Bonolo: A General Digital Library System for File-based Collections

Lighton Phiri, Kyle Williams, Miles Robinson,
Stuart Hammar, and Hussein Suleman

Department of Computer Science,
University of Cape Town,
Private Bag X3, Rondebosch 7701, South Africa
lphiri,kwilliams,mrobinson,shammar,hussein@cs.uct.ac.za

Abstract. There is an ever-increasing amount of digital content being generated that needs to be well-organised, preserved and made accessible. The majority of generic repository software tools that currently exist are, arguably, overly complex, thus making collections difficult to manage and maintain in resource constrained environments. A possible solution to this problem would, in part, require designing digital library tools and services that are simple and easy to manage. This paper describes a digital library system that is based on a set of design decisions aimed at simplifying repository software architectures. The proposed system makes use of a hierarchical file-store for storage of digital objects. Evaluation of the system by means of a user experience study was conducted to investigate the usefulness of the system, its relative ease of use and what effect, if any, the architecture would have on the user experience. Experimental results showed that users found the system useful, effective and easy to use and that the architecture did not appear to negatively influence the user experience.

Key words: Bonolo, Digital Collections, Digital Libraries, Digital Preservation, Developing World

1 Introduction

There have been significant technological advancement focused on developing tools and services to help preserve the abundant amount of digital content being generated and eventually make it more accessible and distributable. However, there are a number of challenges that are specific to developing countries that make it difficult to make use of these solutions. For the most part, the majority of digital collections are being set up by organisations that have limited funding, making it difficult for them to purchase the necessary hardware infrastructure and to hire technical experts to set up and manage such systems. In addition, the limited availability, and in many cases unavailability, of Internet bandwidth hinders access to collections that are set up using existing tools. Furthermore, Internet bandwidth is expensive, making it effectively difficult to use some conventional tools, such as cloud-based solutions. Previous studies have highlighted these issues and presented techniques and solutions to begin addressing them (Suleman, 2008). Earlier attempts at designing tools and services, specifically tailored for environments with resource limitations, resulted in a specialised system for managing historical Bushman manuscripts (Suleman, 2007) and a hybrid solution for managing online-offline digital collections (Suleman et al., 2010). The design of these tools and services was largely influenced by explicit user requirements and the environments in which they would be used and, as a result, they are not general-purpose tools. In this paper, a general solution is proposed to address the issues that arise in creating and managing digital collections in resource-constrained environments. The proposed solution takes the form of a simple and lightweight XML-centric digital library system that is bandwidth friendly and minimalistic, effectively making it easy to manage. The system, called Bonolo[1], makes use of a file-store repository and provides a curator and end user interface. The solution draws inspiration from Project Gutenburg's philosophical premise (Hart, 1992) of making information, books and other materials available in the simplest, easiest to use forms.

The remainder of this paper is structured as follows. Section 2 begins with a review of related work. Section 3 presents Bonolo, a simple and minimalistic digital library tool, as well as the design principles on which it is based. Section 4 presents the evaluation of Bonolo and, lastly, conclusions and plans for future work are presented in Section 5.

2 Related Work

Various solutions have been devised to facilitate ubiquitous access to knowledge by overcoming problems experienced in resource constrained environments. The One Laptop Per Child (OLPC) project (Kraemer et al., 2009) is one such example, as it runs an operating systems that can easily be deployed and updated; and is generally easy to operate. In addition, it is designed to easily synchronise with other machines in the same peer network, making it possible for content to be shared locally.

A number of software systems for building and managing digital collections are available. Greenstone (Witten et al., 2001) arguably stands out, due to its ability to publish and distribute collections on a self-installing CD-ROM. However, Greenstone requires installation of third-party components, and most importantly, stores metadata in a proprietary Greenstone Archive Format; this could pose some challenges in migrating digital objects to a different platform. EPrints (Gutteridge, 2002) is another popular tool of choice, and is commonly used for building institutional repositories. However, setting it up and subsequent maintenance and management tasks require technical expertise. Omeka

[1] Sesotho word for simple

(Kucsma et al., 2010), an open-source Web-publishing platform, is another promising solution; and while designed with non-IT specialists in mind, it stores its metadata in a relational database management system, effectively complicating the migration of digital objects to different platforms.

Other studies include an XML-centric solution (Suleman, 2007) that stores metadata in flat files, as an alternative to storing the metadata in databases; and a hybrid online-offline solution (Suleman et al., 2010) that enables current content updates to be seamlessly integrated with existing larger offline collections, a potentially useful feature in bandwidth constrained environments.

In general, most of the existing conventional digital library systems are Web-based and designed to make use of platform-dependent components, such as relational database management systems as the underlying storage layer for metadata. The fundamental difference between these solutions and Bonolo is that the architecture of Bonolo is largely influenced by a set of design decisions aimed at building tools and services that can operate in environments with resource limitations.

Based on the shortcomings in existing solutions, the next section presents the design of Bonolo and the design principles that were derived in order to begin to overcome these shortcomings.

3 Design of Bonolo

The design goals of Bonolo are centered on the premise that a simple and minimalistic approach to designing tools and services reduces the complexity of tools, making them easy to use and maintain.

As part of this study, a set of design principles were derived to form a basis for the overall design of simple architectures. A Grounded Theory (Glaser and Strauss, 1967) approach was used by conducting a case study of existing software tools relevant to the study; the goal was to infer possible design decisions of the tools that would be suitable for simple architectures. These design principles are that:

- It should be possible to operate tools and services on a wide variety of hardware and software platforms.
- There should be explicit support for integration of any object type, metadata format or new service.
- The design of tools and services should take into account community-based standards and international standards.
- The design should be flexible enough to enable end users to adapt the tools and services to their own needs.
- To simplify the design, there should be minimal use of external software components.

– The preservation process should be simplified as much as possible to make it possible to easily migrate digital content.

– There should be explicit support for hierarchical organisation of information.

– There should be support for access to digital collections i n environmen ts with limited Internet connectivity, for instance, by aking collections distributable on external media.

Bonolo is a generic digital library system based on these design principles. It is composed of three high-level components: a repository, a curator interface and an end user interface, which are used to facilitate storage of digital objects, management of the repository and access by end users, respectively. The system is Web-based, was developed using the Java programming language and thus runs within a servlet engine. In this section, the components that make up Bonolo will be discussed.

3.1 Repository

The Bonolo repository architecture is based on a simple file-store. In this section, the data model and way in which data are stored in the repository is described.

```
archive/
├── notebooks
│   ├── wilhelmbleeknotebooks
│   │   ├── BC_151_A1_4_001
│   │   │   ├── A1_4_1_00001.JPG
│   │   │   ├── A1_4_1_00001.JPG.metadata
│   │   │   ├── :
│   │   │   │
│   │   ├── BC_151_A1_4_001.metadata
│   │   └── :
│   └── wilhelmbleeknotebooks.metadata
├── drawings
└── stories
```

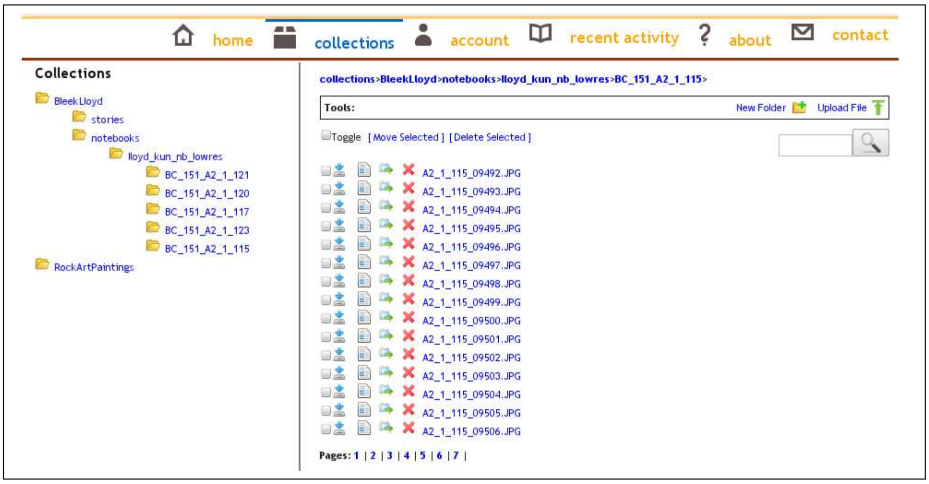
Fig. 1. Repository hierarchical structure

Data Model. Data in Bonolo is stored on the filesystem, in hierarchical directory structures called containers. The role of hierarchical structures is to help preserve the structure and semantics of the original data and to facilitate the discovery of information by end users. This data model has the added benefit of making it relatively easy to manage the repository through bulk actions that can be performed on digital objects contained within the same container.

Storage. An archive name, represented as a directory, defines the root container of the hierarchical structure formed by digital collections. A container structure, represented as a directory on the filesystem, has a corresponding metadata file that contains the title and description of the container and a manifest of the container's contents. The digital objects are stored on the filesystem with the metadata descriptors stored in plain text files alongside the objects that they describe using the same filename, but with a *.metadata* extension. The metadata files are encoded using qualified Dublin Core, but any other metadata schema could easily be used. Figure 1 shows part of a hierarchical structure of a sample collection integrated with Bonolo.

The design decision of storing metadata descriptors on the filesystem alongside the digital content makes it easier for the collection to be migrated to different hardware and software platforms, thus simplifying the preservation process.

3.2 Curator Interface

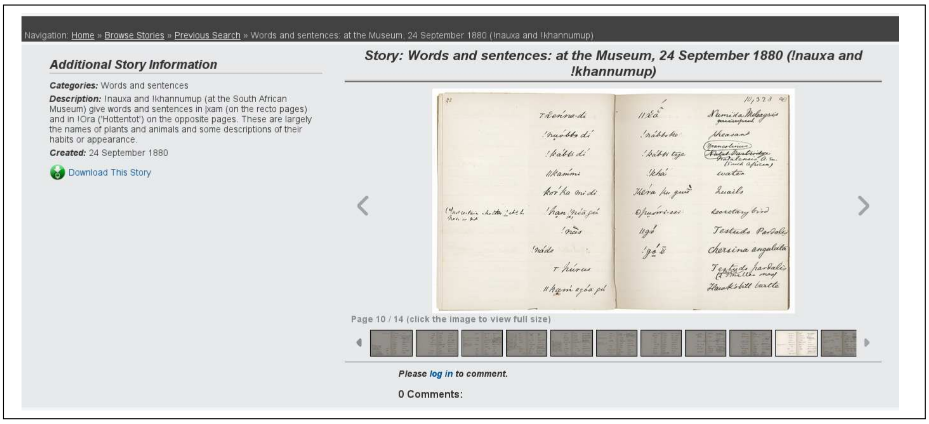


A2_1_115_09492.JPG
A2_1_115_09493.JPG
A2_1_115_09494.JPG
A2_1_115_09495.JPG
A2_1_115_09496.JPG
A2_1_115_09497.JPG
A2_1_115_09498.JPG
A2_1_115_09499.JPG
A2_1_115_09500.JPG
A2_1_115_09501.JPG
A2_1_115_09502.JPG
A2_1_115_09503.JPG
A2_1_115_09504.JPG
A2_1_115_09505.JPG
A2_1_115_09506.JPG

Pages: 1 | 2 | 3 | 4 | 5 | 6 | 7 |

Fig. 2. Curator interface

The Web-based curator Interface enables curators to effectively manage digital collections by rendering and presenting the hierarchical repository structure. The general interface, as shown in Figure 2, was intentionally designed to mimic the file manager interfaces that are used by most operating system so as to present the end users with a familiar view. Furthermore, this interface makes the discovery of digital content easy through browsing. Access to the curator interface is restricted through an access control mechanism that grants access to authenticated users. The access control is defined at the collection level, with users who are not the original owners of a particular collection being required to explicitly request access to it. The curator interface also provides the necessary functionality for the easy management of digital objects, such as: adding items; updating items; moving items; and deleting items. Furthermore, the interface also allows for these operations to be performed on multiple items simultaneously as a batch. The curator interface makes use of the Apache Solr search engine[2] to enable searching for digital objects in a collection.



Navigation: Home » Browse Stories » Previous Search » Words and sentences: at the Museum, 24 September 1880 (!nauxa and !khannumup)

Additional Story Information

Categories: Words and sentences

Description: !nauxa and !khannumup (at the South African Museum) give words and sentences in |xam (on the recto pages) and in !Ora ('Hottentot') on the opposite pages. These are largely the names of plants and animals and some descriptions of their habits or appearance.

Created: 24 September 1880

Download This Story

Story: Words and sentences: at the Museum, 24 September 1880 (!nauxa and !khannumup)

Page 10 / 14 (click the image to view full size)

Please log in to comment.

0 Comments:

Fig. 3. End user interface

3.3 End User Interface

The Web-based end user interface acts as an entry point for end users to explore and discover digital content stored in the repository. In essence, it provides core services for facilitating the discovery of digital content through search and browse features, as well as additional features that add value to the entire user experience (UX). The ability to search and browse is made possible by the Apache Solr search engine that is integrated with Bonolo. An end user is able to conduct a filtered search request by specifying search facets, so as to effectively find desired content. Figure 3 shows a snapshot of the end user interface.

[2] http://lucene.apache.org/solr/

4 Evaluation

Evaluation of Bonolo was done in two ways. Firstly, user experience (UX) studies were conducted for the curator interface and the end user interface. The purpose of these UX studies was to determine how the users felt about Bonolo, given its design considerations and to determine what effect, if any, the file-store repository had on the user experience. The second way in which Bonolo was evaluated was by means of a performance study in order to determine how the data store affected the performance of the system.

4.1 Curator Interface

The curator interface was evaluated using the Intrinsic Motivation Inventory (IMI) (*Intrinsic motivation inventory (IMI)*)(*Intrinsic motivation inventory*) with only three subscales: Interest/Enjoyment, Perceived Competence and Value/ Usefulness. The user study was aimed at evaluating: the usefulness of the tool;the experience of novice users; the ease that users had completing tasks; and the pleasure or displeasure users derived from interacting with the tool. Participants were recruited via social networking sites[3] and a total of 77 responded. Of the 77 respondents, 23 successfully completed all survey tasks. 78.26% of the participants had no prior experience working with digital library systems; however, 91.31% of the participants had experience working with online tools with content manipulation features similar to those of Bonolo.

The survey required participants to watch a 4 minute introductory screencast that showed them how to use the tool. They were then asked to create an archive, build two collections using different datasets and perform a series of manipulation tasks on the built collections. Once these tasks were completed, the participants were asked to complete a post-experiment questionnaire that was designed to elicit their overall experience using the Bonolo curator module.

The IMI related questions in the questionnaire were rated on a scale of 1 (Strongly Agree) to 5 (Strongly Disagree). The IMI subscale scores were computed by averaging all the items in the IMI subscales.

Figure 4 shows the average ratings for the IMI subscales by participants. As can be seen from the figure, the majority of users found Bonolo easy to use Bonolo and generally enjoyed using it. The results from the Value /Usefulness subscale are particularly interesting in that 8 of the participants strongly agreed that they found Bonolo very useful and none of them strongly disagreed.

[3] Facebook, Google Plus and Twitter

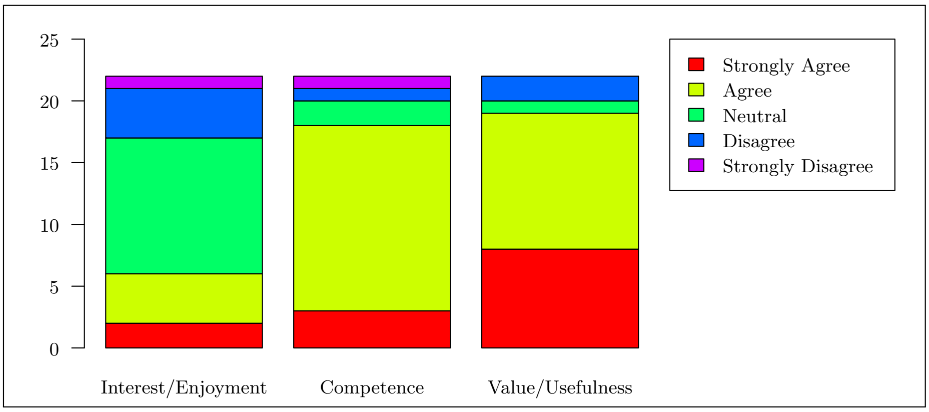


Fig. 4. Average UX ratings of the curator interface

4.2 End User Interface

A UX study was conducted in order to evaluate the effect that a hierarchical file store might have had on the overall end user interface and to determine if the services offered by the interface were comparable to those of other digital library systems. 17 users, the majority of whom were students and one who was a UX expert, took part in the evaluation. As part of the study, users were asked to perform several tasks related to the searching and browsing of content and navigating the interface. Users were then asked to rate the interface on a scale of 1-5 (in this case a high rating is positive) in terms of: intuitiveness; simplicity; satisfaction; expectations; responses; effectiveness and benefits. Users were then also asked to rate the level of similarity that the end user interface had with other systems that they were familiar with and to provide overall ratings on using the system. The average ratings of all users is summarised in Figure 5.

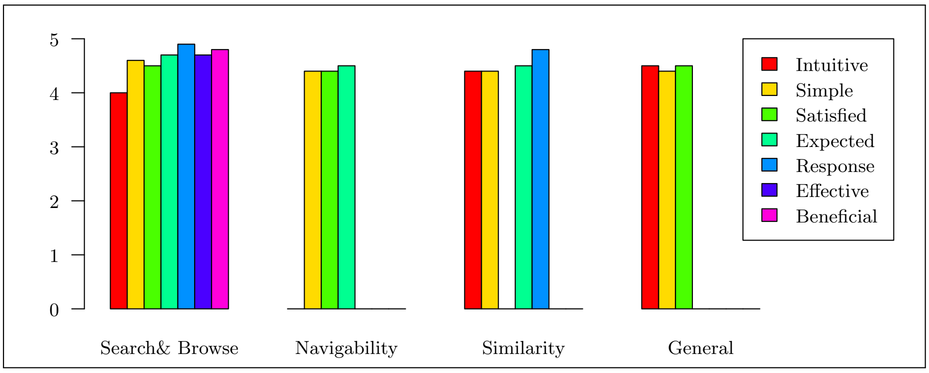


Fig. 5. Average UX ratings of the end user interface

As can be seen from Figure 5, users rated the interface highly from a UX perspective. Search and browse, two of the main digital library services, were rated as being intuitive, with an average rating of 4, while the interface responded as users expected, was effective and was simple. Furthermore, users were clearly satisfied with the navigability of the system, with an average satisfaction rating of 4.4. In terms of comparing the interface to other systems that users were familiar with, the end user interface was rated as being intuitive, simple and worked as expected. Lastly, the system was rated highly in general, with intuitiveness, simplicity and satisfaction ratings of around 4.5. Furthermore, the average of all ratings for all components was also 4.5.

This evaluation suggests that the use of a simple hierarchical file store did not have a negative impact on the user experience.

4.3 Performance

A performance evaluation was conducted in order to determine how well the system scaled when the amount of data was increased exponentially. Scalability was evaluated by considering the amount of time that it took to load the contents of a directory in the collection. In investigating this, two cases were considered: a case where a collection is carefully structured, such that each directory in the hierarchical file system contained exactly 256 digital objects; and a case where all the objects were in a single directory. Figure 6 shows the results of the performance evaluation.

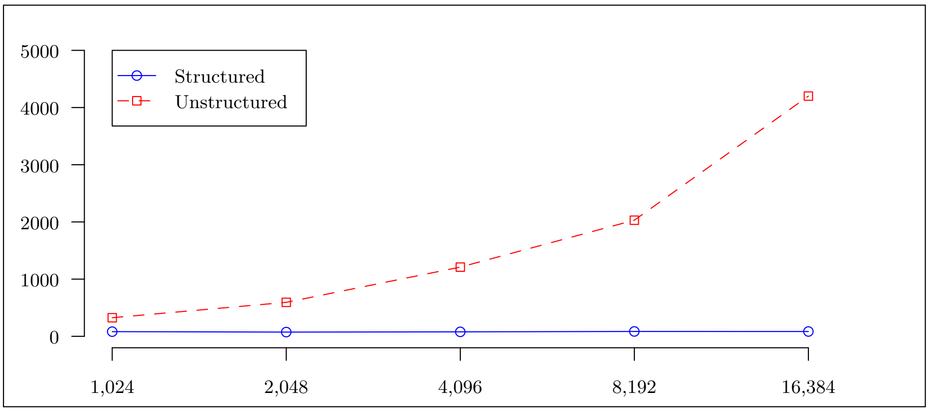


Fig. 6. Page loading times for structured and unstructured collections

As can be seen from Figure 6, when a collection is balanced, the page load times remain constant and it takes less than a second to load a page, regardless of collection size. However, when the collection is not balanced, the page load times scale linearly. These results suggest that, with structured data, the interface scales well. However, it also reveals one of the shortcomings in using a file-store that occurs when a collection is not balanced. A possible solution to addressing this could make use of a pre-processor in order to generate indices for directories in a collection that contains a number of files above some predefined threshold.

5 Conclusions

Current digital library systems are, arguably, not well situated to resource constrained environments. Thus, this paper proposes designing simple architectures for digital libraries that make use of file-store repositories and that have the potential to result in lightweight and portable tools and services. A digital library system called Bonolo was built based on this design and evaluated in terms of system performance and overall user experience.

Evaluation of the curator and end user interfaces by means of an user experience study showed that Bonolo was easy to use and generally useful, thereby suggesting that the file-store architecture is a viable one and does not have a negative effect on the user experience. Furthermore, the hierarchical structure could potentially have a positive impact on the ease of management of the repository, since most operating system file managers organise files in a similar manner and thus users would be familiar with the data representation. However, the performance evaluation showed the importance of a well balanced collection in order to ensure that the system scales well.

The design of simple digital library architectures has the potential of being useful in resource-constrained environments since: a file-based store has the potential to simplify the preservation process as simply copying the data is all that is necessary to migrate content; the digital objects are stored in a manner that users may already be familiar with; and resulting systems could potentially run off of a wide range of hardware and software platforms. In testing these ideas, ongoing work is focussing on the application of Bonolo and related systems for different collections in order to see how well the approaches generalise.

However, as future work, it would be interesting to evaluate the overall extensibility of such architectures by building services on top of Bonolo's core architecture. In addition, future work could also evaluate how well Bonolo performs for very large data sets.

References

[1] Barney G. Glaser and Anselm L. Strauss. *The Discovery of Grounded Theory: Strategies for Qualitative Research.* Chicago: Aldine Publishing, 1967.

[2] Christopher Gutteridge. “GNU EPrints 2 Overview”. In: *11th Panhellenic Academic Libraries Conference.* 2002. URL: http://eprints.soton.ac.uk/256840/.

[3] Michael Hart. *Project Gutenberg. The History and Philosophy of Project Gutenberg.* 1992. URL: http://www.gutenberg.org/wiki/Gutenberg:The_History_and_Philosophy_of_Project_Gutenberg_by_Michael_Hart.

[4] *Intrinsic motivation inventory.* URL: http://selfdeterminationtheory.org/questionnaires/10-questionnaires/50.

[5] *Intrinsic motivation inventory (IMI). All About UX.* URL: http://www.allaboutux.org/intrinsic-motivation-inventory-imi.

[6] Kenneth L. Kraemer, Jason Dedrick, and Prakul Sharma. “One Laptop Per Child: Vision vs. Reality”. In: *Communications of the ACM* 52.6 (June 2009), pp. 66–73. DOI: 10.1145/1516046.1516063.

[7] Jason Kucsma, Kevin Reiss, and Angela Sidman. “Using Omeka to Build Digital Collections: The METRO Case Study”. In: *D-Lib Magazine. The Magazine of Digital Library Research* 16.3/4 (Mar. 2010). DOI: 10.1045/march2010-kucsma.

[8] Hussein Suleman. “An African Perspective on Digital Preservation”. In: *Proceedings of the International Workshop on Digital Preservation of Heritage and Research Issues in Archiving and Retrieval.* Kolkata, India, 2008.

[9] Hussein Suleman. “Digital Libraries Without Databases: The Bleek and Lloyd Collection”. In: *Proceedings of the 11th European Conference on Research and Advanced Technology for Digital Libraries.* Ed. by László Kovács, Norbert Fuhr, and Carlo Meghini. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 392–403. DOI: 10.1007/978-3-540-74851-9_33.

[10] Hussein Suleman et al. “Hybrid Online-Offline Digital Collections”. In: *Proceedings of the 2010 Annual Research Conference of the South African Institute of Computer Scientists and Information Technologists.* Bela Bela, South Africa: ACM Press, 2010, pp. 421–425. DOI: 10.1145/1899503.1899558.

[11] Ian H. Witten, David Bainbridge, and Stefan J. Boddie. “Greenstone: open-source digital library software with end-user collection building”. In: *Online Information Review* 25.5 (2001), pp. 288–298. DOI: 10.1108/14684520110410490.