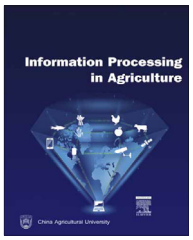
HOSTED BY

ELSEVIER

Available at www.sciencedirect.com

INFORMATION PROCESSING IN AGRICULTURE 2 (2015) 93–100

journal homepage: www.elsevier.com/locate/inpa



Thematic harvesting of agricultural resources from generic repositories

Devika P. Madalli *

Documentation Research and Training Center (DRTC) Indian Statistical Institute (ISI), 560059, India

ARTICLE INFO

Article history:
Received 2 January 2015
Received in revised form
11 April 2015
Accepted 6 May 2015
Available online 17 July 2015

Keywords:
Thematic harvesting
Agricultural data
Issues in harvesting

ABSTRACT

Metadata aggregators and service providers harvest entire collections or they restrict harvesting by date or sets. However most often user approach to collections is not by dates or set names but by domain based keywords. Harvesting by domains is an issue when service providers attempt to collect data from multiple sources. The main problem is that harvesters, at present, do not have the facility to distinguish themes such as domains. In the present work, an attempt has been through Tharvest, a thematic harvester model using the proposed methodology harvesting agricultural resources from generic repositories. Tharvest encompasses a process where technical terms of the domain of agriculture are taken from AGROVOC, a multilingual, structured controlled vocabulary designed to cover concepts and terminologies in the agriculture domain. AGROVOC is deployed to provide the basis for selective harvesting. The system components and workflows are presented and described. Metadata aggregators provide end-users a single platform discovery facility to resources collected from various data providers. It is observed that aggregators such as INDUS [www.drtc.isibang/ac.in/indus] dealing with agriculture and related domains facilitate aggregating metadata from not only repositories but also other sources such as journals and enable a centralized access to full text and objects. While harvesting can be fairly simple and straight forward, it is not without its challenges. This paper intends to highlight some of the issues in harvesting metadata in agricultural domain. The particular focus is to identify agriculture related metadata from generic sets.

1. Introduction

Digital repositories especially open access repositories, usually expose records through OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) and thus ensure that repositories of similar content may be interoperable. However, there are several challenges for service providers aggregating resources from various digital repositories.

Metadata harvesting mainly depends upon the use of standards by data providers. Yet for various reasons repository managers, while populating the collections, often do not follow global standards for metadata [1]. In some cases they adopt a standard such as Dublin Core[1] which involves a set of vocabularies for resource description or AGRIS,[2] a multilingual bibliographic database for agricultural science, but make some deviations to comply with local needs. In particular, problems arise when the vocabulary used in metadata by different repository maybe different. Repositories use non-standard terminology for names of elements. For

* Tel.: +91 80 28483002; fax: +91 80 28482711.
E-mail address: devika@drtc.isibang.ac.in

Peer review under the responsibility of China Agricultural University.
http://dx.doi.org/10.1016/j.inpa.2015.05.002

[1] http://dublincore.org/.
[2] http://agris.fao.org/agris-search/index.do.

instance – use of the term '*contributor*' instead of element '*author*', even when there is a provision for using '*collaborator*' where the '*contribution*' can be specified. It is due to such arbitrary variations that data cannot be harvested in a straight forward way [1]. The other issue in the content or value against the metadata element is in non-standard terminology. For example, rice 'varieties' referred as 'species of rice' or 'cultivars' of rice. This poses a problem for a person trying to access articles just by 'varieties'. The straight forward solution to the problem of synonymy is to follow a standard vocabulary in description. But often it is found that different vocabularies sets are used. It is even more complicated when arbitrary keywords are entered. Hence the approach here is to adapt AGROVOC as a standard reference for vocabulary in agriculture. In the following sections we describe Tharvest, a thematic harvesting facility for agricultural resources from generic repositories [2].

2. Harvesters

Harvesters facilitate centralized access, browsing and search facilities to resources that may be part of different collections. Harvesting is based upon the OAI-PMH standard and protocol for interoperability of repositories as prescribed in [3]. According the standard two important providers are *data providers* and *service providers*. Data providers are basically repository initiators/owners. They usually collect and organize resources in repositories. The requisite to be interoperable is that data providers must be compliant with the OAI-PMH standard. Service providers collect resources from such data providers in order to facilitate centralized access and search to resources exposed by the providers. They also facilitate access to full text or full resource wherever the data providers offer open access resources. Service providers can in turn be data providers and thus act as links to meta aggregators.

2.1. *General features of harvesters*

Harvesters have several features to facilitate identifying repositories and their collections and aggregating records. Usually the entire collection can be harvested. In case of very large collections data providers have the facility to restrict by number of records harvested. In such cases they provide a resumption token so that service providers can resume harvesting when they revisit the site.

Harvesting tools also provide facility for selective harvesting. One way to specify is by sets. Sets are normally collections that are specified by the repository. Repositories organize their resources in categorized collections and each such collection is uniquely identified as a set.

The other option provided is harvesting by date. This feature allows harvesting record from a certain date/ year etc. Other than these selective harvesting options, harvesters do not provide other means of filtering.

One of the popular harvester software is the Open Harvester System of Public Knowledge Project[3] (pkp). DSpace[4] is a popular Open Source software for hosting and managing repositories. It also facilitates harvesting metadata records. We describe here two services implemented at Indian Statistical Institute based on the above two harvesters. Indus is a DSpace based harvester that covers repositories in Asia collecting records in agricultural domain and Demeter is a pkp harvester based aggregator covering 22 data providers.

3. Indus

Indus[5] is an aggregator for agricultural information resources in Asia. At present it contains about 35,000 records from 8 Asian countries in 89 sets. It deploys the DSpace harvesting facility. Indus covers both repositories and journals in agriculture and related domains. For open access repositories full text is available and for some material with restricted access, resources for which metadata level access is provided, are also included (Fig. 1).

Collections in Agricultural domain are sourced from OpenDOAR, Grainger and Open Archives sites. For journals the main source is the pkp harvester official site, DOAJ site and also the OJS (Open Journal System) lists. In addition to these, resources are identified from generic repositories such as SodhGanga, where there are only a few agriculture related sets. We also conducted a Google based search to collect journals in agriculture and related domains that are not listed in OJS and DOAJ sites.

Three levels of harvesting are possible:

- harvesting metadata only
- harvesting metadata plus references to bitstreams
- harvesting metadata and bitstream (for this repository must be ORE complaint) (Fig. 2)

4. Demeter

pkp is a popular harvester used commonly to aggregate metadata from different sources. Demeter (at Indian Statistical Institute) is based on the PKP harvester. At present Demeter is implemented as a test bed. Preliminary study of performance of Demeter against Indus is ongoing. As per observations, there are some shared features and strengths in both the harvesters and some differences that we point out in the following sections (Fig. 3).

5. Comparison of pkp and DSpace based harvesting

At the Documentation Research and Training Center, Indian Statistical Institute, we have implemented both pkp (for Demeter service) and DSpace based harvesting (for Indus). Mainly they are implemented as service providers but they can both be data providers also. Though Dublin Core is used as default it is possible to crosswalk data between other metadata formats in both these systems (see Table 1).

[3] https://pkp.sfu.ca/ohs/.
[4] http://www.dspace.org/.
[5] http://drtc1.isibang.ac.in/indus/.