Thematic harvesting of agricultural resources from generic repositories

Devika P. Madalli *

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

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 straightforward, 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.

Peer review under the responsibility of China Agricultural University.

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 which involves a set of vocabularies for resource description or AGRIS, 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
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 straightforward 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 straightforward 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 Project3 (pkp). DSpace4 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

Indus5 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/.
Fig. 1 – INDUS: an aggregator for agriculture resources.

Fig. 2 – Harvesting modes offered by DSpace.
6. Thematic harvesting from generic repositories

In the process of developing Indus and Demeter, several sources were searched to identify agriculture and related domain resources. It was noted that there were general focus repositories within which some sets were related to agriculture. In some other cases such as in some ETD⁶ (Electronic Theses and Dissertations) repositories, the organization was not by themes or domain and disciplines but by dates. In such cases the challenge is that of harvesting thematically by key terms in order to comprehensively include agricultural resources. We observed that most of the repository and interoperability work has been undertaken among partners, institutes and projects that deal with agriculture domains. In such cases harvesting resources is straightforward. Some attempts have also been noticed in thematic harvesting but they

---

⁶ http://www.ncsu.edu/grad/etd/.

![Demeter collections](image-url)
involve manual verification that may be both time consuming and subject to arbitrariness owing to human intervention and decision making.

6.1 Harvesting Indus by sets

With present technologies for harvesting, it is possible to achieve thematic harvesting by ‘sets’ in OAI-PMH. For example for Indus, the command http://drtc.isibang.ac.in/indus/oai/request?verb=ListSets, lists 89 sets as shown below (Fig. 4):

6.2 Harvesting by date

Harvesting by date is relevant in cases where the aggregator is simply adding records by particular year or month or after a particular date. For example, in ETD we could source by the datestamp for each year the theses were awarded by a university. This may also be relevant to trace research output per year of an institute or university as represented by publication date in their institutional repository.

The above options provide a mechanism to harvest by date and by sets. However, an issue arises in thematic harvesting for agriculture records when the sets are not organized by domains and sub-domains. For example: When we consider a repository of ETDs [Electronic Theses and Dissertations], the sets may be organized by the year of award of the theses, or by the university that contributes the theses to a centralized repository such as a national repository. In such cases theses of different domains may be in the same set. Hence the present facilities provided by harvesters to retrieve records just by sets or by date are inadequate to harvest specifically by domain. There have been some attempts to harvest by domains as discussed in the section below.

7. Related work

Studying literature and Internet based reports, it is noted that there is very little work ongoing towards for domain specific harvesting or thematic harvesting in agriculture and related domains from generic repositories.

Most of the work in repository building is anchored in the agriculture domain itself. For instance Project Harvest that focuses on information resources in agricultural journals for development of the e-journal repository [4] [http://www.library.cornell.edu/harvest/proposal.htm]. Similarly EC project VOA3R deals with agriculture and aquaculture. Abian and Cano describe a system for managing learning resources in agriculture especially in domain of agro-ecology and organic agriculture [5]. It is straightforward to harvest records from such projects provided they are OAI compliant data providers. However, the question arises how we extract agriculture related records from generic repositories such as university repositories or institutional repositories which may not be organized by disciplines and topics of resources may be in diverse domains.

In an experience report by Simek et al. [6] a comparative analytical study of different agricultural repositories is presented where different workflows and metadata schema were followed. Projections are given regarding the compliance or deviation from the standard Dublin core elements so that based on common minimum shared elements harvesting can be achieved.

Metadata Interoperability issues often crop up when harvesting records into a central facility. Chan and Zeng [7] examine different methods adapted to improve interoperability among metadata schemas for cross-domain metadata
harvesting. Halshofer and Wolfgang [8] while categorizing existing interoperability techniques, describe their characteristics, and compare their quality by analyzing their potential for resolving various types of heterogeneities.

The work in domain based harvesting that is rather closest to the present effort are Avano [9] and High North Research Documents [HNRD] [10] that harvest thematically from generic repositories based on OAI sets. Avano and HNRD use keyword research system and deploy manual verification to ascertain the suitability of the resources to added. In contrast to this, we propose in our model a step for verification with an AGROVOC lookup tool that verifies terms from agricultural and related domains that are indexed from the ‘title’, ‘keyword’ and ‘abstract’ fields. Details are discussed the sections that follow.

8. Proposed model for thematic harvesting

Indus only deals with aggregating resources from agriculture and related areas. It also functions as a data provider. It is possible using appropriate OAI command to harvest by specific sets under the domain ‘Agriculture’ in Indus. For example, if one wants to harvest only Indian Agricultural Repositories from Indus then the ListSets above displays ‘Identifiers’ against the Indian Agricultural repositories using which the relevant sets only could be harvested. However in generic repositories where sets are not based on domains, it is difficult to identify agriculture specific resources.

8.1. Harvesting by domain

As stated already in a generic repository such as a university repository where resources from all departments are uploaded it is difficult to identify resources by domains. It may be an ETD repository organized sometimes by different departments and sometimes simply by years. One example is provided below. In one of the ETD repositories, resources related to agriculture are under keywords as shown below and on closer examination these belong to many collections or sets including ‘Economics’, ‘Marketing’ among others (Fig. 5).

Proposed is a model for identifying agriculture related resources from generic repositories and re-harvesting selectively to provide access to such aggregated collections in agriculture.

Data collection: The process of data collection involves assembling and evaluating information on variables based on the interest. Gathering data in a systematic manner will enable one to achieve several goals, as it helps to answer research questions, test hypotheses, and evaluate outcomes. In order to achieve these goals, we collect the sets from generic repositories such as open access university repositories. Resources from harvested sets are also targeted. We mostly use directories such as DOAJ, OpenDOAR and OpenArchives to source data providers that are OAI-PMH compliant. In addition we conducted random Google based searches to check repositories that may contain agriculture related resources. These are included in the input sets.

Indexing: The datasets are indexed to enhance the speed of data retrieval operations on a database table. Indexes are used to identify the data location in order to avoid searching of every row in a database table, when each time the database is accessed. In this case all such sourced sets are indexed using Apache Solr. Solr is an indexing and search platform with a powerful retrieval mechanism (http://lucene.apache.org/solr/#intro). For our thematic harvester we use full-text indexing so that all words of the keywords, title and abstract fields are indexed.

Fig. 5 – Scattered agriculture related resources in a repository.
AGROVOC lookup: AGROVOC is a multilingual controlled vocabulary containing more than 40,000 concepts in the agriculture domains, including food, nutrition, fishery, animal husbandry, forestry, environment [11]. It is a structured thesaurus of terminology in agriculture and related domains and is used a default standard for vocabulary control in agricultural systems worldwide. AGROVOC is expressed as a concept scheme using SKOS (Simple Knowledge Organisation System), a data model for representation of classification schemes, taxonomies, thesauri and every structured vocabulary (Fig. 6).

In order to ascertain whether a certain record belongs to agriculture domain or not, we adapt an AGROVOC lookup tool. The indexed terms are compared with the terminology to find matches. The records of the matched terms are collected into an interim database.

Tharvest: The Theme harvester, Tharvest, then collects and organizes the records into Indus. Thus the records which match the AGROVOC terminology are ingested into the harvester. An interface that enables search and browsing by title, author, date or subject is presented to the end users though at present it is only as a proof of concept (Fig. 7).

We followed the steps described above in developing the model, Tharvest using AGROVOC for verifying records belonging to agriculture related domains automatically. However, in the process several issues were thrown up and some of these are highlighted below.

9. Issues in harvesting

In the process of thematic harvesting several data providers were accessed. There are a number of providers in agricultural domains as well several others that do offer metadata of agriculture resources but these may be in generic sets mixed with other domain records. Also harvesting depends mainly on two factors; compliance with OAI-PMH standard and the metadata standard/s used by repositories though there are also several other factors that affect harvesting [12]. Some of the salient issues are:

Linking: Linking mechanism in harvested records is based on URIs. Hence the standard of URI itself comes into focus. It is observed that URIs is not a problem with OJS (Open Journal System) based journals as also with ePrints (a repository software – http://eprints.org) based repositories. These use data providers’ address directly in the URI. Whereas DSpace uses CNRI handles for persistent IDs for data providers who are registered with CNRI service. However, a significant number of agricultural repositories that use DSpace have not registered with CNRI or any other such service for persistent IDs and simply use the default format, http://handle.net/123.9.html which leads nowhere. In such cases only metadata can be seen since the URI will not link to the actual

Fig. 6 – Tharvest flowchart.

Fig. 7 – Tharvest: Thematic harvest for agriculture.
full resource. During the process of populating Indus many records are missed due non-compliance of the data provider with a standard handle system.

BaseURLs: It is difficult to trace the baseURLs of repositories unless they are registered with one of the directory services such as OpenDOAR or Grainger. We are left to guess based upon whether the data provider is using something like DSpace or ePrints and what could be the baseURLs. In some cases the guess maybe correct but in others the implementers change the content or syntax for the baseURLs. It then becomes impossible to access such repositories for harvesting. The irony is that though OAI-PMH is standard for the interoperability there is no strict standard imposed for the syntax and content of baseURLs. In the process of thematic harvesting the standard directory services such as ROAR could only lead us to a few repositories. The others listed in Indus were collected by physically contacting the repository owners through social and professional networks.

9.1. Other issues

Some other issues are generic and not any platform specific and mostly to do with metadata.

In qualified Dublin Core there is a way to specify the vocabulary used and to what standard it is associated. But if plain Dublin Core is harvested it becomes difficult to associate the terminology with the standard. Further there is the issue of poor quality metadata. Many authors uploading resources are not aware of the metadata standards and they provide very minimal metadata. Jung-ran and Tosaka [13] in a study that explored the current state of metadata-creation practices across digital repositories observed that metadata interoperability remains a major challenge as there is a lack of exposure of locally created metadata and metadata guidelines beyond the local environment. Authors may often use non-standard vocabulary while providing metadata as they are not trained in the metadata standard and standard vocabulary. It then becomes the purview of the repository managers to implement metadata quality check by appointing metadata editors before committing resources to repositories.

10. Conclusion

Tharvest is a model for harvesting thematically from generic repository. At present the system is a test bed and successfully tried to harvest resources related to agriculture from three generic repositories more to provide proof of concept discussed here. The next step will be the full-fledged implementation and populating it with more records. In the end the harvested records will be populated into Indus. In the steps following its implementation it is intended to run evaluation for performance and also for comparing the success rate of the procedure of using AGROVOC based lookup tool for identifying agriculture related resources versus manual checks as adopted by other systems. The experience with building Tharvest brought forth several issues as discussed above and these are broadly applicable to thematic harvesting in any other domain.

REFERENCES