Work package 4 : Metadata creation
Rodney Brunt, Elizabeth Guest

Progress report 13 vii 2007

Introduction

Initial explorations have addressed:
- the nature and content of the metatag surrogate (i.e. document representation)
- the indexing of the tag to aloe access to it and thence to the learning object (LO) itself:
  - indexing to allow 'known item access'
  - indexing to allow 'unknown item access'[1]

the proposed approach is to explore methods and processes in existing repositories to see if any might be adopted with little or no change or adapted and applied in light of technological development since their first (or most recent) appearance.[2]

Objectives
The package will involve the achievement of the following objectives:

1. Document representation

Metadata tags are essentially catalogue records embedded in the electronic documents to which they relate much as the title page and CIP[3] records provide the metadata of books.

In order to establish the metatags some means have to be found to create the basic surrogate to provide first an authoritative and dependable source for cataloguing purposes and second a source of 'derived indexing'[4] to facilitate retrieval on an interim basis while awaiting further processing to create the enhanced final metadata.

2. Surrogate indexing

Once some form of representation has been provided it will be possible to consider its indexing to provide for more exhaustive retrieval.
This will probably initially involve 'assigned indexing'[5] to cover the subject or unknown item approach; but might also be concerned with known item access as the repository grows and the need for authority control of creator/author names becomes apparent.[6]

---

[1] 'known item access' permits retrieval of specific items by identifying criteria such as author name, title, control number, &c.; 'unknown item access' retrieval usually by subject or presentational form, language, &c.
[2] Brunt, R. (2006) *Getting, hiding, finding: information management perspectives on the Replika resource base*. Unpublished abstract submitted to the proposed Replika conference, Murcia, May, 2006
[3] Cataloguing in publication
[4] 'derived indexing' allows free text searching of all or sections of electronically stored documents
[5] 'assigned indexing' is that provided via human intervention usually involving reference to vocabulary control standards such as classification schemes and thesauri
[6] authority control involves the establishment of definitive forms of names and identification of variants

3. Processes

Some means by which the authors of the LO might be enabled to create the initial metadata must be found. This will involve the design of a suitable and simple PRO forma with guidance to encourage the production of an extended 'title page' which could include not only author, title and date of creation but also a brief indicative abstract and a limited number of key indexing terms.

Concern will be not to overburden creators with tasks not usually give to authors of books, but which will be acceptable to those used to supplying such data to journal editors in submission of articles for publication.

Having got the creators to produce the 'title pages' a procedure will need to be established whereby professional upgrading to full meta tags might be achieved.

4. Pitfalls

Avoidance of certain difficulties, including resistance on the part of LO creators, failures of consistency, problems with key words, &c must be assured. This might be achieved by establishing a quality control procedure and careful design of the guidance notes.

5. Control of costs

Economic processing must be achieved in the overall scheme, not only in the creator-produced initial metadata but also in the full version of the tags. Questions about who should generate the final versions must be addressed and will include consideration of various types of models such as auto generation (via submission of URL), professional generation (by cataloguers) and even author generation where creators can show competence in applying appropriate data standards.
Again, appropriate quality control mechanisms will need to be designed and incorporated.

6. Testing and evaluation

Testing and evaluation will have to be achieved under two headings:

• the metadata creation process

• retrieval performance

Te first will address the production process itself to ensure that meta tags adhere to prescribed standards covering identities, titles, edition/version statements, control numbers, dates, &c as supplied in the 'title page' as well as any need for revision in light of the editorial process (e.g. initial scan of the LO to check for presence, quality and appropriateness of the 'title page' and if absent or deficient its return to creator).

The second will entail use of a test bed (c.f. Cranfield, SMART, TREC[7]) to test and refine the information retrieval elements of the system for both derived and assigned indexing performance.

7. Methodology

A literature review will involve an exploration of existing information retrieval systems and processes to see what might be already available. Specific aspects for consideration will include:

- DC-dot, Nordic Template, Warwick initiative[8], &c for generation of full tags;
- PRECIS,[9] View-based searching,[10] ULM, FRBR[11] for indexing and retrieval;
- review of the work of Greenberg et al in metadata creation;

though this list is not exhaustive and will doubtless be expanded in the light of the review.

Other potentially useful sources might be found in content management systems and integrated digitised library systems associated with e-library initiative, including those being explored currently by Leeds Metropolitan University Library and attendance at workshops on e-library building and bibliographic control such as those organised by UKeIG.

An evaluation of these will result in identification of potentially useful existing processes or point to the need for new design.

Review of processes explored in the projects will offer additional insights into potentially useful approaches to the creation and indexing of the full meta tags.

It is believed that this work package represents advance on HSLI and Replika in that it incorporates control of the design and production of the author produced initial surrogate which will give, first, a robust tag for initial retrieval; and, second, a sound basis for the enhancement of the 'title page' by professional cataloguers to the full tag should that method be adopted.

8. Testing and evaluation processes

The testing and evaluation process will address:

- Retrieval:
    - by browsing
    - by purposeful search

[7] Ellis, D. (1996) *Progress and problems in information retrieval.* LA Publishing, provides a useful outline of a number of potentially useful approaches
[8] based on Dublin Core Metadata Initiative
[9] Preserved Context Index System
[10] Pollitt, AS, Smith, MO & Ellis, GP. HIBROWSE for bibliographic databases. *Journal of information science*, 20(6):413-426
[11] IFLA Study Group on the Functional Requirements for Bibliographic Records (1998) Functional requirements for bibliographic records : final report. K.G. Saur,.

- Training for use
    - degree of intuitiveness
    - need for learning on part of users

These would be used to devise a set of criteria against which the repository IR mechanisms could be tested.

One possible approach would be to produce a number of different versions of a pilot repository with different degrees of bibliographic control, e.g. one with only the initial 'title page'; one with professionally enhanced tags; one with professionally produced tags; one produced automatically, say by DC-dot. Since the contents of the repository would be a known the performances of these versions could be matched against the criteria established.

Conclusions

The above are initial thoughts based on a limited amount of deliberation so far. It is expected that the greater resources expected to be available after September will facilitate further and more rapid progress.

rmb
eg
July, 2007