapman, A. (2005). Collection descriptions: state of play. *Library & Information Update, 4* (4). Retrieved November 28, 2006, from http://www.cilip.org.uk/publications/updatemagazine/archive/archive2005/april/collectiondescriptionsapril05.htm

Chowdhury, S. & Chowdhury, G, G. (2004). Using DDC to create a visual knowledge map as an aid to online information retrieval. *ISKO8: Knowledge organization and the Global Information Society, July 13-16,*

2004, London. Retrieved November 28, 2006, from http://www.cis.strath.ac.uk/research/publications/papers/strath_cis_publication_333.pdf

Dewey, M. (2005). *Dewey decimal classification and relative index. Edition 22*. Dublin, Ohio: OCLC.

Dublin Core Collection Description Working Group. (2006). *Dublin Core Collection Description Application Profile*. Dublin, Ohio: DCMI/OCLC. Retrieved November 28, 2006, from http://dublincore.org/groups/collections/collection-application-profile/2006-08-24/

Dunsire, G. (2002). *Extending the SCONE collections description database for CC-interop*. Glasgow: Centre for Digital Library Research. Retrieved November 28, 2006, from http://ccinterop.cdlr.strath.ac.uk/documents/CCIExtendSCONE.pdf

Dunsire, G. (2004a). *Collection landscaping in the common information environment : a case study using the Scottish Collections Network (SCONE)*. Glasgow: Centre for Digital Library Research. Retrieved November 28, 2006, from http://ccinterop.cdlr.strath.ac.uk/documents/ccicldlandscaping.pdf

Dunsire, G. (2004b). *Output formats for collection-level descriptions from the SCONE database*. Glasgow: Centre for Digital Library Research. Retrieved November 28, 2006, from http://ccinterop.cdlr.strath.ac.uk/documents/CCISCONEOutput.pdf

Dunsire, G. (2006). Conspectus and the Scottish Collections Network: landscaping the Scottish common information environment. *Signum*, 3. Retrieved November 28, 2006, from http://pro.tsv.fi/stks/signumnew/200603/4.pdf

Dunsire, G. & Macgregor, G. (2003). Clumps and collection description in the information environment in the UK with particular reference to Scotland. *Program: electronic library and information systems*, *37*(3), 218-225.

Foulonneau, M., Cole, T, W., Habing, T, G. & Shreeves, S, L. (2005). Using Collection Descriptions to Enhance an Aggregation of Harvested Item-Level Metadata. In *5th ACM/IEEE-CS Joint Conference on Digital libraries 2005* (pp.*32-41).* New York: ACM Press. Retrieved November 28, 2006, from http://portal.acm.org/citation.cfm?id=1065385.1065393

Gabriel, C. (2002). Subject Access to Archives and Manuscript Collections: An Historical Overview. *Journal of Archival Organization*, *1*(4), 53-63.

Garshol, L, M. (2004). Metadata? Thesauri? Taxonomies? Topic Maps! Making sense of it all. *Journal of Information Science, 30*(4), 378-391.

Geisler, G., Giersch, S., McArthur, D. & McClelland, M. (2002). Creating Virtual Collections in Digital Libraries: Benefits and Implementation Issues. In W. Hersh & G. Marchionini. (Ed.), *2nd ACM/IEEE-CS Joint Conference on Digital libraries 2002* (pp.*210-218).* New York: ACM Press. Retrieved November 28, 2006, from http://portal.acm.org/ft_gateway.cfm?id=544265&type=pdf

Gilmore, M, B. (1988). Increasing Access to Archival Records in Library Online Public Access Catalogs. *Library Trends, 36*(Winter), 609-523.

Heaney, M. (2000). *An Analytical Model of Collections and their Catalogues*. Bath, UK: UKOLN. Retrieved November 28, 2006, from http://www.ukoln.ac.uk/metadata/rslp/model/amcc-v31.pdf

Heaney, M. (2005). *Users and Information Resources : An Extension of the Analytical Model of Collections and their Catalogues into Usage and Transactions.* Bath, UK: UKOLN. Retrieved November 28, 2006, from http://www.ukoln.ac.uk/cd-focus/model-ext/CD2-principles-v2-2.pdf

Hill, L, L., Janée, G., Dolin, R., Frew, J. & Largaard, M. (1999). Collection Metadata Solutions for Digital Library Applications. *Journal of the American Society for Information Science, 50*(13), 1169-1181.

ISO. (1985). *Documentation - Methods for examining documents, determining their subjects, and selecting indexing terms*. Geneva, Switzerland: ISO.

Lancaster, F, W. (2003). *Indexing and Abstracting in Theory and Practice (3rd Ed.)*. Champaign, Illinois: Graduate School of Library and Information Science Publications Office.

Lee, H-L. (2000). What is a collection? *Journal of the American Society for Information Science, 51*(12), 1106-1113.

Macgregor, G. (2003). Collection-level descriptions: metadata of the future? *Library Review, 52*(6), 247-250.

Maltby, A. (1975). *Sayers' Manual of Classification for Librarians*. London: André Deutsch Limited.

Nicholson, D. (2002). *CURL Study of the OCLC/Lacey iCAS Software : External Evaluator's Report :Final Report of the RSLP SCONE project, annexe A.3*. Glasgow: Centre for Digital Library Research.

Nicholson, D. & McCulloch, E. (2006). Investigating the feasibility of a distributed, mapping-based, approach to solving subject interoperability problems in a multi-scheme, cross-service, retrieval environment, *International Conference on Digital Libraries, 5-8 December 2006, India Habitat Center, New Delhi*. New Delhi: TERI.

Nicholson, D., Dawson, A. & Shiri, A. (2006). HILT: A pilot terminology mapping service with a DDC spine. *Cataloging & Classification Quarterly*, *42*(3/4), 187-200.

NISO. (2005). *Collection Description Specification [Draft – Z39.91]*. Bethesda: NISO. Retrieved November 28, 2006, from http://www.niso.org/standards/standard_detail.cfm?std_id=815

OCLC. (2006). *Learn more about the DeweyBrowser*. Dublin, Ohio: OCLC. Retrieved November 28, 2006, from http://www.oclc.org/research/researchworks/ddc/browser.htm

Powell, A. (2000). *RSLP Collection Description Schema*. Bath: UKOLN. Retrieved November 28, 2006, from http://www.ukoln.ac.uk/metadata/rslp/schema/

Powell, A., Heaney, M. & Dempsey, L. (2000). RSLP Collection Description. *D-Lib Magazine*, *6*(9). Retrieved November 28, 2006, from http://www.dlib.org/dlib/september00/powell/09powell.html

Ribeiro, F. (1996). Subject Indexing and Authority Control in Archives: the need for subject indexing in archives and for an indexing policy using controlled language. *Journal of the Society of Archivists, 17*(1), 27-54.

Saeed, H. & Chaudhry, A, S. (2002). Using Dewey Decimal Classification scheme (DDC) for building taxonomies for knowledge organisation. *Journal of Documentation*, *58*(5), 575-583.

Shreeves, S, L. & Cole, T, W. (2003). Developing a Collection Registry for IMLS NLG Digital Collections. In *DC-2003: Proceedings of the International DCMI Metadata Conference and Workshop* (pp.241-242). Dublin, Ohio: DCMI/OCLC. Retrieved November 28, 2006, from http://www.siderean.com/dc2003/705_Poster43.pdf

Vizine-Goetz, D. (2002). Classification Schemes for Internet Resources Revisited. *Journal of Internet Cataloging, 5*(4), 5-18.

Vizine-Goetz, D. (2006). DeweyBrowser. *Cataloging & Classification Quarterly*, *42*(3/4), 213 – 220.

Wiesman, F., Van den Herik, H, J. & Hasman, A. (2004). Information Retrieval by Metabrowsing. *Journal of the American Society for Information Science and Technology, 55*(7), 565-578.