Building Digital Collections Using Open Source Digital Repository Software: A Comparative Study

George Pyrounakis, Department of Informatics and Telecommunications, University of Athens, Athens, Greece

Mara Nikolaidou, Department of Informatics and Telematics, Harokopio University of Athens, Athens, Greece

Michael Hatzopoulos, Department of Informatics and Telecommunications, University of Athens, Athens, Greece

ABSTRACT

The last decade a great number of digital library and digital repository systems have been developed and published as open-source software. The variety of available software systems is a factor of confusion when an organization is planning to build a repository infrastructure to host its collections. To simplify the decision process five widely used open-source repository software systems are compared, namely DSpace, Fedora, Greenstone, EPrints and Invenio. In addition to the comparison of these software systems and their characteristics’ description, the authors propose the most suitable systems for different cases of digital collections. Using five collection paradigms that represent case studies of different content and functionality, an organization can be directed to select a repository software matching its criteria.

Keywords: Digital Collection, Digital Library (DL), DSpace, EPrints, Fedora, Greenstone, Invenio, Open-Source Software, Repository

INTRODUCTION

The last decade a great number of Digital Library (DL) and Digital Repository (DR) systems have been developed and published as open-source software. The variety of available software systems becomes a headache when an organization plans to build a repository infrastructure to host its collections. Fortunately, there are many articles and surveys that evaluate or compare open-source DR and DL software. One of the first guides for selecting open-source repository software, based on the features and benefits of 9 different repositories, is provided by the Open Society Institute (2004). An extensive checklist for evaluating DL software is drafted by Goh et al. (2006). Also, two recent papers compare and evaluate some current open-source DR and DL software (Masrek & Hakimjavadi, 2012; Tramboo et al., 2012). The main scope of
these papers is the comparison of the software systems based on some quantitative and quality characteristics, in order for interested organizations to select the proper system for their digital collections. In our study we try to go a step further and in addition to the comparison of DR software systems and their characteristics' description, we propose the most suitable systems for different collection types. Using five collection paradigms that represent case studies of different content and functionality, an organization can be directed to select a repository software matching its criteria.

We used the following three restrictions in order to select the repository software systems that participate in the comparative study. The repository systems:

1. are publicly available using an open-source license,
2. are compliant with the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) (Lagoze & Sompel, 2001),
3. have a large number of installations worldwide.

Using these restrictions we selected the five (5) widely used repository systems shown in Table 1. These systems are between the 10 most used open-source software participating in the Directory of Open Access Repositories—OpenDOAR (n.d.). Each of these systems has been thoroughly studied based on the core repository characteristics and supported features. We used the latest software versions which are (at August 2013): DSpace 3.2, Fedora 3.6, Greenstone 3, EPrints 3.3 and Invenio 1.1. In the following section, the characteristics needed by a current repository software are listed and described. In the third section, the five repository systems are compared based on each of the characteristics and the results are summarized in a score table. In the fourth section, one or two repository systems are proposed for the hosting of each of five different collection types.

**REPOSITORY SOFTWARE CHARACTERISTICS**

In our approach the essential characteristics and features that are expected from a modern repository software are analyzed for each system. The following 14 characteristics are selected based on models for repository and DL systems, like the Reference Model for an Open Archival Information System (CCSDS, 2012) and DELOS DL Reference Model (Candela et al., 2007).

1. **Object Model:** The internal structure of the digital object, which according to Kahn and Wilensky (2006) is the entity that integrates metadata, digital content and relationships with other objects. Existence of unique identifiers for the digital object and every

---

Table 1. The five repository software systems selected for the comparative study

<table>
<thead>
<tr>
<th>Software</th>
<th>Initially developed by</th>
<th>License</th>
<th>Website</th>
</tr>
</thead>
<tbody>
<tr>
<td>DSpace</td>
<td>MIT Libraries and Hewlett-Packard Labs</td>
<td>BSD open source license</td>
<td><a href="http://www.dspace.org/">http://www.dspace.org/</a></td>
</tr>
<tr>
<td>Fedora Commons</td>
<td>Cornell University and the University of Virginia Library</td>
<td>Apache License, Version 2.0</td>
<td><a href="http://www.fedora-commons.org/">http://www.fedora-commons.org/</a></td>
</tr>
<tr>
<td>Greenstone</td>
<td>University of Waikato</td>
<td>GNU General Public License</td>
<td><a href="http://www.greenstone.org/">http://www.greenstone.org/</a></td>
</tr>
<tr>
<td>EPrints</td>
<td>University of Southampton</td>
<td>GNU General Public License</td>
<td><a href="http://www.eprints.org/">http://www.eprints.org/</a></td>
</tr>
<tr>
<td>Invenio</td>
<td>CERN Document Server Software Consortium</td>
<td>GNU General Public License</td>
<td><a href="http://invenio-software.org/">http://invenio-software.org/</a></td>
</tr>
</tbody>
</table>