Hidden Data Sets, an Untapped Resource

Pickett, T., N. Marsh and B. Farthing

Queensland Environmental Protection Agency, E.P.A. Building, 80 Meiers Road, Indooroopilly, QLD, 4068.
E-Mail Trevor.Pickett@epa.qld.gov.au

Keywords: *metadata, data sets, web-based tool*

EXTENDED ABSTRACT

Planning and some modelling decisions in natural resource management are made based on available information, where the information could be in the form of expert opinion, data or technical reports. While there are many formal repositories of published data sets, there is also an untapped wealth of information in unpublished or hidden data sets. This hidden data is often either collected by members of the public or small focus groups and is not housed within Government databases. There is much valuable information which can be obtained from this data and there is a necessity to develop methods or tools to accurately catalogue and provide exposure for it.

Traditionally, a data warehouse would be used to store data, but for the storage and access of information in hidden data sets, data warehousing techniques have a number of limitations. To store and distribute data owned by others, a data warehouse provider must negotiate data sharing agreements with the data owners. These agreements will vary from data set to data set and require data warehouse providers to implement data access restrictions. The management, maintenance and facilitation of data access is a large task and data warehouses require dedicated managers and administrators to fulfill this function. Data warehouses are an excellent means of providing access to data sets or electronic reports, but lack the capacity to provide access to other information. For example, information that is only available in hard copy must be digitised or it can not be stored. Data storage is often limited to a specific file format, particularly if visualisation is required. These and other limiting factors make traditional data warehousing an inadequate method for the storage and retrieval of hidden data.

What is required is a method to expose hidden information, without the limitations of traditional data warehousing. Ideally, this method should incorporate strategies such as: unrestricted access, easy data retrieval and formal data set descriptors. Unrestricted access and easy data retrieval encourage user participation, while formal data set descriptors are necessary for standardisation and integration with existing resource repositories.

This paper will describe one such method, the Natural Resource Info Pool, developed by the National Action Plan for Water Quality and Salinity's Water Quality State Investment Program. This is an Internet based tool, which provides a facility to search and capture descriptions of data sets (metadata). It is designed as a minimal maintenance portal to a metadata database, which accurately captures information describing resources such as data sets, reports, images and maps. It provides three necessary functions for handling metadata and these are: a means of entering and storing meta-information regarding resources; a means of modifying entered meta-information when necessary; and a means of searching stored meta-information and extracting resource descriptions including data retrieval particulars.

Storing and distributing meta-information has distinct advantages and disadvantages when compared with traditional data warehousing techniques. Since the physical data is not housed in-situ, there is a significant reduction in storage space, Internet bandwidth, security concerns and user access restrictions, and these lead to a significant reduction in cost. Resource (particularly data set) cataloguing, classification, verification and validation is also unnecessary, since cataloguing and classification is an automated part of metadata submission and all verification and validation claims are made by the resource contributor. One clear disadvantage of distributing meta-information is the increased complexity of resource retrieval. Another disadvantage is that resource custodianship is a dynamic process a constant update of meta-information is required to keep this information current.

With traditional data warehousing techniques, data sets are surrendered to a central body and stored in a repository. Data set distribution control is the responsibility of the database administrator acting on behalf of the central body. The Natural Resource Info Pool provides a mechanism whereby resource contributors can maintain total control of their assets and reports. A resource contributor is only required to provide a comprehensive description of their data, and retains the right to distribute their resources as they see fit.

1 INTRODUCTION

Natural Resource Management is an information hungry process. Often, available data sets form the basis of many planning and modelling decisions. Additional data will generally strengthen the decision making process. But collecting additional data is often expensive and time consuming. Even low confidence data has its place in testing model responses and better understanding model results. A much better solution would be to utilise data collected by others. A recent survey was conducted in Queensland (Cawley 2004) on community-based groups such as waterwatch groups; landcare and landholder community catchment groups; and communities involved in externally managed natural resource and scientific projects. It showed that they are collecting a wide range of data on a variety of parameters including: sediment concentrations; clarity and turbidity; nutrients; macroinvertebrates; algae/blue green algae; ionic concentrations; pesticides; riparian condition; flow; birds; fish; instream habitat; geomorphology; catchment impacts; traditional owner; and groundwater. The survey indicated that there is a high degree of variation in data confidence, and that data with higher confidence levels was generally produced by groups participating in collaborative monitoring. The data sets and reports produced by these groups usually do not find their way into formal data repositories and often do not see the light of day. These hidden (or grey) data sets, represent an untapped wealth of information if used appropriately. What is required is a method which provides exposure for these data sets. If collaborative efforts in general produce better quality results, community groups require a means to identify others in their region undertaking similar monitoring and ideally, any method providing exposure for data sets would also facilitate this. This paper describes the Natural Resource Info Pool, an Internet based metadata tool created to provide exposure for and access to, these data sets and other resources.

1.1 Traditional data warehousing vs. metadata storage

The Natural Resource Info Pool is a metadata tool, storing only descriptions of resources rather than the resources themselves. There are several reasons for adopting this method over traditional data warehousing techniques. A traditional data warehouse may contain large documents, images, GIS files, complex data sets, large spreadsheets, or other large files, thus storage space is always an issue. Adequate storage space must be allocated for existing resources and extra space must be available to allow for growth. If large files are being stored then high bandwidth is required for access, as multiple users need to be able to quickly download these files. Security measures are required to protect the integrity of sensitive data sets stored in the warehouse and to keep out malicious intruders. Data contained in a central warehouse also needs to be catalogued, classified, verified and validated before distribution. Once established a data warehouse require dedicated data managers and administrators to perform regular maintenance and housekeeping. Each of these factors has an associated cost and when combined they make data warehousing an expensive exercise. Additionally, data distribution is the responsibility of the data warehouse owner. Resource owners usually relinquish control over their data sets when they are placed in a data warehouse, an unappealing prospect for many community-based groups. Finally, data warehouses provide no easy means of connecting community-based groups with other like-minded people.

A metadata database on the other hand, only contains descriptions of resources, and therefore requires much less storage space. For example 1000 metadata records in the Natural Resource Info Pool only requires about 1.5Mb of storage space. Since metadata database users only need to upload and retrieve very small records, much less bandwidth is necessary. Security measures and access restrictions can be kept to a minimum as well, since no sensitive information is stored in-situ. Metadata records are classified and catalogued automatically as part of the metadata submission process and verification and validation of resources becomes the responsibility of the end user and the resource custodian. All of these features make a metadata database much more cost effective when compared with traditional data warehousing techniques. In addition the Natural Resource Info Pool is designed to require a minimum of maintenance, significantly reducing ongoing costs. More importantly for community-based groups a metadata database allows them to retain complete custodianship of their data, distributing it only as they see fit. Additionally, the Natural Resource Info Pool requires contact information as a mandatory metadata descriptor, providing community-based groups in a region with a list of potential collaborators.

An additional and considerable constraint in the development of a data warehouse, is the need to establish data sharing agreements with each and every contributor. This onerous need to establish data sharing agreements would usually result in only information collected as part of large ongoing programs such as those run by state agencies to be housed in data warehouses. A metadata database does not require data sharing agreements, but simply allows the custodians of information to advertise its availability. Potential users of the information can negotiate directly with the data custodian.

Traditional data warehousing also requires a rigid data storage and retrieval structure, hence novel or small data sets or associated information such as data interpretation and reports are difficult to house and retrieve in a flexible way. Metadata descriptors are not prescriptive when describing information, hence meta-fields can relate to data sets, reports, assessments, web sites and other resources. This allows a user to search for 'information' on a given topic or a specific location and be provided with a list of not simply data sources but other valuable information.

Metadata storage does have one drawback when compared to traditional data warehousing techniques and this is the speed of data retrieval. However, if meta-tags are used correctly and users regularly update their stored metadata, then resource retrieval speed will be greatly improved. Furthermore, the Natural Resource Info Pool incorporates a metadata field pointing users to a web address, if the custodian has made the resources available online.

2 METHODOLOGY

The Natural Resource Info Pool was originally designed and produced by the National Action Plan Queensland's Water Quality State Investment Program in conjunction with the Queensland Environmental Protection Agency and the Queensland Department of Natural Resources and Mines to provide Regional Bodies in Queensland with access to existing water quality information about their catchments. Since its inception, the scope has been broadened somewhat to provide unrestricted access to general natural resource information with a specific focus on water quality. It is envisaged that subsequent versions of this tool will have a lesser focus on water quality resources in particular and a greater focus on all natural resources in general. The Natural Resource Info Pool is a series of Internet based forms, interfacing with a metadata database and providing users with a facility to enter, maintain, search and retrieve formal descriptions of data sets, reports, maps and other resources. Formal descriptions include retrieval information for each resource. The form components of the Natural Resource Info Pool have been written using a combination of HTML, JavaScript, and PHP technologies, while the relational metadata database is written in MySQL and resides on a remote server.

2.1 Metadata descriptors and field selection

The Natural Resource Info Pool is a metadata tool, incorporating a series of mandatory, semi-mandatory and optional meta-information fields to accurately capture resource descriptions. Fields were chosen so that mandatory and semi- mandatory fields provide the minimal information necessary for accurate resource retrieval and optional fields provide a comprehensive and unique description of the resource. These fields are based upon the metadata element set described in AS5044, the AGLS Metadata Standard (National Archives of Australia, 2002), an Australian standard for resource description. Table 1 contains all of the fields used in the Natural Resource Info Pool and a brief description of each. A user must complete all mandatory fields. Semi-mandatory fields provide an either-or choice. So, either the "Contact Corporation" field and at least one of the "Contact Address" or "Contact Telephone" or "Contact Email" fields must be completed or the "Identifier" field must be completed. Optional fields should be completed if the information is known. Mandatory and semi-mandatory fields provide the absolute minimum information regarding a resource, as recommended by the AGLS metadata standard.

As shown in Table 1, the fields used in the Natural Resource Info Pool are optimised for capturing information describing water quality resources. Fields such as "Water Body Type", "River Basin" and "Usage" have been included to specifically target this type of information. It should be noted however, that these fields are optional and contributors wishing to upload descriptions of information other than that regarding water quality can choose to ignore them.

2.2 Functionality

The Natural Resource Info Pool provides three necessary functions for handling metadata;

1. entering and storing meta-information regarding resources
2. modifying entered meta-information when necessary and
3. searching stored meta-information and extracting resource descriptions including data retrieval particulars.

Each of these functions is delivered through a separate Internet-based form. These forms are respectively, The "Metadata Upload Form", the "Metadata Modify Record Form", and the "Natural Resource Info Pool Search Form".

1. Entering and storing meta-information
The "Metadata Upload Form" is accessible at http://www.wqonline.info/FORMS/upload.php. Access to this form requires an initial registration process and a login step. Upon successful submission, the metadata record is amended with three additional

Table 1. Metadata fields used in the Natural Resource Info Pool

Field Name	Obligation	Description
Creator	Mandatory field	The author of the resource.
Search year	Mandatory field	The creation or availability year of the resource.
Title	Mandatory field	The title of the resource.
Publisher's name	Mandatory field	The resource publisher's name.
Publisher's address	Optional field	The resource publisher's address.
Subject	Mandatory field	Keywords for the resource.
Contact corporation	Semi-mandatory field	The resource custodian's corporation.
Contact address	Semi-mandatory field	The resource custodian's contact address.
Contact telephone	Semi-mandatory field	The resource custodian's telephone number.
Contact email	Semi-mandatory field	The resource custodian's email address.
Identifier	Semi-mandatory field	A unique identifier for the resource (eg. ISBN).
Alternative	Optional field	An alternative title for the resource.
Date created	Optional field	The creation date of the resource.
Date modified	Optional field	Any modification dates for the resource.
Date issued	Optional field	The publication date of the resource or similar.
Date valid	Optional field	A date (often a range) of validity of the resource.
Extent	Optional field	The physical size or duration of the resource.
Sites	Optional field	The number of sites sampled.
River basin	Optional field	As defined by AWRC (2005).
Water body type	Optional field	The water body type associated with the resource.
Scale	Optional field	The sampling scale of the resource.
Usage	Optional field	The scale at which the results of the resource can be applied.
Frequency	Optional field	The sampling frequency of the resource.
Samples	Optional field	The number of samples taken.
Monitoring	Optional field	The monitoring date range of the resource.
Abstract	Optional field	An abstract associated with the resource.
Content	Optional field	The parameters measured.
Methods	Optional field	The sampling methods used.
Format	Optional field	Current format of the resource (eg. hard copy, electronic, etc).
Rights	Optional field	Property rights or copyrights applying to the resource.

fields and added to the Natural Resource Info Pool's relational database. The three additional fields are: a date/time stamp field containing the current date and time; a record owner field containing the name of the user who submitted the record; and a moderated field. The Natural Resource Info Pool administrator checks the recently submitted record for inappropriate or malicious content and if none is found, changes the moderated field entry to "yes" and the entry is publicly accessible. The moderated field is a simple security measure designed to stop inappropriate and malicious content, it is not intended to be used for resource validation. Resource validation is the responsibility of the end user and the resource contributor.

2. Modifying entered meta-information

Modifying an existing metadata record is a two step process, comprising record selection and record modification. A user cannot modify the metadata records submitted by someone else, but users may change any field information in the records that they "administer" and resubmit the record. Upon successful submission, the Natural Resource Info Pool administrator must once again check that they contain no malicious content before making them publicly available.

3. Searching and extracting stored meta-information

Unlike the previous two forms, the "Natural Resource Info Pool Search Form" requires no login. It is accessible to anyone with an Internet connection at http://www.wqonline.info/FORMS/search.php. The primary function of the "Natural Resource Info Pool Search Form" is to facilitate metadata retrieval. It is a search form designed to retrieve metadata records from the database, based upon input from a user. The "Natural Resource Info Pool Search Form" allows both simple and complex search techniques in a single form. The user has the ability to create searches using expressions ranging from generic, which will return everything, to complex, which combines three distinct search terms with logical operators. There is also the option of applying the search term expressions to either a single field or a number of fields in the database. A user can restrict a search by selecting values for "Year", "River Basin" and "Water Body Type", from drop down boxes.

Thus, the Natural Resource Info Pool provides adequate functionality for users who want to

enter, modify, and search descriptions of resources, including hidden data sets.

2.3 Verification and validation

Software verification and validation are an important part of the software lifecycle process and may be defined as follows:

- Software verification answers the question "Are we building the software right?"
- Software validation answers the question "Are we building the right software?"

Verification of the Natural Resource Info Pool has been an integral part of its construction. End users will be disinclined to utilise a tool which is error prone or unreliable. It is important to identify and remove as many bugs as possible before the tool is released. To achieve this, a series of testing procedures was adopted during the construction of the Natural Resource Info Pool. Each individual component of the Natural Resource Info Pool was tested as it was constructed. When bugs were discovered, they were rectified and re-tested. Testing procedures incorporated both "White Box Testing" (the testing of program statements, logic, loops and functions) and "Black Box Testing" (input/output testing). Individual components were then integrated into the Natural Resource Info Pool and further testing carried out. Because the Natural Resource Info Pool is an Internet based tool, platform and browser-specific issues needed to be uncovered and resolved. Before it was released, the Natural Resource Info Pool was trialed on several machines, with different operating systems and web browsers. This ensures that users with contemporary and legacy technology are able to use the tool.

Validation of the Natural Resource Info Pool is currently an ongoing process. A number of small workshops and one-on-one tutorials are being conducted with potential end users. These workshops have a dual purpose. Firstly, they introduce the Natural Resource Info Pool to potential users and secondly and more importantly from a validation perspective, at the end of the workshop users have an opportunity to provide feedback on the Natural Resource Info Pool. They answer important questions about the usefulness and functionality of the tool and are encouraged to suggest changes which could be made, and to alert the development team to other similar tools which may be able to be incorporated into the Natural Resource Info Pool.

3 CONCLUSION AND FUTURE DIRECTIONS

This paper has presented a method (The Natural Resource Info Pool) for capturing, storing and retrieving resource information, particularly information regarding "hidden" data sets owned by community-based groups. It has touched briefly on the motivation for sharing these data sets, both for the owner and other potential users. It has also demonstrated that the Natural Resource Info Pool can provide a list of contacts for community-based groups seeking like-minded people within their region. There has been discussion and reasons given for choosing metadata storage over traditional data warehousing techniques including cost and accessibility. Metadata fields used in the Natural Resource Info Pool have been listed, together with a small description of each field. Each of the three primary functions required by metadata storage and distribution (entering and storing metadata records, modifying existing metadata records and retrieving metadata records) has been explained as has the delivery mechanisms for each function incorporated into the Natural Resource Info Pool. Finally, the validation and verification techniques used in the creation and modification of the tool have been explained. In short, the discussion so far has shown that the Natural Resource Info Pool is an easily maintained and cost effective tool for capturing, storing and distributing description of resources. It is robust and reliable and adequately suited for exposing hidden data sets.

To round out this discussion, one further important point regarding the Natural Resource Info Pool should be discussed, and this is: How successful is the Natural Resource Info Pool?

How is the success of a tool such as the Natural Resource Info Pool judged? If the Natural Resource Info Pool were a traditional data warehouse, then it might be argued that the number of users or the number of records in the database could be good metrics. There are however, several factors which must be taken into consideration, when quantifying the success of the Natural Resource Info Pool this way. Firstly, the Natural Resource Info Pool was created with a specific target audience in mind. It has been designed to target users who have data sets which would not usually appear in formal data repositories and this restricts the size of its user base. Secondly, the Natural Resource Info Pool is a relatively new innovation, and many members of the natural resource community are unaware of its existence. While it is true that workshops and tutorials have promoted the tool and increased community participation, it will still require significant effort for it to gain widespread acceptance.

A much better measure of success for the Natural

Resource Info Pool, is to assess whether it fulfills its intended purpose: does it provide exposure for hidden data sets and if so, are these data sets being used in modelling and decision making? Initial response indicates that it does provide exposure for data sets. Since its release early in 2005, response to the tool has been very positive and metadata record numbers in the database have been steadily growing from 0 in March 2005 to 1260 records in September 2005). Although this clearly shows that the Natural Resource Info Pool is being used, there is no current method to determine whether the search and retrieve functions of the Natural Resource Info Pool are being utilised. There is also currently no evidence to suggest that the datasets in the Natural Resource Info Pool are being used for decision making and modelling. However the tool is still reasonably new and it is felt that with repeated workshops and additional exposure, the Natural Resource Info Pool will be used to its full potential.

4 REFERENCES

AWRC. (2005), Australia: Drainage Divisions and Basins, Map 5, http://www.deh.gov.au/water/wetlands/database/directory/appendix3.html#map-drainage.

Cawley, R, (2004), Community Water Quality and Stream Health Monitoring Survey: Review Document. National Action Plan for Salinity and Water Quality, Queensland Water Quality State Investment Projects. Unpublished technical document.

The National Archives of Australia, (2002), AGLS Meta Data Standard, Australian Standard AS 5044, Standards Australia.