# Greenstone: a platform for semantic digital libraries

Annika Hinze[1], George Buchanan[2], David Bainbridge[1], and Ian H. Witten[1]

[1] University of Waikato, New Zealand
[2] University of Swansea, United Kingdom
hinze@cs.waikato.ac.nz, g.r.buchanan@swansea.ac.uk,
davidb@cs.waikato.ac.nz, ihw@cs.waikato.ac.nz

## 1 Introduction

This chapter illustrates the impact on a well–known digital library system – Greenstone – when it is moved from fixed modules and simple metadata–based structures, to open semantic digital library modules. This change has profound effects on the tools available to end–users to retrieve relevant content from the library, and an equally significant impact on the digital library (DL) architecture. Most current DL systems contain protocols for internal communication that define information exchange solely in terms of searching, browsing, and document retrieval. These communications reflect traditional user interactions in the library. However, this regimented approach results in inflexible systems that are difficult to extend to support other retrieval techniques. Furthermore, simple field–based metadata limits the ability of the DL to connect or disambiguate key items of information, impeding the precision of retrieval.

Greenstone, an open source digital library toolkit that has developed over the last 10 years [14], forms the basis for the work described here. The software comes in two flavours: Greenstone 2 and Greenstone 3. The former exemplifies the classic form of digital library, with the added twist that (through UNESCO involvement) it is capable of running on primitive computing platforms that are common in developing countries (e.g. Windows 3.1 using Netscape 4). The latter is a reimplementation that is backwards compatible with the earlier version but far more ambitious in its goals. Particularly germane to the present work is the fact that it adopts an open protocol that works in tandem with a dynamic, componentised architecture [2].

In this article we describe the semantic aspects of the Greenstone 3 design and compare it with the earlier version, as a representative example of the archetypal approach. To illustrate key design elements, we draw upon three examples—an alerting service, ontology-enhanced representation and retrieval, and interoperability with a tourist information system—where many of the required user tasks and system features cannot be achieved through traditional digital library capabilities alone. These build upon the Greenstone 3 infrastructure. The details are developed in two steps:

1. *Semantics of documents and collections.* In the first step, we describe and analyse typical user tasks in the Greenstone Alerting Service. These tasks

call for a semantic model for digital library collections and content. We show how this requirement can be met by providing an ontology-based extension to the Greenstone Librarian Interface, an interactive subsystem for creating and maintaining digital library collections.
2. *Semantics for services and collaboration.* In the second step, we address the semantic issues of collaboration within a software framework. We study interoperability between the mobile Tourist Information Provider (TIP) information system and Greenstone, which together provide location-based access to digital library documents. This example shows how the software architecture of Greenstone 3 supports different semantic models.

The structure of the chapter is as follows. We begin with an introduction to the Greenstone system. Then we detail Greenstone's solutions for semantic issues, both at the collection level (Section 3) and for collaborating modules and systems (Section 4). We conclude with a general discussion about interoperating between ontologies/semantic models and general-purpose digital library software (Section 5).

## 2 Greenstone Digital Library Software

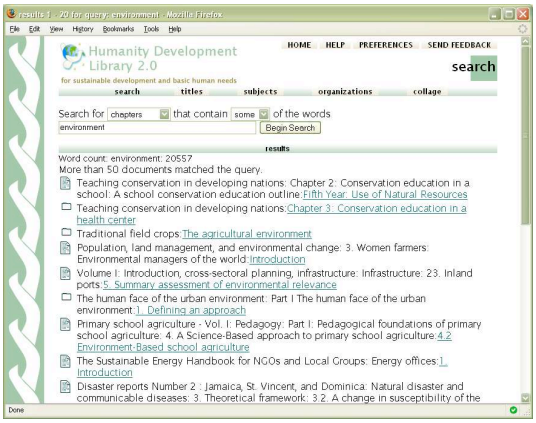
Greenstone is an open source digital library toolkit [13]. Used out of the box it provides the ability to create collections of digital content, to display the content in a web browser and to access and search the collections that have been built. Through UNESCO sponsorship the software is fully documented in English, French, Spanish, and Russian; in addition, its web interface has been translated into over 40 languages through volunteer efforts.

Countless digital libraries have been built with Greenstone since its public release on SourceForge in 2000: from historic newspapers to books on humanitarian aid; from eclectic multimedia content on pop-artists to curated first editions of works by Chopin; from scientific institutional repositories to personal collections of photos and other document formats. All manner of topics are covered: the black abolitionist movement, bridge construction, flora and fauna, the history of the Indian working class, medical artwork, and shipping statistics are just a random selection.

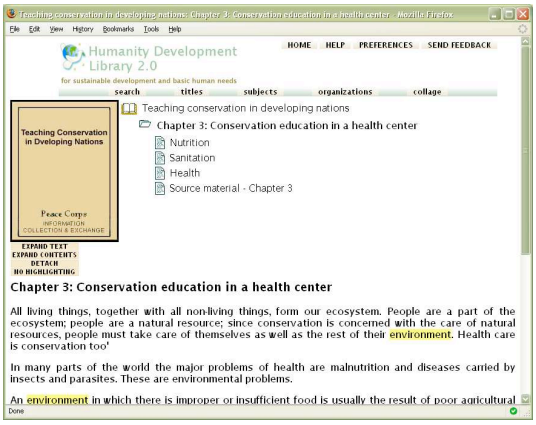
A wide variety of formats are accommodated, including HTML, PDF, OpenOffice, Word, PowerPoint, and Excel document formats; MARC, Refer, Dublin Core, LOM (Learning Object Metadata) and BibTeX metadata formats; as well as a variety of image, audio, and video formats. Greenstone also supports numerous standards including OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting), Z39.50 and METS (Metadata Encoding and Transmission Standard) to assist interoperability. Export options include Fedora, DSpace and MARC. See our web-site www.greenstone.org for more details.

An end-user's experience of Greenstone is through a web interface, such as the one shown in Figure 1, taken from the Human Info NGO's *Humanity Development Library*.[3] Documents in this collection can be searched by chapter title,

[3] http://www.nzdl.org/hdl




(a) Greenstone Collection interface




(b) Greenstone Librarian Interface

**Fig. 1.** Screenshots of Greenstone readers' interface.

in addition to full text searching by chapter or entire document. Alternatively, users might choose to browse alphabetically by title, or hierarchically by subject or organisation. In Figure 1(a) the user has searched within chapters for the word "environment" with a ranked listed of matches displayed; in Figure 1(b) the user is viewing the document that results from selecting the second matching item: Chapter 3 of *Teaching conservation in developing nations*.

Figure 2 shows the Greenstone Librarian Interface (GLI), a graphical application for creating and maintaining collections such as the *Humanity Development Library*. Through a system of tabbed panels accessed along the top of the interface, the digital librarian decides what files to include in the collection, what metadata is manually assigned (in addition to that automatically extracted by Greenstone from the source files), the collection's searching and browsing capabilities, and the customisation of presentation details.

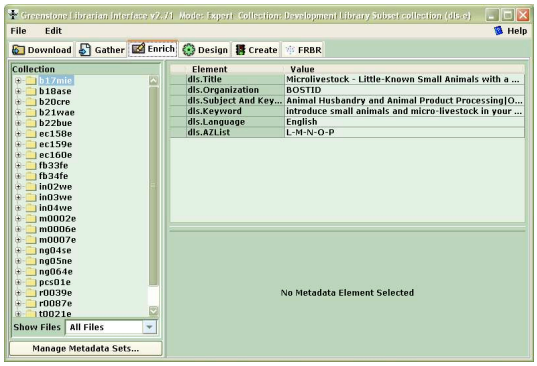



**Fig. 2.** Screenshot of the Greenstone Librarian Interface, including the FRBR ontology based extension.

Both Greenstone 2 and 3 provide the above capabilities—the archetypal digital library—but differ radically in implementation. Version 3 is fully built on open standards technology. Its predecessor was designed before many standards (such as XML and SOAP) existed, and although it has kept pace with developments, this has involved building onto a framework that was never designed for such dramatic extensions. Additionally, Greenstone 2 follows a traditional client server model, the protocol of which, as noted earlier, is fixed. In contrast, Greenstone 3 is based upon a distributed network of modules and uses SOAP to stream XML messages between them, with the option of customisation of messages at any point through the use of XSL Transforms. Dynamically loadable modules are layered on top of this communication channel. A "describe-yourself" call is a mandatory fixture that allows service discovery in a heterogeneous world of communicating applications.

Figure 2 is actually a Greenstone 3 version of the Greenstone Librarian Interface, built using this very capability. Notice the rightmost tab (labelled "FRBR") which indicates that GLI has been initiated with the *Functional Requirements for Bibliographic Records* [10] extension loaded—the ontology enhanced feature discussed in Section 3.2. Further technical details of the two architectures are discussed in Section 4.2. Henceforth we use "Greenstone" to denote the semantic-capable version of the software, and write "Greenstone 2" when it is necessary to refer to the earlier version.

## 3 Semantic model for documents and collections

In terms of a semantic model for documents and collections, we first describe the issues that were identified when implementing the Greenstone alerting service. Then we explain how these issues are addressed by the FRBR ontology support that has been introduced into the Greenstone Librarian Interface.

### 3.1 Greenstone Alerting Service

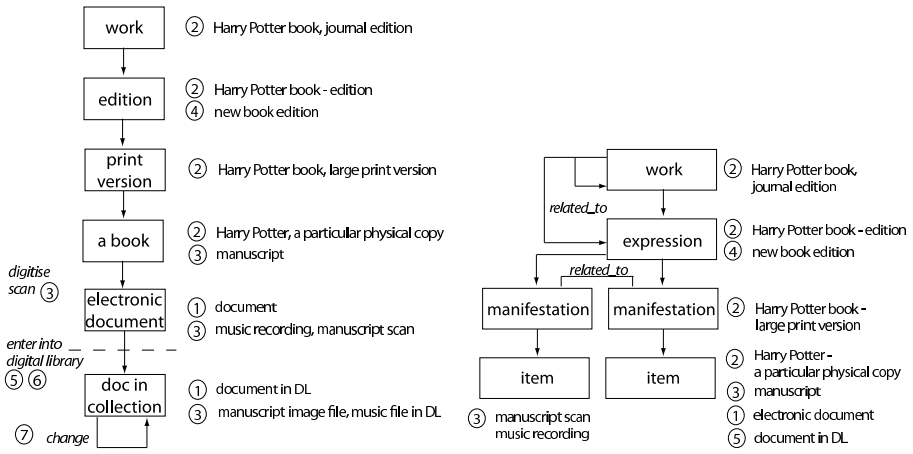
Users of digital libraries often have ongoing information needs: a doctor may need to track new publications in their field of speciality; an academic may wish to identify new articles on an individual whose biography they are writing. In either case, "stored" searches, using filtering or event-based technologies, are a critical addition to the digital library tool-set.

The Greenstone Alerting Service (Greenstone-AS) is a generic event-based alerting system that is provided as an optional extra to the Greenstone toolkit package [5, 8]. The system implements alerting across several libraries, including Fedora [9] and DSpace [11]. It supports the creation of profiles that represent a user's information needs. These profiles are subsequently used to match the user's needs against new or modified documents that are added to any DL server across a federated and/or distributed network of heterogeneous DL servers. In other words, in a network of DL servers the user can keep track of any document changes on any server. Furthermore, Greenstone-AS supports the notification of other types of events—such as the creation of a new classification within a topic heading, or the addition of an entirely new collection of material on an existing server.

The following examples are typical tasks that a user may be interested in. They can be defined as profiles that can be registered with the alerting service.

1. A new *electronic document* has been made available, for example, an electronic document about climate change. A user wants to be notified once the document is entered into a DL collection and available via the digital library.
2. A new physical work has been published, e.g., the latest Harry Potter book. A related task is a user keeping track of new issues of a journal (again in physical form).
3. An old manuscript is newly digitised by scanning, in a higher resolution, or in a different format. A music recording has been re-sampled.
4. Another edition of the same book is published.
5. An electronic document has been *newly published in the digital library.* The document is not new but has been newly added to the collection.
6. A document is deleted from the digital library. This event may be of interest, for instance, to a professor who wants to keep track of the DL documents available to students.
7. An electronic document has changed: for example, online software documentation may be continually written and adapted. Similar properties hold for blogs and wikis.

These different user profiles have implications for the semantic model used by the digital library. For illustration, we show the semantic hierarchy of our example items as perceived for alerting in Figure 3(a). We see the view of a book (in the example, a volume of "Harry Potter") from the perspective of an alerting system. The circled numbers correspond with the alert types listed immediately above. On the right, in Figure 3(b), is a depiction of a simple FRBR




(a) Alerting profile view (b) FRBR view

**Fig. 3.** Semantic hierarchies of documents and DL items. Numbers refer to the enumerations of profile tasks on the previous page.

data model of the same needs. The figures are compared in detail in Section 3.2 below.

Different media types (e.g., text, music, film, maps) require and allow different tasks to be carried out (e.g. scan, sample, change after digitising) and allow for different types of profiles (see above). We need to distinguish actions within the digital library and actions outside; both types may need to be captured. For example, the concept of a new document may carry the following semantics: new 'real world' document, scanning (i.e., making an electronic document), adding the document to the library.

There are several semantic consequences from this. Across the DL network involved, all items (works, editions, documents, scans) must have a consistent identifier over time. A new version of an item may be assigned a new identifier while the old one retains its identifier. Alternatively, the new document may take the identify of the previous version, while the old document is archived (and obtains a new identifier) or is deleted. In addition, both the new and old version could be assigned new identifiers. Semantically, the abstraction of the work and edition from a document or recording is crucial to the successful execution of alerts.

### 3.2 FRBR in Greenstone

The recommendation on Functional Requirements for Bibliographic Records [10] (FRBR) is an important and relatively recent recommendation for the enriched description of creative works in digital indexes. FRBR can be used to improve the features and functionality of the reader's experience of using a digital library [3].

As an example of a simple ontology developed by the Library of Congress, FRBR is highly significant to the DL research community.

The FRBR model is based upon four entities: *works, expressions, manifestations* and *items*. A work is a unique creative product (e.g. James Joyce's "Ulysses" or the latest Harry Potter book), available in one or more *expressions* (commonly termed editions or versions). Each expression has a particular selection of content, and may be produced in several different *manifestations*. For example, a book may be printed in a number of different bindings (paperback, hardback), or electronically (in PDF). An *item* is a single copy of a particular manifestation—the file on your computer, or the volume on the library shelf. This simple framework creates a tree for each work.

Beyond the core entities, there are others for expressing the identities of people and organisations concerned with the creation of the work or expression, such as authors, performers and publishers. Similarly, the subject of books can be encoded, including again people and organisations, but also events, places, and so forth. Works can also be related. For example, "West Side Story" can be encoded to identify that it is derived from William Shakespeare's "Romeo and Juliet".

Figure 3(b) illustrates some potential examples. In the case of a music manuscript, both the physical print and the electronic scan reflect the same *expression* of the same work. However, the printed version and the related scan are best modelled as different *manifestations* of this expression. Both the printed and scanned manifestations may have a number of different actual instances, or *items*. In contrast, the user's view of a scanned copy, when defining an alert, is probably different. Looking left to Figure 3(a), the scanned copy is seen as a derivative of the physical (printed) copy it was taken from. Whilst in both cases the logical modelling could be altered, or represented in a different manner, digitisation often throws up such possibilities for different outcomes to arbitrary decisions, and a good DL system should include the ability to resolve these – particularly when including material hosted elsewhere. Referring to our example alerting tasks, we observe that the manuscript as well as the scanned manuscript are items to different manifestation that refer to each other. Note that tasks 6 and 7 (actions within the DL) are not represented in FRBR. Similarly, the relationship between a manuscript and its electronic representation (see alerting task 3) can be made explicit by relating the two manifestations (see Figure 3(b)).

The initially simple framework of FRBR provides powerful tools for resolving some common user requirements. As we noted in Section 3.1, users often want to track accessions to a library, such as the arrival of a new issue of an important journal. Whilst traditional metadata can track some relatively simple requirements, complexities in the wider world or in the user's needs often mean that metadata-only methods lack critical levels of precision. To take some simple examples, if a journal that is being tracked for new issues changes its title, alerting may fail; authors with common names may be difficult to disambiguate; translations of a work may have entirely different titles. These are just

some cases where only a richer semantic approach can hope to support the user's actual information goals.

**Collaboration** Users of digital libraries often use several different digital libraries to fulfil their information needs. Developing the dataset to support a rich information environment such as a sizable FRBR repository may similarly require the inclusion of material from several different sources. As any given *work* may appear in dozens or hundreds of individual *items*, FRBR is quite capable of including the content of several libraries. Such an approach also offers the opportunity to support the discovery of content from traditional DLs that are built on regular metadata. Thus, much of the semantic-technology benefits of FRBR can be extended to older DL architectures as well.

**Greenstone FRBR implementation** FRBR support in Greenstone includes several different facets: the encoding of FRBR information using the Greenstone Librarian Interface (cf. Figure 2); the discovery of material from metadata-based DLs; the support of pure "FRBR" retrieval; novel user interactions and improved alerting. Space precludes a full discussion of these different aspects, so here we will focus on GLI's support for ingest, searching metadata-based DLs through FRBR, and alerting. For each aspect, we first discuss its semantic challenges and then discuss the details of Greenstone's approach.

*1 FRBR Ingest*

The use of any ontology or semantic model, such as FRBR, requires the population of the model with actual data. Interactionally and architecturally, adding FRBR support to the GLI is not trivial as the application was originally developed to support a metadata-driven build process, albeit with a configurable workflow and the ability to handle any chosen metadata standard (e.g. MARC, Dublin Core). However, most metadata schemes have little or no hierarchical aspects, and none take the object-oriented, ontologial approach of FRBR.

The differences are easily highlighted when one compares the ingest of a new document. One key step is the addition of author metadata to the document. Using Dublin Core in a traditional DL architecture, one would simply add data to the DC.creator field. However, in FRBR one creates a relationship between a *work* and a *creator*. The creator will often be a specific person, who is represented by a specific object in the FRBR repository. Thus, what in one approach is represented by a metadata field is in the other represented by an object-to-object relationship. The former simply requires text input, whilst the latter requires a query to be executed against the FRBR repository, and perhaps a new object to be created.

These differences have clear impacts on the Greenstone 3 system and interaction design. For example, an author query will result in the system selecting one or more matching author objects and their corresponding documents, whilst in Greenstone 2 a simple search against the metadata of all documents is sufficient. In interaction terms, a traditional DL architecture has no system concept of the

author as an object, the author cannot be represented in the interface directly. In contrast, Greenstone with FRBR explicitly represents the author in a data object. Consequently, a specific page can be created for each author with a list of their works and biographical details. An analogue of this could be created in a traditional architecture but the biographical details etc. must be encoded in a document in the collection and similarly named authors would have to be distinguished by careful manual creation of the author document. Worse, author biographies for collection access would now now mixed with original material in a collection.

FRBR is only one example of ontological support. Other schemes could provide further or complementary advantages. Ontologies can also support traditional metadata libraries: for example, ingest of new documents can be assisted by extracting relevant data from an ontology. Once a particular *expression* or *manifestation* of a FRBR work is associated with a new digital document, FRBR data can be ingested into a metadata-based library through the simple step of outputting the FRBR data in a compatible metadata format such as MARC. Further, library specific, data can subsequently be added, such as subject classification and accession date. The Greenstone-FRBR module provides such a facility for Greenstone 2.

*2 FRBR Retrieval*

Interactive retrieval lies at the heart of any digital library system. Traditional metadata-based digital libraries, such as Greenstone 2 and DSpace, provide document retrieval through the document metadata and/or full text. As in the case of FRBR ingest in Greenstone, however, this simple mechanism alters. Whilst the syntax can appear to be similar—e.g. retrieval of a document by author name—the corresponding operations underneath have changed. Furthermore, the range of possible retrieval operations expands beyond what is simply expressed by traditional search.

A semantic approach also simplifies the index encoding issues for more complex retrieval tasks. A search for books by "Winston Churchill" may seem straightforward. However, there were in fact two well-known Winston Churchills who were nearly contemporaneous: Sir Winston Leonard Spencer-Churchill, the wartime Prime Minister of the United Kingdom, and American novelist Winston Churchill. In fact, there are a number of other Winston Churchills of other dates who have written well-known texts. Using document metadata alone, all that can be done is to maximise the data encoded about any individual document—giving as much author information as possible. However, doing so consistently is problematic, and identifying possible confusions fraught with potential misunderstandings or simple lack of knowledge. Worse, some authors write under more than one name (Agatha Christie also published as Mary Westmacott, for example) or with various spellings of their name (a common problem with Russian writers translated into English). Using a document-by-document approach produces many problems of scale (multiple inputting) and validity (e.g. ensuring consistent encoding of all documents by an author). A structured, semantic approach yields immediate benefits.

However, how can this be achieved when a considerable investment has already been made in an existing metadata-based DL, such as Greenstone 2? Using a separate FRBR database, documents in Greenstone 2 are stored as particular FRBR *items* in the FRBR database, connected to Greenstone 2 by their unique document identifiers. Subsequently, the FRBR database can be used as a supplementary tool to disambiguate metadata through the FRBR system, rather than endeavouring to use complex encoding in the metadata features of the DL. With Greenstone an installation can be customised using its open architecture, to provide this functionality directly within the running DL system. Alternative or additional further ontologies can be supported in by exactly the same method: for instance, a bespoke ontology in Greenstone has been created for a commercial collection of 18th century literature.

*3 Alerting*

In Section 3.1 we looked at a number of the difficulties that can emerge when a user attempts to define an ongoing alert requirement. As with many metadata-centred methods, particular requirements can be difficult to express precisely in the original Greenstone-AS alerting system. The use of FRBR or an equivalent ontological method provides one technique for increasing precision or recall for many alerting tasks. Simple alerts directly correspond to the retrieval requirements identified above, but are triggered whenever the Greenstone ingest process performs changes on a collection. Though additional data is required to distinguish current and prior state of the DL to establish when a change occurs, these simply provide a second filter to reduce the final result set. Greenstone with FRBR covers all of the alerting tasks detailed in Section 3.1.

However, the vocabulary and language of potential alert requirements is vast. Alerts are not necessarily only the product of readers' information needs about documents. Alerts could correspond with subject classifications, and may be the product of managing the library content, rather than use of the library content. In other words, *librarians* themselves have information needs that do not correspond with traditional document retrieval. For example, a librarian may wish to know when a subject classification exceeds a particular size, suggesting that it may require new sub-classifications to be added.

Any one ontological model is unlikely to cover all possible requirements, and FRBR contains no direct representation of subject hierarchies, though particular information on the subject of individual documents (e.g. of people, places or topic) is supported. Just as FRBR can provide a method for better modelling the connection of documents, authors and publishers, classification hierarchies in turn can better model structured maps of document topics.

**Semantic challenges** In developing FRBR support and alerting in Greenstone, further semantic challenges have emerged. Some reflect significant issues in the modelling of library content. One such challenge is the question of *aggregate works*. Aggregates are publications—such as an anthology of poetry—that contain a number of separate *works*. Researchers from library science argue for

different approaches to modelling aggregates in FRBR. One such approach is to identify an aggregate as a particular *manifestation* of many different works. In consequence, the aggregate can only then be identified as a relationship between the many manifestations of different works that constitute it. In other words, as a distinct entity, it becomes implicit. There are serious shortcomings to this and other approaches that attempt to model aggregates within the existing FRBR model. Our approach was to model aggregates as a separate entity type, and support for this new type was added to Greenstone [4].

## 4 Semantics for Collaboration

The collaboration of Greenstone with the TIP service for location-based access identifies further semantic challenges. We describe the Greenstone 3 architecture and explain how its open framework addresses a variety of these challenges.

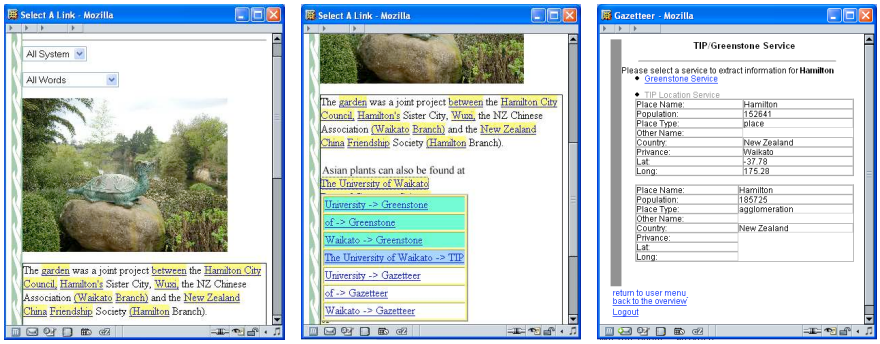
### 4.1 Location-based access to Greenstone

The TIP/Greenstone Bridge [6] provides location-based access to documents in a digital library. TIP is a mobile tourist information system that gives context-aware access to information about sights that are in the vicinity of the user. In a typical interaction, a user starts from a TIP information page, e.g., about the University of Waikato, and decides to look up the digital library collections that refer to their current location. When they switch to the page from the TIP/Greenstone Bridge, the system will display nearby regions and places that they might want to search for in the collection repository provided by Greenstone. This reflects their location at the University of Waikato, in the city of Hamilton, in the Waikato region, on the North Island of New Zealand etc. All these locations could be used to search the library; the user can guide the selection.

Based on the user's selection, the system triggers a location-based search of the DL collections. The user is presented with a list of all collections that refer to the selected region. After selecting a collection (e.g., Hamilton Gardens) and a document they are interested in (e.g., a description of the Chinese Garden), the user is presented with the digital library document with the place name highlighted. These names serve as anchor points that can link to further documents within the Greenstone collection, to a gazetteer of placements, and to TIP pages. An overview of the interactions is given in Figure 4.

For location-based access, documents must be pre-processed in order to identify any locations that they mention. Our current implementation of the TIP/Greenstone bridge uses a gazetteer to identify place names annotated by country. A simple location-aware mark-up of the documents is used in the software version described in [6]; a design using more complex access via spatial indexes is described in [7].

In either case, the challenge lies in identifying location or place names. For example, a document may be about Hamilton New Zealand, or Hamilton Canada.




(a) highlighted text (context world) (b) back link pop-up (context world) (c) gazetteer (context New Zealand)

**Fig. 4.** Overview of example interaction with TIP/Greenstone [6]

The term Hamilton may also refer to Captain Hamilton, after whom the city was named, or to any other person with the first name or surname Hamilton. Similar challenges apply for other contextual semantic information, such as the distinction between documents *by* Shakespeare and documents *about* Shakespeare.

Different systems collaborating with the digital library may have different notions of semantic information. For example, the digital library bridge combines Greenstone, TIP, and the Gazetteer. Collaboration requires the semantics of place names to be made explicit: correct hyper-links to related pages within the three systems (as shown in Figure 4(b)) are only possible if the respective concepts are aligned. In addition, the semantic context needs to be considered: in Figures 4(a) and 4(b), all place names are identified; whereas in Figures 4(c), only place names in New Zealand are taken into consideration.

### 4.2 Open framework in Greenstone

In the earlier version of Greenstone, communication between the underlying services and the user interface was provided through the Greenstone protocol [1]. This was a closed protocol, in common with other contemporary DL protocols such as SDLIP and Dienst. In Greenstone 3, each service (module) can provide its own interface, and the operational DL system can be composed from the desired mix of individual services and modules. To support simple adoption of the standard features, an analog to the earlier Greenstone protocol is provided. However, this is no longer the only method of providing communication, and it is this very mechanism that was used to implement the TIP/Greenstone bridge.

Suleman [12] has produced a powerful argument for modularising digital library systems. Greenstone 3 follows this philosophy (which is already present in many areas of Greenstone 2) throughout its entire architecture, from the ingest of new material to runtime services. This openness, which naturally embraces

semantic technologies, also brings pointed challenges. For instance, co-ordinating multiple services at runtime becomes difficult when new services can readily be added or existing ones removed.

Certain elements must remain constant to provide sufficient rigour and regularity for implementation to remain reliable and efficient. The protocols of Greenstone 2, Dienst and SDLIP share common features that mean that coupling a client of one protocol with a server of another is relatively straightforward [1]. These similarities apply at both document and service levels: all protocols use unique document identifiers, and all protocols support query-based retrieval.

## 5 Discussion

Ontologies and semantic models provide more powerful levers for information retrieval than do classical metadata-based digital libraries. The simple metadata structures such as Dublin Core (or MARC) used by standard digital library systems provide only limited means of expressing particular information needs. Greenstone 3 allows for the creation of a simple metadata-based digital library collection. It supports the use of rich semantic data and corresponding services that transcend traditional metadata. Thought this semantic framework complex ontological methods can be built to annotate documents, classifiers and the collections that contain them.

As discussed in this chapter, documents in a digital library can play a critical role in the retrieval process. The semantics that occur *within* documents can be expressed simply as flat text, but at the cost of precision during retrieval. Just as FRBR can disambiguate between different authors of the same name, detailed mark-up and semantic modelling of document content can distinguish places of the same name in a text, or between a place name and a personal name.

Supporting a richer range of retrieval methods requires corresponding runtime services and ingest time indexation. Greenstone has responded to both these challenges by extending the use of open standards within its architecture and implementation. At ingest time, the system provides the opportunity to tailor the accession process, including the use of additional indexers, applying metadata and content validation and running data extraction processes such as summarisation. At runtime, the system retains the provision of a simple Greenstone 2-like protocol, but in addition each module and service can define its own messaging options, and the entire communication between user interface and runtime services can be fully componentised.

The application of semantics to digital library systems is not cost free. Considerable time investment is required to achieve richer mark-up, and accession time costs on metadata creation are already high. Such costs also apply to the installation and configuration of a digital library system: whilst the configuration of Greenstone 2 for a collection of Dublin-Core encoded XML documents for later retrieval by metadata and full text is straightforward, the task of preparing a Greenstone 3 collection for retrieval using FRBR for cross-document retrieval and internal semantic modelling within documents requires an entirely

different level of commitment. The process of configuring an ontology-enabled Greenstone 3 collection is more complex, requires more effort, and is resource intensive. Providing API-style access at runtime to the first system is also simple, whilst the latter is open to much greater variation between installations. Currently, such complexities directly challenge the widespread adoption of the richer retrieval technologies that semantic DLs can provide. A considerable body of further research is required to lower these costs, both at installation and runtime.

## References


1. D. Bainbridge, G. Buchanan, J. McPherson, S. Jones, A. Mahoui, and I. Witten. Greenstone: A platform for distributed digital library applications. In *European Conference of Digital Libraries*, pages 137–148, Darmstadt, Germany, 2001.
2. D. Bainbridge, K. J. Don, G. R. Buchanan, I. H. Witten, S. Jones, M. Jones, and M. I. Barr. Dynamic digital library construction and configuration. In *Research and Advanced Technology for Digital Libraries*, pages 1–13, 2004.
3. G. Buchanan. Frbr: enriching and integrating digital libraries. In *JCDL '06: Procs. 6th ACM/IEEE-CS joint conference on Digital libraries*, pages 260–269, New York, NY, USA, 2006. ACM Press.
4. G. Buchanan, J. Gow, A. Blandford, J. Rimmer, and C. Warwick. Representing aggregate works in the digital library. In *Procs. ACM/IEEE JCDL*, pages 247–256, 2007.
5. G. Buchanan and A. Hinze. A generic alerting service for digital libraries. In *Proceedings of the JCDL*, June 2005.
6. A. Hinze, X. Gao, and D. Bainbridge. The tip/greenstone bridge: A service for mobile location-based access to digital libraries. In *ECDL*, pages 99–110, 2006.
7. A. Hinze and W. Osborn. Location-based indexing for mobile context-aware access to a digital library. Technical report, University of Waikato, Hamilton, New Zealand, August 2007.
8. A. Hinze, A. Schweer, and G. Buchanan. An integrated alerting service for open digital libraries: Design and implementation. In *Procs. 13th Int. Conf. on Cooperative Information Systems (CoopIS 2005)*, volume 3760 of *Lecture Notes in Computer Science*, pages 484–501. Springer, 2005.
9. C. Lagoze, S. Payette, E. Shin, and C. Wilper. Fedora: An architecture for complex objects and their relationships. 2005.
10. S. G. on the Functional Requirements for Bibliographic Records. *Functional requirements for bibliographic records*. K.G. Saur, 1998.
11. M. Smith, M. Bass, G. McClella, R. Tansley, M. Barton, M. Branschofsky, D. Stuve, and J. Wakler. DSpace: An open source dynamic digital repository. *D-Lib Magazine*, 9(1), 2003.
12. H. Suleman and E. A. Fox. Designing protocols in support of digital library componentization. *Proc. European Conference on Digital Libraries*, 2458:568–582, 2002.
13. I. H. Witten and D. Bainbridge. *How to Build a Digital Library*. Elsevier Science Inc., 2002.
14. I. H. Witten and D. Bainbridge. A retrospective look at greenstone: lessons from the first decade. In *JCDL '07: Proceedings of the 2007 conference on Digital libraries*, pages 147–156, New York, NY, USA, 2007. ACM Press.