Comparison of digital libraries systems

MICHAL KÖKÖRČENÝ, AGÁTA BODNÁROVÁ
Department of Informatics and Quantitative Methods
University of Hradec Králové, Faculty of Informatics and Management
Hradecká 1249/6, Hradec Králové
CZECH REPUBLIC
{michal.kokorceny,agata.bodnarova}@uhk.cz

Abstract:
In the last decades digital libraries play an important role in knowledge sharing. There are many digital library systems for building of digital collections and repositories. Comparison of digital libraries systems was widely discussed in many scientific papers. Nevertheless, these papers are usually aimed to description and comparison of features and capabilities of digital libraries. In this paper we will compare several systems for building of digital libraries from perspective of architecture of information systems. We will focus on comparison of these systems and theirs flexibility instead of comparing theirs features and functionalities. Suitable architecture of any information system (not only a digital library) is important aspect for all institutions and can reduce cost of system implementation, modification, integration and maintenance.

Key-Words:
digital library, Fedora, DSpace, Greenstone, EPrints, CDS Invenio, DILLEO, service oriented architecture

1 Introduction

Comparison of digital libraries systems was discussed for example in the paper [1]. There were selected five open source systems and were defined several features of any digital library [1]. Selected systems were compared based on the previously defined characteristics.
In this paper we will compare several systems for building of digital libraries from perspective of architecture of information systems. In general, digital libraries systems are – in point of view of architecture – a kind of information systems. We will focus on comparison of these systems and theirs flexibility instead of comparing theirs features and functionalities.

2 Problem Formulation

Digital libraries are controlled collections of digital objects, including services for storing, manipulation, accessing, searching etc. Therefore digital library systems provide relevant uses cases. These features and requirements have influence on architecture and design of whole information system. From the point of view of an institution managing collection(s) of digital objects is important to choose such system and architecture, which represents:

- low costs of implementation and maintenance of the digital library system,
- possibility to easily integrate the system into the IS/ICT environment of the institution,
- sufficient flexibility of the information system for future changes.

We have selected for comparison six widely used free digital library systems: Fedora, DSpace, Greenstone, EPrints, CDS Invenio and DILLEO.

3 Problem Solution

3.1 Fedora

Fedora (Flexible Extensible Digital Object Repository Architecture) was developed at Cornell University. It is architecture for storing, managing and accessing digital content in the form of digital objects [2]. Fedora is based on the Kahn-Wilensky Framework [3]. The Fedora Repository Project is an open source implementation of the Fedora architecture. Nevertheless, Fedora does not contain the user interface (presentation tier). Fedora is not a complex system, which is possible simply to install and to use. An institution should implement the user interface by other means.
The main difference between other systems is that Fedora is based on principles of service oriented architecture (SOA). Service oriented architecture is a concept for building and integration of information systems and applications [4]. Fedora provides repository service exposed as web services with well-defined application interfaces via REST or SOAP protocols [2]. The key feature of Fedora is that repository can store all types of digital content and its metadata [2]. Due to service oriented architecture Fedora provides high level of flexibility of content (not whole system).

3.2 DSpace

DSpace is open source software for building of digital libraries, primarily intended for academic and research institutions. The system was developed in cooperation MIT (Massachusetts Institute of Technology) and HP (Hewlett Packard) [5].

DSpace – in oposite to Fedora – is a complex software solution that includes repository and complete user interface. Therefore it is possible to deploy the software in a relatively short time period. DSpace uses as unique identifier Handle system (via URN), metadata are stored using Dublin Core standard [6], and metadata can be exported into METS format [7].

DSpace supports metadata searching as well as full text searching. Indexing of documents for full text searching is possible for these file formats:

- Plain text (TXT),
- Microsoft Word (DOC),
- Adobe Portable document format (PDF),
- Hyper text markup language (HTML).

All these file formats – except plain text – are for purposes of indexing converted using filters into plain text. Indexing of documents proceed always on plain text data. If the system is extended with new filters, it will be possible to automatically index other file formats and documents [8].

DSpace supports "collections" for storing of data objects. Each data object must be inserted into at least one collection (or more). Furthermore, these collections may be divided into "communities". Communities are organized into tree structure. That way DSpace provides means for dividing of digital sources into logical domains [8]. Due to these features DSpace differs from most of other digital library systems.

The biggest disadvantage of DSpace system is low flexibility of whole system. DSpace is a monolithic system, which does not allow modifying certain parts of software. DSpace is not based on principles of service oriented architecture (SOA).

DSpace also supports workflow processes for inserting new digital objects into repository. Newly inserted object is not published automatically and immediately – it must pass through an approval process, which contains a few predefined phases (metadata approval, content approval etc.). For each phase of the approval process is responsible certain (defined) user. An approval process is defined for a collection – not for whole repository. It means that each collection can have its own approval process [8].

3.3 Greenstone

Greenstone is a system for building and distributing digital library collections [9]. It is open source software released under the GNU General Public License. Greenstone has been developed at the University of Waikato in cooperation with UNESCO [9]. This system provides a way for organizing and publishing information on the internet as well as on removable media (e.g. DVD, CD etc.) – just this feature is not usual. Greenstone is the only system, which supports distributing collections via removable media.

Metadata are stored using Dublin Core standard and can be exported into METS format. Digital objects are organized into "collections", which contains relating objects. A library contains one or more collections.

3.4 EPrints

EPrints is open source software for building repositories of digital sources [10], primarily intended for scientific publications. EPrints has been developed at the University of Southampton.

For metadata the system uses its own internal format. In opposite to other digital library systems EPrints supports multiple "archives" under one instance of the software. An archive stands for standalone logical (not physical) digital library. So, you can have multiple digital libraries running on one instance of the software [8].

Unusual feature is that EPrints is based on statically generated pages of user interface (partially). This is just because of intention of the software – publishing of scientific sources. In this case is not supposed, that digital objects are inserted or updated very often [8]. When a digital object was inserted/updated in a repository, all changes are visible after regenerating of user interface (static pages). Therefore there is a delay between editing and publishing of a digital object.

3.5 CDS Invenio

CDS Invenio is a complex system for building and managing digital libraries. CDS Invenio is developed at the CERN and it is primarily intended for purposes of this institution. It is free software licensed under the GNU General Public Licence (GPL) [11].

CDS Invenio uses MARC [12] format for metadata and allows defining mapping between other metadata formats. Similar to EPrints, this system uses statically generated pages (user interface) [8].

CDS Invenio is a comprehensive solution; nevertheless for many functions it is necessary to install third party products. This system is not based on principles of service oriented architecture. There are strong dependencies on other products, which produces tight coupling relations [8]. Furthermore it brings problems with incompatibility between different versions of these products. CDS Invenio is relatively flexible product, but on the other hand, modification of the system is usually complicated and often very expensive. In opposite to other digital library systems CDS Invenio allows

changing of searching algorithm including results presentation.

3.6 DILLEO

DILLEO is a digital library of sharable teaching materials primarily for faculty and students of higher education [13]. DILLEO was developed at University of Hradec Králové in cooperation with other partners and universities under project E-DILEMA.

Digital objects have SCORM compliant metadata format. The information system is based on three-tier architecture with thin client as browser. The library is implemented on .NET platform, presentation tier is created in ASP.NET, data tier is realized on Microsoft SQL Server. Fulltext searching is implemented by means of Microsoft Index Server. Logic tier consists from set of application objects, which ensure processing of requests received from presentation tier [13]. This is typical architecture which produces tight coupled relations between application objects (components). Flexibility of such system is very low.

4 Conclusion

Most of presented systems are specific oriented (e.g. EPrints or CDS Invenio), whereas usually respect needs and requirements of institutions, where were created. There are a few digital library systems, which are complex; nevertheless the solution is not optimal and sufficient enough. Table 1 contains comparison of architecture and features of selected digital libraries systems.

Table 1 – Comparison of digital library systems

	Fedora	DSpace	Greenstone	EPrints	CDS Invenio	DILLEO
User interface		*	*	*	*	*
Dynamically generated pages	*	*	*			*
Customizable searching					*	
Customizable metadata formats	*				*	
Flexibility of the system	*				*	
SOA principles	*					
Process based approach		*				
Removable media (DVD, CD…)			*			

In general, flexibility of described digital library systems is very low. Only Fedora is based on SOA principles, CDS Invenio is a flexible system, nevertheless customization is relatively complicated. None of these systems (except DSpace) support process based approach, typically for the approval process. Very problematic is also customization of searching algorithm and modifying of used metadata format. Any customization and modification of these systems may be complicated and represents high costs for an institution.

4.1 Basic characteristics

Fedora – it is a flexible system based on principles of service oriented architecture. Nevertheless, flexibility is concerning only to content of the repository, Fedora is capable to store and access any digital content – this does not stand for flexibility of whole information system. Fedora does not contain user interface.

DSpace – is complex system that contains repository including user interface. Nevertheless, DSpace is not flexible system – it is not possible to customize certain parts and behavior of the software.

Greenstone – supports distributing collections via removable media. In this paper we discuss service oriented architecture, process based approach etc. – nevertheless contemporary technologies do not allow creating applications, which are based on these principles, which are multiplatform and which supports running from removable media.

EPrints – is primarily intended for scientific publications (digital objects are not modified very often). This system uses statically generated pages of user interface. EPrints supports multiple archives under one instance (logical libraries).

CDS Invenio – is a complex and flexible system, which uses third party products for certain use cases. Customization of the system is complicated and may be expensive. There are strong dependencies on other products, which can cause problems with incompatibility between different versions.

DILLEO – is primarily intended for educational purposes. It is typical three-tier architecture with thin client. The application consists from set of components producing tight coupled relations. DILLEO – similarly to the most of other digital library systems – does not support service oriented architecture and does not stand for flexible system.

4.2 Summary

Typical disadvantages of contemporary digital library system in opposite to current company information systems are:

- not enough flexible architecture,
- tight coupling of application components (instead of loose coupling architecture),
- higher costs of customization of a digital library system,
- digital libraries usually cannot share application components with other information systems,
- difficult integration of a digital library system with other information systems and applications in an institution,
- higher costs of this integration,
- usually there is not process based approach.

Contemporary digital libraries are very often monolithic systems/applications, which cannot fulfill new requirements on information systems.

4.3 Recommendations

New digital library systems and theirs architecture should follow ways and trends of development in the area of company information systems. Actually it comprises these concepts and principles:

- service oriented architecture (such as Fedora),
- process based approach (such as DSpace),
- automated processes including human workflow,
- using standard infrastructure and technologies (e.g. Enterprise Service Bus etc.).

In the area of digital libraries these approaches were not applied (except a few systems).

One of the many possibilities how to achieve high flexibility of the whole system is to implement a digital library as a composite (SCA) application. Service component architecture (SCA) is a relatively new concept or framework for creating applications (composites) built from services [14]. Services are basic elements of SCA composites which perform a specific function. Services may be implemented using different technologies and programming languages [14]. SOA/SCA based architecture provides loosely coupled suite of services which can reduce costs of system modification.

References:

[1] PYROUNAKIS, G., NIKOLAIDOU, M., Comparing Open Source Digital Library Software, *Handbook of Research on Digital Libraries: Design, Development, and Impact*, IGI Global, 2009.

[2] *Fedora, Flexible Extensible Digital Object and Repository Architecture,* http://www.fedora.info

[3] KAHN, R., WILLENSKY, R., *A Framework for Distributed Digital Object Services,* http://www.cnri.reston.va.us/home/cstr/arch/k-w.html

[4] DAVIES, J., SCHOROW, D., RAY, S., RIEBER, D., *The Definitive Guide to SOA,* Apress, 2008.

[5] *DSpace, Digital Archive Project,* http://dspace.org

[6] *Dublin Core Metadata Initiative,* http://dublincore.org

[7] *Metadata Encoding & Transmission Standard,* http://www.loc.gov/standards/mods/

[8] KREJČÍŘ, V., Systémy pro tvorbu digitálních knihoven. *INFORUM 2006: 12. konference o profesionálních informačních zdrojích,* 2006.

[9] *Greenstone digital library software*, http://www.greenstone.org

[10] *Open Access and Institutional Repositories with EPrints*, http://www.eprints.org

[11] *CDS Invenio,* http://cdsware.cern.ch/invenio/index.html

[12] *MARC Standards,* http://www.loc.gov/marc

[13] MIKULECKÝ, S., *Digital library of learning objects,* University of Hradec Králové, 2005.

[14] MARINO, J., ROWLEY, M., *Understanding SCA (Service Component Architecture),* Addison-Wesley, 2010.