Author

Dr Marie-Louise Ayres
Executive Manager, AustLit: Australian Literature Gateway
Academy Library
UNSW@ADFA

Co-Authors

Kent Fitch, Project Computing
Annette Scarvell, UNSW@ADFA
Kerry Kilner, UQ

Title

AustLit: A Gateway on Steroids[1]

http://www.austlit.edu.au

Abstract

AustLit: Australian Literature Gateway provides access to bibliographic citations and a developing body of full text for more than 350 000 Australian creative and critical works (regardless of format), and to biographical and organisational information on more than 40 000 Australian authors and literary organisations. The Gateway was formed by a consortia of eight universities and the National Library, incorporates records from a number of previously existing databases, and aims to provide Australian students and researchers with a single access point for their Australian literature information needs. The Gateway system was custom built and employs leading edge knowledge models (including IFLA's *Functional Requirements for Bibliographic Records* for works; the INDECS and Harmoney models for agents and their relationships with works; and Topic Maps for creating flexible relationships) and enabling and delivery technologies such as Z39.50, XML and XSL. The Gateway is the first large scale implementation of IFLA's *FRBR* model, and is an early adopter of INDECS, Harmony and Topic Maps. This paper will report on the reasons behind the choice of models, how these models were implemented, and what the implications of adopting these models have been from both the production system and user perspectives.

Keywords

Australian literature; Functional Requirements for Bibliographic Records; Literary databases; Subject Gateways; XML.

Introduction to the AustLit Gateway

In late 1998, a group of eight Australian universities – UNSW@ADFA, UQ, Monash, Flinders, Deakin, UWA, Canberra, and our hosts, the University of Sydney – together with the National Library of Australia, commenced planning a joint application for an Australian Research Council 'Research Infrastructure and Equipment Fund' grant to establish an Australian Literature Subject Gateway.

All eight universities had made long term investments in Australian Literature biographical and bibliographic 'products', some of which were available to the public (either in print or electronic formats), and some of which had not yet been made so available. The first Gateway grant application, submitted in May 1999 for 2000 funding, was successful - in fact, it was a 'landmark' grant amount for a humanities project. We submitted another application in May 2000 and were again successful, receiving a further year's funding for 2001. We are currently awaiting the outcome of a further application for 2002 funding. The partners' commitment to the project is perhaps best demonstrated by the fact that all eight universities and the National Library have committed cash and in kind contributions to at least the end of 2003, regardless of the outcome of this or other applications.

The various Australian literature information 'products' all came from different information traditions. UNSW@ADFA's AUSTLIT database - by far the largest of the databases, and available commercially to the public for more than a decade – was maintained within a library information space, and was very robust. Unfortunately, the AUSTLIT software – a customised version of URICA – was obsolete, and was unable to support Z39.50 or other interoperability protocols. In addition, its structures and practices were incompatible with those of major information sources such as Library catalogues and the National Bibliographic Database. The other eight bibliographic and biographical databases had generally arisen from print based bibliography projects, and were therefore more oriented towards description than to providing access or revealing complex relationships.

As was perhaps inevitable, the database producers had all started to duplicate each other's work, even though the single AUSTLIT database took a generalist, mostly contemporary monograph and journals sources approach, and the other databases tended to be in specialist areas, and to have arisen from specific areas of academic enquiry. These included author-centric monograph based bibliography (the Bibliography of Australian Literature Project and the List of Australian Writers) produced jointly by Monash and UQ), children's literature (Lu Rees Archive at the University of Canberra), multicultural literature (Deakin), Western Australian and South Australian writing (at the local universities), and drama (Monash and UQ). The University of Sydney's SETIS service[2] was in a different information space again: one that provided high quality and robustly encoded and searchable electronic editions of early Australian literary works, but did not provide those editions in either an author based or works based context.

Resources in this area are scarce and always under threat. By 1998, it had become imperative to adopt a collaborative and cooperative approach to ensure that all known resources could be described in a single information space, and to ensure that none of these scarce resources was being wasted through duplication.

Choice of Models

From the outset of our project, it was clear that although our service would be called a Gateway, and we have benefited significantly from membership of the Australian Subject Gateways Forum[3], the AustLit Gateway would be very different to other Australian and international Gateways in both its size and its scope. We realised that we had a rare opportunity to go back to 'first principles', from which we developed a clearly articulated and common set of values:

- we wanted to provide a single access point to information about Australian writers and their writing *regardless* of whether that information was in print or electronic format;
- we were not interested in being a 'catalogue': our whole *raison d'etre* was to provide enhanced and enriched research-conducive information;
- we were very concerned to represent the publishing histories of all works, and, as far as possible, to contextualise the works;
- we wanted to draw rich relationships between a whole range of entities such as authors, organisations, works, places, times, subjects, settings, and publishers.

We were very fortunate to begin with this strong sense of our professional and information values. We were also very fortunate to have the assistance of two exemplary thinkers and implementers: Judith Pearce, Director of Web Services at the National Library of Australia, and Kent Fitch, of Project Computing. Judith had strong experience in standards-based library web services. Kent had no previous *library* experience when he started developing the AustLit Gateway, but had a great deal of experience in database design, especially XML for large information spaces, and brought a great depth of modeling experience to the project.

Our modeling period was intense, and mostly involved our development team of four: myself, Kent, Annette Scarvell (Content Manager at UNSW@ADFA) and Kerry Kilner (Content Manager at UQ). We were committed to spending as much time as necesssary to undertake this task, rather than make the mistake of leaping to a solution. As an aside, I believe that both our modeling and the succeeding implementation have benefited considerably from the fact that this development team had representatives from all possible information sectors feeding into the service. My background is academic (in Australian literature), but my major work experience has been in libraries; Kent is a programmer and developer of more than twenty years standing; Annette is a librarian and holder of an additional Business Administration degree; and Kerry is an experienced research bibliographer of Australian literature and drama. Our data model is publicly available on the AustLit website [4].

Quite early on, as we were articulating our desire to represent publishing histories for works, Judith pointed us to the International Federation of Library Associations' *Functional Requirements for Bibliographic Records* (*FRBR*) model, published in 1998[5]. This was definitely our *Eureka!* [6] The *FRBR* model includes the concepts of:

- the Work: an abstract concept (e.g. the novel *Voss* by Patrick White)
- the Expression: how that Work is realised (e.g. White's original version of the novel in English or the German translation by John Stickforth of the novel *Voss* by Patrick White)
- the Manifestation: how that Expression is made concrete (e.g. the 1958 Kiepenheuer & Witsch publication of Stickforth's translation of the novel *Voss* by Patrick White)
- the Item: the individual item on the Library shelf (e.g. the copy of the 1958 Kiepenheuer & Witsch publication of the John Stickworth translation of the novel *Voss* by Patrick White, held at the National Library)

The model was strong and robust, and seemed to do most of what we wanted to achieve with our bibliographic description. We augmented this model with 'event modeling' (based on the ABC Harmony and INDECS models[7]):

- works have a **creation** event
- expressions have a **realisation** event
- manifestations have a **manifestation** event

Works can be expressed one or many times, Expressions can be manifested one or many times, and manifestations can result in one or many items. Works, Expressions and Manifestations all have attributes, and Creation, Realisation and Manifestation **events** all have **attributes.** Works, for example, can have subject attributes – they can be *about* things - and work creation events can have creators and places and dates of creation - although, as attributes of an abstract concept, these are rarely known.

The user view of the AustLit Gateway record for *Voss* at:

http://www.austlit.edu.au:7777/presentations/staticHTMLSnapshots/voss.html

shows many of these events and attributes.

Examining the Kinetica holdings for *Voss* [8]:

http://www.austlit.edu.au:7777/presentations/staticHTMLSnapshots/vossHoldings.html

quickly reveals the fundamental differences between *FRBR* and traditional MARC cataloguing.

One of the great advantages of this model – and the reason the model is of significant interest to large cataloguing organisations – is that electronic editions of works can be

represented in their own right *and* in relation to originating print editions[9]. Our record for Marcus Clarke's *His Natural Life*, for example:

http://www.austlit.edu.au:7777/presentations/staticHTMLSnapshots/hisNaturalLife.html

allows us to adequately represent that the SETIS 1997 electronic edition of the novel is a **manifestation** of the 1888 Richard Bentley and George Robertson **expression** of the 1870 **work**.

The model was also augmented with our concept of a **SuperWork**, for example to represent the relationship between two works, such as the novel *Voss* and the opera *Voss*, which cannot be seen as merely an expression of the Voss work:

http://www.austlit.edu.au:7777/presentations/staticHTMLSnapshots/vossSuperWork.html

and our representations of agents (authors and organisations), other entities such as subjects, settings, and awards[10] and relationships between all the Gateway entities through the use of Topic Mapping[11]. In the Gateway, agents, for example, also have events and attributes associated with them. Our record for Patrick White includes birth and death events, date and place attributes of those events, award events and attributes of those award events, and attributes such as gender, pseudonyms, nationality and biographical information (other agents have cultural heritage attributes where these have been claimed by the author) and, of course work relationships including creator and subject relationships:

http://www.austlit.edu.au:7777/presentations/staticHTMLSnapshots/agentWhite.html

All AustLit entities, including events and attributes, are **topics**, and relationships between those entities are also topics. The AustLit Gateway includes more than 3.3 million topics.

Of course, once we had decided that this was what we wanted to achieve, it was clear that we would need to build, rather than buy, a system: there are currently no commercial systems which support all these models. The basic design documents relating to our custom built system are publicly available at our website [12]: the following diagram is a simple representation of the elements of our system:

http://www.austlit.edu.au:7777/design/custom2.gif

With the exception of the Oracle database – which our University licence made available to us – all other software is open source. We purchased a Sun Blade Server on which to run the system.

Although the topics and their relationships are stored in conventional (but unusally highly normalised) relational database tables, the system converts the data into a common XML

format at an early stage of output processing. From this common XML format, information is transformed into the desired final output format (typically HTML) using XSL (eXtensible Stylesheet Language). The XML representation contains enough information to generate alternative encodings such as MARC or to augment the HTML with Dublin Core or RDF metadata.

Implementation of Models

At the outset of the implementation phase, we believed that the major risks lay in the complexity of designing a database to accommodate the *FRBR*, INDECS and Harmony models along with all the multitudinous relationships we had mapped out, and the likely performance of a highly normalised (ie consisting of some millions of 'topics') database. As it turned out, thess was not the major hurdles we had envisaged, and we have been extremely pleased with the design outcome and database performance.

By February 2001 – 10 months after Kent started working with us, and just 7 months after acceptance of the proposed model and design by the Gateway consortium - we had designed the database and the maintenance interface, migrated most data and trained our 20+ librarians and bibliographers from around the country to use the new system. The use of IE5.5 as the maintenance interface has been a great boon: there was no need to develop and provide client software, and startup costs for new maintainers are minimal: all that is needed is a reasonable PC, IE5.5 and access to a network.

One of our major worries in adopting the *FRBR* model was that it could prove too expensive to create and maintain *FRBR* records. This has certainly not proved to be the case. Educating our staff about the *FRBR* model certainly took a lot of work, especially because practical implementation raised many issues. But once they were familiar with the model, they loved the fact that it allowed them to represent works in a rich context. They also thoroughly enjoy the maintenance interface which gives them many choices about how to describe works and authors, and gives instant satisfaction: create or edit the record, update it, see it in the browser immediately. We also have a very effective review interface, which allows our two Content Managers to review work and provide timely feedback by email. From a management point of view, I am absolutely thrilled about the productivity of the staff, and our earlier fears about the expense of both implementing and using the model have been emphatically put to rest.

Having said that, there certainly have been issues in implementing the *FRBR* model and the other elements of the AustLit model which intersect with it, such as representations of events and agents. The major issue actually has nothing to do with the models we chose. We substantially underanticipated the risk which lay in migrating a range of existing non-standards based databases to the new structure. To our horror, we found that much of the data we had all thought was pretty robust was highly ambiguous. Kent has spent many programming hours trying to automate as much matching as possible. While we budgeted for significant librarian hours to work on merged and ambiguous records, we have spent nearly twice as much on this task as originally anticipated. Every new database brought new problems and we were not able to reapply the last solutions! While

these problems have now been conquered, they do raise issues regarding future AustLit outcomes. The AustLit Gateway has several further specialist bibliographies and datasets 'on offer'. We are attempting to establish a robust costing for integrating such datasets, and to encourage their 'owners' to seek completion grants to enable their work to be made publicly available, as the Gateway cannot meet these costs despite the clear benefits of enriching the information base. This has raised questions about whether DETYA research points might accrue to academics from such collaboration.

We also encountered significant issues relating to interpretation of the FRBR and the pragmatics of implementation. The model was clearly written with a monograph emphasis (although the model demonstrates that it can be used for other types of works, such as performances), and I think if we were only concerned with monographs, implementation would have been more straightforward. But our information universe includes:

- monographs
- parts of monographs, ie. introductions, critical articles etc.
- selected work and anthology monographs, with many individual inclusions
- journals and newspapers, with many individual inclusions
- ejounals and sites, again with many individual inclusions
- individual items: poems, short stories, reviews, criticisms etc
- relationships between individual items: not just 'part of selected work A' but also 'part of poem sequence A', 'part of publisher series B' etc.

Once we got into really practical issues, we had to deal with complex issues such as:

- a novel first appearing in two different newspaper expressions (different versions of the novel), both under a pseudonym, then a book expression (a different version again), still under a pseudonym, and finally in a new book manifestation of that expression, using the author's 'real name'
- poem sequences which appear in one expression containing six poems, in the next expression containing four poems, and in the third, five poems (we would like to ban poets from ever changing their sequences and admit candidly that we've 'fudged' the *FRBR* logic on this one)
- trying to figure out whether 'part of series'-ness is a work, expression or manifestation attribute (logically, we decided, expression/manifestation, but pragmatically, we've left it at work)
- Do we have a new expression of a selected works or anthology if it includes 100 of the same poems as the first expression, but adds 3 new ones and omits 6 old ones? We decided yes - but that has raised issues: do we record the contents of each of these expressions? Only the differences? What's the pragmatic solution?

We have also spent quite a lot of time refining our three 'upper level' work attributes:

- workType (basically to do with number: is this a single thing, a collection of things or part of a thing. We have around 12 workTypes: single work, selected work, periodical issue, website, author series etc.)
- formType (we have around 30 forms: novel, poetry, review, obituary etc.)
- genreType (we have around 15 genres: romance, science fiction, young adult etc.)

all now documented in our online manual[13], but no doubt subject to further adjustment as the 'real world' forces us to reconsider both logical and pragmatic issues. We have also put a great deal of our resources into redeveloping what was a non-hierarchical and less than wieldy thesaurus for subject indexing. We now have subject authority hierarchies for general concepts, place names, time periods, and literary awards.

AustLit Outcomes

The goals we set out in our original grant application documents have all been met. Indeed, many of them have been far exceeded, as we certainly did not anticipate being among the earliest large-scale implementers of several new models. Nor did we anticipate how rich our data model and design have allowed our service to be. Of course, we need to find futher resources to collect and record the data required to maximise its utility.

Our major outcomes have been:

- The establishment of a truly collaborative and cooperative framework, not only between different institutions, but between professions or 'information traditions' which had not previously worked together so closely: academics, librarians, bibliographers, programmers and service developers.
- The concurrent development of a strong **data** model and a strong **business** model, and effective implementations of these models.
- The implementation of new and powerful data models, with a very large dataset as an effective testbed.
- Simultaneous amalgamation of all past and future work into a single database *and* successful retention of the established 'research identities' of contributing institutions through providing subset views[14].

Our service outcomes to date are:

- migration of 350 000+ work and 40 000+ author records.
- more than 4500 new author records have been added, and 4500 existing author records updated in the first six months of operation (including adding rich biographical data)
- more than 14 000 new work records have been added, and a further 28 000 existing work records updated in the same period (including adding rich Expression and Manifestation information).
- comprehensive online manual for our staff, which addresses many of the finer points of *FRBR* decision making.
- the service will be released for beta testing (as a free trial) in late September 2001.

- full text work, some of it encoded and mounted through SETIS, is in production, and while it will take some time to finalise permissions, we aim to provide access to 20 000+ full text documents by the end of 2001.

Our business outcomes to date are:

- An agreed business model[15]. AustLit is not for profit, but we will be charging annual subscriptions to cover our central 'publication' costs and to try to return some royalties to the eight universities providing data . Basic information on every Australian author will be available free: the 'rich' data will be available by subscription only.
- An agreed legal agreement[16]. The agreement is based on the partners providing the Gateway with a perpetual, non-exclusive licence to publish records, and on a formula for meeting central costs before royalties.
- The selection of a vendor. Subscription-based access will commence at the beginning of January 2002.

We look forward to the opportunity to achieve new goals over the next 2-3 years, and in particular, to:

- develop the Gateway's innovative design infrastructure to offer sophisticated interoperability with other discovery services[17];
- maximise the full service capacity of the Gateway by enriching its content base, especially in those areas which can enhance relationship mapping;
- exploit the capacity of this powerful set of research resources for a range of tertiary needs, including integration with flexible learning services such as student portals and distance education programs; and
- deliver significant full content in a discipline which is inadequately supplied by commercial providers.

We are particularly enthusiastic about developing the 'relationship' mapping which our Topic Map basis facilitates, and on a browsing interface which can 'expose' those relationships. Examples include being able to represent things such as:

- writers born in the Hunter Valley, or the Darling Downs, or the WA 'wheatbelt' or 'the outback'; all works published in Gippsland, or the South Coast, or the 'Top End', to assist with studies of regionality.
- writers who were members of the Jindyworobak movement, or the Realist Writers, or the Lindsay circle, or the Generation of '68, to assist with studies of literary movements.
- writers and works in the context of historical events of the day.

The list is really only limited by our imaginations: the ability to implement, of course, will be limited by our available resources!

Outcomes for the Humanities Information Community

We believe this project has achieved powerful outcomes for AustLit Gateway users. The Gateway is an authoritative service delivering quality information on Australian literary authors and works. It draws together information on writers and works in powerful new ways which facilitate understanding of the *contexts* in which Australian literature exists, and has the capacity to develop those contexts much more richly. Our mission now is to augment this contextual environment with as much full text as is possible: not because print versions of works are difficult to locate in Australian libraries (unlike science journals, Australian literary monographs and journals are not expensive), but because the developing user community *expects* full text, and may only appreciate the rich context the Gateway provides if that context is synchronised with instant access to materials.

The benefits of the Gateway project, however, are not limited to Gateway users. The AustLit Gateway has demonstrated the power and effectiveness of collaboration between academe and the information profession, and the power and effectiveness of collaboration between a range of stakeholders. It has provided the first large-scale implementation of the *FRBR* model in the world, and the first large-scale augmentation of *FRBR* with the event modeling and other extensions proposed in responses to that model. It has proven that it *is* possible to implement these models, and the underlying normalised database design without compromising either performance for users, or incurring unsustainable creation and maintenance costs. The authors of this paper are proud of their involvement in this innovative project – and keen to work with the next generation of implementers to improve on and 'multiply' this success.

Conclusion

AustLit is a *literature* Gateway. We do not forget what made us passionate about literature in the first place, and it was perhaps inevitable that at least one of our wonderful team of librarians and bibliographers would find a literary way to express some of our experiences in implementing the *FRBR* and other models. I will conclude by reading you a poem written by Carol Hetherington, of the University of Queensland team. By way of explanation, a 'buggered thread' relates to a Java error thread…

The AustLit Lament (Apologies to T.S.Eliot[18])

Here we go round the FRBR
FRBR, FRBR
Here we go round the FRBR
In our bibliographic dilemma?

Between the creation
And the realisation
Between the expression
And the manifestation
Comes the hesitation

Between the selection
And the collection
Between the description
And the illustration
Comes the indecision

Between the translation
And its origination
(Put it in the collation?)
Comes the procrastination
And the mystification
And the consternation

Is this the way the record ends?
Is this the way the day ends
Not with an update
But with a buggered thread?[19]

Carol Hetherington

[1] The AustLit developer, Kent Fitch, coined the term 'a thesaurus on steroids' to describe the Topic Map standard.

[2] See http://setis.library.usyd.edu.au

[3] See http://www.nla.gov.au/initiatives/sg/index.html

[4] See http://www.austlit.edu.au:7777/DataModel/index.html

[5] The model is available at http://www.ifla.org/VII/s13/FRBR/FRBR.pdf (full model) and http://www.ifla.org/VII/s13/FRBR/FRBR.htm (no tables or figures).

[6] AustLit lists 134 works which include 'Eureka' in their titles or first lines, and 60 works about the Eureka Stockade as at September 2001.

[7] See http://archive.dstc.edu.au/RDU/staff/jane-hunter/harmony/workshop_notes.html and http://www.indecs.org/pdf/framework.pdf

[8] AustLit uses a custom designed Z39.50 client to interrogate and return relevant holdings records from the National Bibliographic Database, and will use a version of this client to do the same with the National Library's *Register of Australian Archives and Manuscripts* in the near future. The AustLit team aims to establish further interoperability, such as the ability to download suitable MARC records directly from the NBD, ready for augmentation and, possibly, the upload of Gateway records back to the NBD.

[9] The addition of event modeling to represent creation, translation, encoding, abridgement, adaptation and publication etc. events will also facilitate AustLit's future integration with Digital Rights Management initiatives.

[10] All represented in the AustLit thesaurus at http://www.austlit.edu.au/common/newHierarchicalSelect.html

[11] See http://www.infoloom.com/tmsample/bie0.htm

[12] See http://www.austlit.edu.au:7777/design/index.html. Our underlying database design is not available to the public.

[13] All at http://www.austlit.edu.au/common/manual/WorksContents.html, and despite some recent misgivings that perhaps 'Works' are so abstract that we can't possibly assign any of these attributes to them!

[14] See http://www.austlit.edu.au/specialistDatasets

[15] See http://www.austlit.edu.au:7777/BusinessModel/bmindex.html

[16] See http://www.austlit.edu.au:7777/BusinessModel/ALEGagreement.htm

[17] For example, the *Guide to Australian Literary Manuscripts* at http://findaid.library.uwa.edu.au

[18] Just for the record, Australian literature's most prolific pseudonymist, John Clarke, born New Zealand 1948, has published two Eliot parodies as T.S. Eliot (also known as Tabby Serious Eliot).

[19] Yes, a record for this poem exists on AustLit.