CERN Document Server: Document Management System for Grey Literature in Networked Environment

Martin Vesely, Thomas Baron, Jean-Yves Le Meur and Tibor Simko
CERN

Abstract:
In this paper we present a technology for networked information services, developed at the CERN Document Server (CDS) research group, called the CERN Document Server Software (CDSware).

Standardization of networked information services in the field of grey literature has recently become a subject of an intensive research in the digital library community. The current state-of-the-art in this area effectively allows to provide various networked information services, such as information brokering or other value-added services upon distributed or federated data. This refers specifically to a variety of newly developed frameworks, such as the Open Archives initiative Protocol for Metadata Harvesting (OAi-PMH).

The CDSware technology offers a comprehensive solution for a document management of a large grey literature document repository, compliant with a variety of networking standards essential for a wide deployment of networked information services.

1 Introduction

The research in the field of digital libraries has recently focused on information interoperability models, the integration of distributed and federated data and networked information management. One of the key features of the networked information is that data, information and knowledge can be gathered, processed, stored and maintained separately from the information services offered by information mediators or brokers. In the scope of scholarly communication the models of networked information involve predominantly distributed and federated data processing, built on top of various widely deployed internet technologies, transfer protocols, their extensions and finally also comprehensive technological frameworks such as the Web Services.

Within this perspective, the Open Archives Initiative developed a Protocol for Metadata Harvesting [OAi-PMH]. The OAi-PMH was set up in 1999 [Van de Sompel, 2000] in order to filter out information heterogeneities that prevented an efficient cooperation between various e-print archives and other grey literature repositories in the networked environment. The interoperability is achieved by sharing the metadata format schemata used by all parties involved. One of the important goals of this protocol is to allow federated archives to harvest references to relevant documents that can then be made available for the archive users.

The CERN Document Server (CDS) group has been active in research focusing on interoperability of digital document storage and retrieval systems, particularly promoting the WWW and related technologies in the digital library community, including the OAi-PMH. Within the last 5 years the digital library research at CERN focused on (i) linking strategies to digital library resources with emphasize on the scientific literature [Vigen, 1999], (ii) data integration from heterogeneous data sources [Vesely, 2000], (iii) specification of the protocol for metadata harvesting [Vesely, 2002] and (iv) automated indexing of scientific documents [Dallman, 1999] [Raez, 2002].

2 Grey literature management

Until recently, the management of grey literature collections and mediation of scientific information has been performed predominantly by specialized *disciplinary repositories* based mainly on a centralized model. One of the pioneering repositories of this type was the ArXiv.org